ProDom: Automated clustering of homologous domains

Florence Servant, Catherine Bru, Sébastien Carrère, Emmanuel Courcelle, Jérôme Gouzy, David Peyruc and Daniel Kahn

Date received (in revised form): 25th June 2002

Florence Servant
is a programmer in the Swiss-prot group at the EBI whose research interests include automated clustering methods of proteins, multiple alignment improvements and graphical user interface design.

Catherine Bru, Sébastien Carrère, Emmanuel Courcelle, Jérôme Gouzy, David Peyruc and Daniel Kahn
work at the Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes (INRA/CNRS) in the ProDom team headed by Daniel Kahn.

Keywords: *protein, database, domain family, clustering, sequence alignment*

Abstract

The ProDom database is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases. An associated database, ProDom-CG, has been derived as a restriction of ProDom to completely sequenced genomes. The ProDom construction method is based on iterative PSI-BLAST searches and multiple alignments are generated for each domain family. The ProDom web server provides the user with a set of tools to visualise multiple alignments, phylogenetic trees and domain architectures of proteins, as well as a BLAST-based server to analyse new sequences for homologous domains. The comprehensive nature of ProDom makes it particularly useful to help sustain the growth of InterPro.

INTRODUCTION

The ProDom database[1,2] has been designed to cope with the large amount of data available in the protein sequence databases. In fact the SWISS-PROT and TrEMBL databases[3] contained more than 700,000 entries in April 2002 and this figure is increasing exponentially. It is therefore useful to be able to analyse such an amount of data in an automated process. ProDom and its close parent ProDom-CG[4] are based on an iterative PSI-BLAST[5] search method in order to cluster protein segments into homologous domain families. This paper describes the construction method used for both types of ProDom, and the graphical interface that enables easy browsing and analysis of biological sequences. Some examples of applications of ProDom are provided.

ProDom CONSTRUCTION

The ProDom database is based on a method of automated clustering of homologous domains into families.[6] The source sequence database used to build ProDom is composed of a non-redundant set created from SWISS-PROT, TrEMBL and the complete proteomes available on the ExPASy server[7] and on the Proteome Analysis pages.[8]

The clustering program MKDOM2 relies on the assumption that the shortest amino acid sequence corresponds to a single domain protein. Therefore this sequence can be used as a query to screen the sequence database with the PSI-BLAST program in order to cluster homologous domains. A ProDom family is created from these homologous segments, which are then removed from the database. If a segment is found within a larger amino acid sequence, the remaining segments are kept in the database. This process is iterated, using the shortest sequence as a query, until all sequences in the database have been exhausted.

Some constraints and refinements must be added for this method to be efficient. First, the starting hypothesis is not applicable to fragmentary sequences as they may infer incorrect domain boundaries. Therefore fragmentary sequences are removed from the database. Second, low-complexity regions must be masked in order to prevent spurious matches. This is achieved by filtering the database with the SEG program.[9] Third, a

Daniel Kahn,
Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes,
INRA/CNRS,
BP27,
F-31326 Castanet-Tolosan cedex,
France

Tel: +33 561 28 53 29,
Fax: +33 561 28 50 61
E-mail: dkahn@toulouse.inra.fr
Requests about ProDom to:
proquest@toulouse.inra.fr

An automated process to extract domains from protein sequences

minimal length of 20 amino acids was set for a domain to be included in the analysis. Fourth, we use a procedure to detect internal repeats which are used as a query instead of the entire sequence. One important adjustment is to set proper PSI-BLAST parameters. ProDom construction is an automated process, and in order to limit the number of false positives, a stringent threshold is chosen for the expect value. For PSI-BLAST version 2.0.11 and a size of about 400,000 sequences of the source database, the *E*-value threshold was set at 10^{-6} which yields a good compromise between sensitivity and specificity. For this *E*-value to be consistent throughout the process, only the initial size of the database is used for *E*-value estimation.

