Evolutionary trace report_maker: a new type of service for comparative analysis of proteins

I. Mihalek*, I. Reš and O. Lichtarge

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received on March 7, 2006; revised on April 4, 2006; accepted on April 20, 2006

Advance Access publication April 27, 2006

Associate Editor: Martin Bishop

ABSTRACT

Summary: Evolutionary trace report_maker offers a new type of service for researchers investigating the function of novel proteins. It pools, from different sources, information about protein sequence, structure and elementary annotation, and to that background superimposes inference about the evolutionary behavior of individual residues, using real-valued evolutionary trace method. As its only input it takes a Protein Data Bank identifier or UniProt accession number, and returns a human-readable document in PDF format, supplemented by the original data needed to reproduce the results quoted in the report.

Availability: Evolutionary trace reports are freely available for academic users at http://mammoth.bcm.tmc.edu/report-maker

Contact: (imihalek,ires,lichtarge)@bcm.tmc.edu

IMPLEMENTATION AND DEPENDENCIES

ET report_maker is implemented as a set of interacting Perl modules, with computationally demanding parts written in C, and visualization of the mapping of results onto primary sequence implemented in Java. Report_maker draws on the work of many researchers outside of our own group. The programs, all free for academic users, that report_maker depends on are the following: alistat (statistical profile of a multiple sequence alignment; part of the HMMER package http://hmmer.wustl.edu), BLAST (sequence database search; (Altschul et al., 1997)), CE [structural alignment of proteins, used in the mapping of geometrically determined ligand binding surfaces between different Protein Data Bank (PDB) entries; (Shindyalov and Bourne, 1998)], ClustalW [multiple sequence alignment for a small number of sequences; (Thompson et al., 1994)], DSSP [determination of protein surface; (Kabsch and Sander, 1993)], LaTex [typesetting of the final text; (Lamport, 1986)], Muscle [multiple sequence alignment for a large number of sequences; (Edgar, 2004)], PyMol [structure visualization; (DeLano, 2002, http://www.pymol.org)]. It also relies on the following publicly available databases: HSSP (Sander and Schneider, 1991), PDB (Berman et al., 2000) and UniProt (Boeckmann et al., 2005).

METHODS

Real-valued evolutionary trace. To rank the evolutionary importance of residues, report_maker uses real-valued ET, described in Mihalek et al. (2004).

Heuristic suggestions for disruptive mutations. Report_maker makes some heuristic suggestions for mutations, meant to be disruptive to the interaction of the protein with its ligand. They are based on complementarity of the proposed mutation with the physical and chemical properties of a residue and its substituents found in the alignment.

The attempt is made to complement the following 11 properties of an amino acid: small \{AVGSTC\}, medium \{LPNQDEMK\}, large \{WFYHR\}; hydrophobic \{LPVAMWF\}; polar \{GTYC\}; positively \{KR\} or negatively charged \{DE\}; aromatic \{WFYH\}; long aliphatic chain \{ERQ\}; OH-group position \{SDETY\}; and NH2 group position \{NQKR\}. For a given column i, a score $S_i(a)$ is assigned to each of 20 amino acid types $a$:

\[ S_i(a) = \frac{1}{p} \sum_{p=1}^{11} \left( |s_p(a) - s_p(q_i)| + |s_p(a) - s_p(a_j)| \right), \]

where the sum runs over the 11 properties $p$; $s_p(a)$ is a function that assigns 1 to an amino acid $a$ that has the property $p$ (given in square brackets), and 0 to all other types. $q_i$ is the amino acid type at the i-th position of the query protein, and $s_p(a_j)$ is average for all substitution amino acid types found in the neighbor sequence.

*To whom correspondence should be addressed.
the column \( i \). Thus, \( S_i(a) \) is the biggest for the cases when the amino acid type \( a \) differs in the maximal number of properties from the amino acid types already seen in the alignment column. Amino acid types with high \( S_i(a) \) are listed in \textit{report maker} as disruptive mutation suggestions for some top ranking residues.

\textit{Geometric estimate of functional residues.} In \textit{report maker}, protein residues are estimated to be involved in ligand interaction if in the co-crystal they have at least one heavy atom within 5 Å of the ligand.

\textit{Sequence selection.} \textit{Report maker} imposes the following requirements on sequences: they should not be <75% of the query, no pair should be >99% identical and the final homologue set should consist of at least 10 sequences; also to perform meaningful database searches, the minimal length of the analyzed protein sequence be 20 residues.

When the structure is known but no HSSP alignment is available, \textit{report maker} uses Monte Carlo sequence selection procedure described in Mihalek \textit{et al.} (2006).

**FEATURES**

\textit{ET report maker} takes as input a PDB identifier or UniProt accession number. The amount of work it does subsequently, as well as the output, depends on the data available for the query protein (or complex).

The default output is a brief statistical description of the alignment used, in terms of its homologue and taxonomical content, and the estimate of the evolutionary pressure on the protein residues, mapped onto the primary sequence.

If the input is a PDB identifier, or if a related structure can be found, the following additional output is produced: (1) structural map of the evolutionary pressure, (2) discussion of binding sites for known small ligands and protein binding partners, (3) outline of potential novel active sites on the protein surface and (4) suggestions for mutations to block (disrupt) protein function through known and putative binding surfaces.

If the input is PDB identifier and the related entry consists of several protein chains, all chains are discussed.

The whole analysis is presented in the form of a human-readable and printable document.

Finally, \textit{report maker} will produce an accompanying package of data essential for reproducing results given in the report: the alignments in GCG format; a brief description of sequences used; the raw ET output; PyMol scripts with mapping of ET data onto structure (when applicable) and an etvx file for ET Viewer (D. Morgan and O. Lichtarge, manuscript in preparation, http://mammoth bcm.tmc.edu/traceview).

The list of possible extensions for \textit{report maker} is long: a better evaluation of the impact of suggested mutations [see (Capriotti \textit{et al.}, 2005) and references therein], difference analysis (Madabushi \textit{et al.}, 2004), prediction of specificity determinants and an expert input page for interested users, to name a few possibilities. At the same time it is our hope that the current version of \textit{report maker} will serve as an invitation for collaboration with experimental groups, which may be aware of the possibility for customization of the report, such as by providing a model structure or relevant selection of sequences for the comparative analysis.

**CONCLUSION**

In the current era of information overflow, \textit{ET report maker} provides a new data accumulation and presentation service, allowing the users to focus quickly on interesting features of protein under the investigation. The resulting printable report, is our hope, will help bridge the time-consuming gap between multiple databases and services for comparative analysis of proteins, and the laboratory bench.

**ACKNOWLEDGEMENTS**

The authors thank Mr José Rivera for the technical help in setting up the server, and the members of the Lichtarge laboratory for testing and criticizing the reports. This work was supported by grants from the National Institute of Health (GM066099), the National Science Foundation (DBI #0318415) and the March of Dimes (MOD FY03-93).

**Conflict of Interest:** none declared.

**REFERENCES**


