Recent improvements of the ProDom database of protein domain families

Florence Corpet*, Jérôme Gouzy1 and Daniel Kahn1

Laboratoire de Génétique Cellulaire, INRA, BP27, F-31326 Castanet-Tolosan cedex, France and 1Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes, INRA/CNRS, BP27, F-31326 Castanet-Tolosan cedex, France

Received October 2, 1998; Revised October 7, 1998; Accepted October 14, 1998

ABSTRACT
The ProDom database contains protein domain families generated from the SWISS-PROT database by automated sequence comparisons. The current version was built with a new improved procedure based on recursive PSI-BLAST homology searches. ProDom can be searched on the World Wide Web to study domain arrangements within either known families or new proteins, with the help of a user-friendly graphical interface (http://www.toulouse.inra.fr/prodom.html). Recent improvements to the ProDom server include: ProDom queries under the SRS Sequence Retrieval System; links to the PredictProtein server; phylogenetic trees and condensed multiple alignments for a better representation of large domain families, with zooming in and out capabilities. In addition, a similar server was set up to display the outcome of whole genome domain analysis as applied to 17 completed microbial genomes (http://www.toulouse.inra.fr/prodomCG.html).

INTRODUCTION
The ProDom database of domain families aims at systematically collecting families of homologous domains on the basis of an automated analysis of available protein sequence data (1,2). This is useful for analysing the domain arrangements of proteins, particularly in large and complex protein families. It helps the dissection of homology relationships for modular proteins, and the clustering of homologous domains provides a rational way to organise protein sequence data. An interactive graphical user interface was designed to allow for easy navigation between schematic domain arrangements, multiple alignments, SWISS-PROT entries (3), PROSITE patterns (4) and 3-D structures in the PDB (5). New sequences can be searched against ProDom and readily aligned with existing domain families, or modelled on the basis of homologous domains in the PDB. Such uses of ProDom were described previously (2).

In order to improve ProDom quality we introduced two important changes in the ProDom construction procedure. First, we have included some expert validated domain families in order to improve the accuracy of domain boundaries. Second, we have developed a completely new version of the MKDOM program used to build ProDom (6), exploiting features of the recursive PSI-BLAST homology search algorithm (7). Not only does this new version of MKDOM process large families more efficiently, with a much decreased tendency towards domain fragmentation, it also scales much better with database size.

In addition, two major new developments are now available on the ProDom WWW server. The first development aims at facilitating the analysis of sequence relationships within large domain families. Indeed multiple sequence alignments become impractical when handling several hundreds of homologous sequences. They have been replaced by summary alignments and trees which can be developed or reduced according to user-specified parameters. The second development is ProDom-CG (‘Completed Genomes’), in which the MKDOM methodology is put to use to analyse domain arrangements in all proteins derived from completed genomic sequences (6).

BUILDING THE DATABASE

Method
Since version 35, the automated process that builds ProDom has been complemented by the result of an expertise. For some domain families defined in previous versions of ProDom, experts have been asked to correct domain boundaries on the basis of both sequence and structural information. In a first step, a position-specific scoring matrix is built for each expert-validated family and the source database, SWISS-PROT (3), is searched with the PSI-BLAST program (7). Domains that are found as members of these expert-validated families are extracted from the source database. If the extracted domain is not terminal in a protein sequence, the remaining sequence is cut into two parts that become two independent entries in the source database. In a second step, the other families are built with an entirely new process [MKDOM2, published elsewhere (6)], also based on PSI-BLAST. Each time a domain family is found, the corresponding sub-sequences are extracted, as in the first step, from the source database, the size of which is thus decreasing. The process stops when PSI-BLAST does not find any similarity between the remaining sequences. Multiple alignments are then systemati-

*To whom correspondence should be addressed. Tel: +33 561 28 54 31; Fax: +33 561 28 53 08; Email: fcorpet@toulouse.inra.fr
cally generated for all families using the MultAlin program (8), and a consensus sequence is calculated as the best weighted average sequence for each multiple alignment. Large families are processed much better with this new procedure than with the former version of MKDOM so that a better coverage of SWISS-PROT is reached (Table 1).

The ProDom building process can be applied to any set of protein sequences, provided there are enough sequences to detect domain boundaries. SWISS-PROT is a suitable source database, because it includes many sequences and fragmentary sequences can be avoided. But SWISS-PROT is not exhaustive and does not include all recent information from large scale sequencing projects. We have thus applied the ProDom building process to a set of available complete genomes and built ProDom-CG.