In addition the ProDom construction process uses external information from some well-characterised domain families. This takes place at an early stage in the process. A position-specific scoring matrix (PSSM) is built for each of these families and used as a PSI-BLAST query in order to systematically recruit homologous domains. Currently, two series of expert validated families are used in ProDom: 21 domain families defined by the ProDom team and 383 additional families selected from Pfam-A 4.3 on the basis of sequence homogeneity.[10]

Transfer of ProDom accession numbers using MatchDom

As ProDom is built anew each time, it can prove difficult to track families across different releases. Moreover domain families may be fused or split in new releases of ProDom. Therefore we introduced accession numbers in order to track domain families across successive releases of ProDom. A procedure to transfer accession numbers from release $n-1$ to release n was developed and implemented in the MATCHDOM program. For each family F_n in ProDom release n, MATCHDOM searches for sequence overlaps in families F_{n-1} from release $n-1$. MATCHDOM evaluates the quality of the overlap between F_n and F_{n-1} using the number of domains involved in the overlap N_{overlap} and the total number of domains in F_{n-1}. The percentage of overlapping domains is computed as follows:

$$P_{\text{overlap}} = \frac{N_{\text{overlap}} \times 100}{\text{Number domains in } F_{n-1}}$$

Overlaps are sorted by decreasing N_{overlap} (first key) and P_{overlap} (second key). For each family in release n, the list of overlapping families from release $n-1$ is built and sorted. Accession numbers are transferred first from families with the best overlaps. For each family F_n in release n, we identify the family F_{n-1} from release $n-1$ exhibiting the best overlap with F_n. The corresponding accession number is assigned to F_n if it has not yet been assigned to another family. Otherwise the next best overlap is used until a suitable accession number can be transferred. If this fails a new accession number is created for F_n.

Current release of ProDom

The latest ProDom release, ProDom 2001.3, was built from SWISS-PROT version 39.27 and TrEMBL version 17.13 (26th September, 2001). The total number of non-fragmentary sequences from this data set is 373,869. The current release of ProDom contains 108,076 families with at least two domains among the 305,465 domain families generated by the clustering method. This makes ProDom an ideal tool with which to identify novel domain families.

The ProDom-CG database

In order to facilitate whole genome studies, ProDom-CG was initially created as an independent database constructed from 17 complete genome sequences.[6] The current ProDom-CG release contains 47 complete genomes and is an extraction of the standard ProDom release. Domains that belong to a complete genome are taken from the standard ProDom release to form ProDom-CG. Accession numbers are

kept identical between both flavours of ProDom except for the 'PD' prefix which is replaced by 'CG' in ProDom-CG. ProDom-CG47 was built from ProDom 2001.3 using 158,245 protein sequences from 47 complete genomes including 34 bacteria, 9 archaea and 4 eukaryotes. This led to 49,943 families with at least two domains among 182,217 domain families. The use of accession numbers makes it possible to compare successive releases of ProDom-CG and to compare ProDom-CG with ProDom.

ProDom graphical interface

A user-friendly graphical interface is required to cope with huge data sets

The ProDom web interface is based on a graphical representation of protein domain arrangements using domain specific cartoons. These representations of domain arrangements are stored as pre-computed images which are put together in a Berkeley database. The ProDom graphical interface[2] gives access to three different ways to query the ProDom database.

First, a query sequence (protein or DNA) can be used to search the ProDom database using the BLAST program (BLASTP or BLASTX)[11] through the BlastProDom wrapper tool. A similarity search is performed using either ProDom consensus sequences as a target database or, more sensitively, using all individual domain sequences present in ProDom. The results are filtered out by BlastProDom to retain only the best hit for each domain family at any given position in the query. This search is as fast and sensitive as a direct search against the primary sequence database, but in addition the filtered output directly provides a possible domain arrangement. Additional functionalities are accessible after a BLAST search on the ProDom server, such as the alignment of each predicted domain with the corresponding ProDom family, or homology modelling with the SWISS-MODEL[12] server.