We request that users of ProDom or ProDom-CG should cite this article.

**Database format**

ProDom is built as a text file, ‘prodom.srs’. Each entry is a domain family with an identification (ID), a line of keywords (KW), the number of domains in the family (ND), the alignment length (LA), the multiple domain alignment (AL), a consensus sequence (CO), and links to PROSITE patterns and PDB entries if relevant (DR). The automatically generated keywords are the most frequent protein names in the family and the most frequent words in the corresponding SWISS-PROT description field (DE). The multiple alignment is presented with one line for each domain; each line includes the SWISSPROT ID and accession number (AC) of the protein, the start and end position of the domain, a weight in the alignment and the aligned domain sequence. PROSITE and PDB links are calculated by direct comparison with ProDom domains using the LASSAP program (9). This new ProDom format allows its inclusion in any SRS server (10); we also provide the necessary SRS icarus files. The previous ProDom file format (‘prodom.mul’ for multiple alignments and ‘prodom’ for consensus sequences) is still maintained for compatibility with previous versions. We provide a tool (FETCHDOM) to retrieve the domain decomposition of any protein that is present in ProDom, or to fetch a ProDom entry. This tool now uses indices and is much faster than before; for compatibility, it also works with previous non-indexed versions of ProDom.

**Content of the current release**

Release 36 of ProDom (August 1998) contains 57 976 domain families. These are sorted by decreasing number of protein sequences in the families. Each non fragmentary sequence from SWISS-PROT release 36 is treated in ProDom 36. More recent sequences and fragmentary ones will be added if they share similarity with a ProDom domain family in secondary releases (i.e., 36.1). The database requires ~60 Mb of disk storage space. The present distribution frequency is one main release for each SWISS-PROT release (with the same release number).

Release 17 of ProDom-CG was constructed by automatic clustering of protein domains derived from 17 complete genomes available on July 23, 1998: four archaea, 12 bacteria and one eukaryote. The present distribution frequency is four releases a year (the release number is the number of completed genomes included in ProDom-CG).

**THE WORLD WIDE WEB ProDom SERVER**

The most efficient and user-friendly way to browse ProDom interactively as well as to perform similarity searches is to use the WWW ProDom server which can be accessed at [http://www.toulouse.inra.fr/prodom.html](http://www.toulouse.inra.fr/prodom.html). As described before (2), the ProDom server displays ProDom entries and their relevant links to other databases. It provides a graphical representation of protein domain arrangements, and it allows for ProDom similarity searches with graphical outputs.

ProDom domain families can be accessed through ProDom numerical IDs, through keywords, or through relevant PROSITE (4) or PDB (5) entries. The display of a ProDom domain family includes its number with its pattern for graphical representation, general information about the family (keywords and size), a list of appropriate PROSITE patterns and PDB structures with hypertext links. This page is linked to the PredictProtein server for secondary structure prediction ([http://www.embl-heidelberg.de/predictprotein/predictprotein.html](http://www.embl-heidelberg.de/predictprotein/predictprotein.html)) (11): the form is automatically filled in with the ProDom family as a multiple sequence alignment, requesting both secondary structure and solvent accessibility prediction; the user should then identify himself and may select other options before running the prediction.

A graphical view presents domain arrangements for proteins sharing homology. Each protein is shown on a single line. It starts with its name, hypertext-linked to SWISS-PROT. Then, the domain arrangement is displayed with schematic boxes hyper-text-linked to corresponding ProDom entries.

ProDom can be searched for similarity with a query sequence using BLAST tools (NCBI-BLASTP, WU-BLASTP or WU-BLASTX (http://blast.wustl.edu/). BLAST results are followed by a graphical representation of a proposed domain arrangement for the query. The user can then request that the query be aligned with one of the proposed ProDom domain families using MultAlin (8), or that 3-D models of domains be generated on the basis of homology using SWISS-MODEL (12), where applicable.

**Table 1. Comparison of ProDom versions 34 and 36**

<table>
<thead>
<tr>
<th></th>
<th>ProDom 34</th>
<th>ProDom 36</th>
</tr>
</thead>
<tbody>
<tr>
<td>non fragmentary SWISS-PROT sequences</td>
<td>52,772</td>
<td>66,756</td>
</tr>
<tr>
<td>size in amino acids</td>
<td>20,065,998</td>
<td>25,365,087</td>
</tr>
<tr>
<td>ProDom domain families</td>
<td>53,599</td>
<td>57,976</td>
</tr>
<tr>
<td>size in amino acids</td>
<td>17,003,787</td>
<td>24,328,094</td>
</tr>
<tr>
<td>SWISS-PROT coverage</td>
<td>85%</td>
<td>96%</td>
</tr>
<tr>
<td>mean domain length in amino acids</td>
<td>126</td>
<td>128</td>
</tr>
<tr>
<td>ProDom domain families with at least two sequences</td>
<td>18,086</td>
<td>17,777</td>
</tr>
<tr>
<td>size in amino acids</td>
<td>12,304,724</td>
<td>19,625,004</td>
</tr>
<tr>
<td>SWISS-PROT coverage</td>
<td>61%</td>
<td>77%</td>
</tr>
<tr>
<td>mean domain length in amino acids</td>
<td>113</td>
<td>152</td>
</tr>
<tr>
<td>ProDom consensus size in amino acids</td>
<td>6,725,485</td>
<td>7,387,625</td>
</tr>
<tr>
<td>ProDom versus SWISS-PROT compression ratio</td>
<td>2.98</td>
<td>3.43</td>
</tr>
</tbody>
</table>
**ProDom domain families (Fig. 1)**

The display of a ProDom entry has been renewed and allows for better representation of large families. A ProDom family can be represented by a multiple alignment of all homologous domains and by a phylogenetic tree calculated from this alignment (details of the method will be published elsewhere). The multiple alignment and the tree are difficult to read when the family is large and we propose to cluster the less distant domains following the tree. By default, the family is represented by a tree with a maximum of 12 leaves which are either single domains or domain clusters and by a multiple alignment of sequences that represent these leaves: the sequence of a single domain or a consensus sequence of a domain cluster. A cluster is named after one of the proteins it includes, followed by the number of domains in the cluster. Domains whose distance is less than 20 PAMs are clustered. The user can change the maximum number of leaves and the minimum distance between two leaves. Many protein names are hidden because only one single name can represent each cluster. Thus, one may request that clusters be named after the proteins containing a given word, if present. The tree and the multiple alignment representations are displayed in two different linked windows. One can get representations of a sub-family of a ProDom entry by clicking on the corresponding cluster node in the tree representation or on the plus sign in the cluster line header in the alignment representation. For a sub-family, the user can again choose the clustering parameters and get tree and alignment representations. Colours have been chosen for protein names, following the four domains of life (Eukarya, Bacteria, Archaea and Viruses); if a cluster bridges two or more domains of life, its name is in a neutral colour. Colours are also used to emphasise sequence positions that are conserved in the whole family or a sub-family. This makes it easier to visualise and analyse large ProDom families and to navigate between partial tree and alignment representations.

**The ProDom-CG server**

ProDom-CG, the database of domain families built from complete genome data, can be accessed at http://www.toulouse.inra.fr/prodomCG.html. This server exhibits functions similar to the ProDom server. In addition, for each ProDom-CG family, a
Figure 2. ProDom-CG output. (a) List of ProDom-CG families meeting the criteria indicated on top (partial). (b) Graphical representation of all proteins containing ProDom-CG domain family 270. Each domain type is represented by a specific colour pattern which is hypertext-linked to the corresponding multiple alignment. Protein names are linked to the corresponding NCBI entries (13), SWISS-PROT (14) for *Bacillus subtilis* and MIPS (15) for *Saccharomyces cerevisiae*. The bottom table indicates the number of sequences in the family found in each complete genome.

The query form gives the same possibility as the ProDom one. In addition, it is possible to get the lists of ProDom-CG families that answer a criterion such as «present in all Archaea, present in some Bacteria and absent in Eukarya».

**ACCESS**

**Anonymous FTP site**


**Email server**

prodom@toulouse.inra.fr

Send the word HELP as the only word in the message body.

**WWW server**

http://www.toulouse.inra.fr/prodom.html
http://www.toulouse.inra.fr/prodomCG.html

**ACKNOWLEDGEMENTS**

We wish to thank Amos Bairoch, Claude Chevalet, Philippe Dessen, Laurent Duret, Jérôme Gracy, Manuel Peitsch and...
Burkhard Rost for stimulating discussions and exchange of information. The ProDom project is supported by the Centre National de la Recherche Scientifique (Genome Initiative) and the European Union (Biotech BIO4-CT980052).

REFERENCES