A second way to access the ProDom database is the main ProDom form. ProDom domain families can be accessed through their accession numbers, through keywords related to a protein or a whole family, or through relevant InterPro,[13] PROSITE,[14] PDB[15] and Pfam-A[10] entries. Moreover, complex queries involving ProDom, SWISS-PROT+TrEMBL and the cross-referenced databases can be performed using the SRS environment.[16]

Finally, the main ProDom form allows the user to get a direct graphical representation of the domain composition of one or several proteins.

A ProDom domain family is structured in several parts as shown on Figure 1. The ProDom entry is characterised by its accession number and the release number. Some information related to the graphical representation is also given, such as the picture used for the graphical representation of this domain family and a link to the simplified graphical output of the proteins belonging to the family. In this simplified view, represented on the top left of Figure 1, all proteins sharing the same domain architecture are represented by only one of them, with a link to their complete list. This allows the user to view at a glance the different contexts in which a domain of interest can be found. The complete graphical output can also be displayed if required. In addition to the domain composition view, a tree representation of the family is available to visualise the relationships between domains in the family. Any internal node in the global rooted tree defines a sub-tree and thus a sub-family. A sequence family or sub-family is represented by summary alignments and trees can be pruned or expanded using DisplayFam.[17] The level of detail is defined by a maximal number of leaves, maxleaves, and a minimal distance between leaves, mindist. Default values are 12 leaves for maxleaves and 20 PAM (82 per cent identity) for mindist. The user can choose how the tree or the alignment is summarised by these two parameters. A leaf is labelled by sequence name or, if it corresponds to a cluster, by one of the sequence names

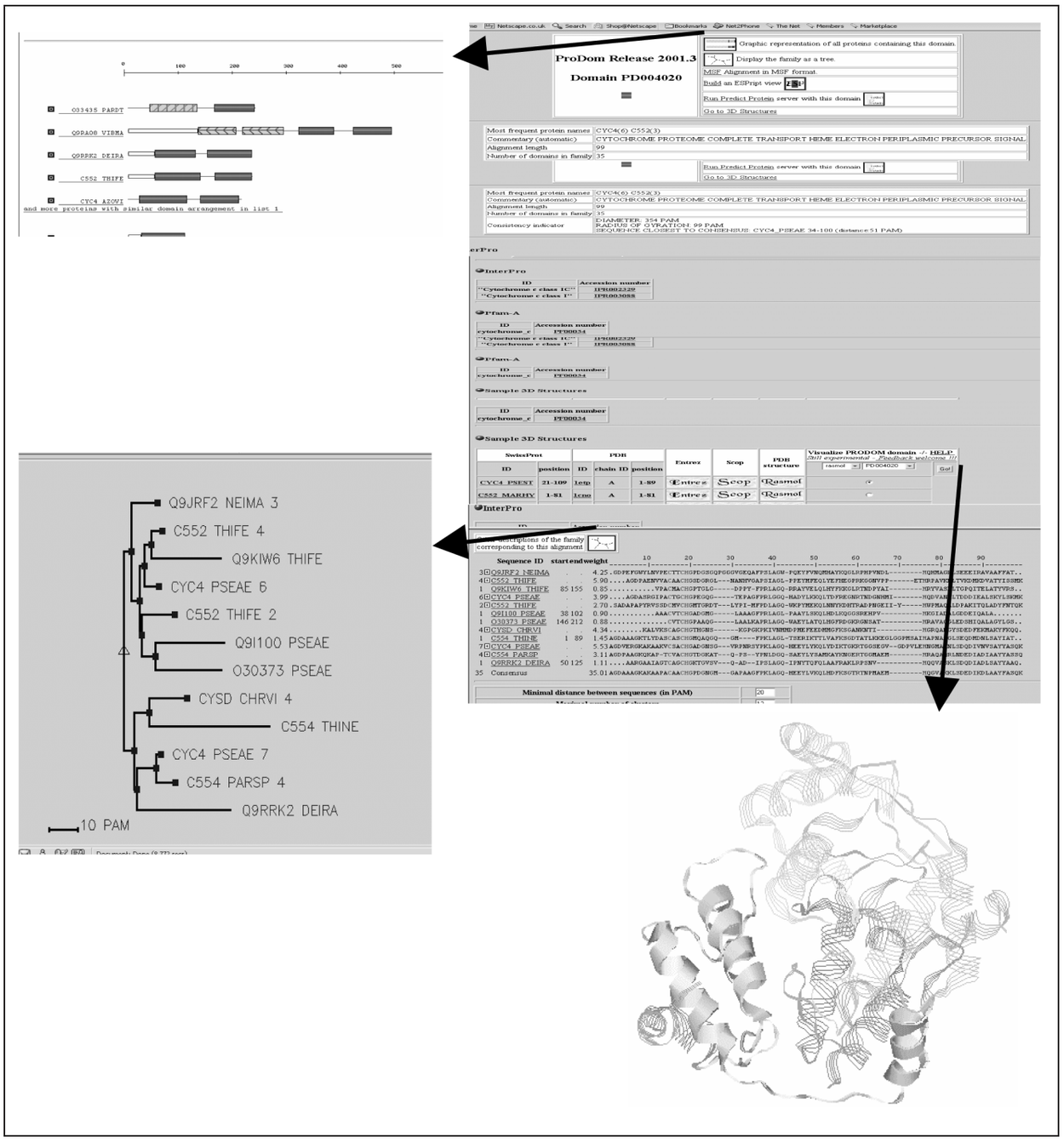

Figure 1: ProDom graphical interface. The main window provides general information on the domain family (*top right*): the entry accession number, the picture used to represent the family on the graphical output, some statistics, as well as useful links. A summarised multiple alignment of all domains in the family is also provided using the DisplayFam alignment browser.[17] Additional views are available: graphical representation of domain arrangements of proteins in the family (top left); tree view of the phylogenetic relationships between domains in the family (bottom left); visualisation of domains on 3D structures when available (bottom right)

followed by the number of sequences in the cluster. Labels and internal nodes are coloured following taxonomic class attributes (bacteria, archea, eukaryote and virus); an additional colour indicates that the cluster contains more than one different class.

General information provided on the

family includes the number of domains in the family, the most frequent protein names and an automated comment line containing the most frequent words found in the SWISS-PROT/TrEMBL description fields. The radius of gyration and the diameter of the family are provided to assess sequence homogeneity. The radius of gyration is the weighted root mean square of the distances between sequences in the family and the consensus sequence. The diameter is the maximal distance between two domains in the family. Additionally, the sequence closest to the consensus sequence is provided to help the selection of a representative member. Moreover, in ProDom-CG, a table provides for the distribution of the domain family across the various species.

Cross-references to PROSITE, InterPro, Pfam-A and PDB are displayed when relevant. Additional tools are provided to allow for a direct visualisation of a particular domain on the 3D structure. It is also possible to visualise all ProDom domains on a 3D structure using different colour codes for each domain type. The tools available for this type of analysis are Rasmol,[18] Chemscape Chime,[19] VRML97 files generated by MolScript[20] using secondary structures calculated with DSSP,[21] and static images of the structure.

ProDom is a source of protein domain targets

ProDom APPLICATIONS

ProDom is a powerful tool to analyse protein domain relationships. For instance, ProDom domain decompositions were used by Marcotte *et al.*[22] in large-scale predictions of protein–protein interactions on the basis of 'Rosetta Stone' sequence combinations. A broad spectrum of applications is made possible with ProDom. For instance ProDom was used extensively for the annotation of the *Sinorhizobium meliloti* genome.[23]

Another application of ProDom is to select candidate proteins for structural genomics projects. As determination of protein structures is difficult and as some structures are already available in the PDB database, it is relevant to construct a minimal set of protein families with unknown structures and to identify target proteins for the identification of novel structures. ProDom is well adapted to this task. Proteins were selected on the following criteria:

- no 3D structure available in the family;
- they belong to a family with at least two members (true family);
- the most distant sequences in the family are at least 10 per cent identical (family homogeneity);
- they must be single domain proteins, shorter than 500 amino acids (avoid multi-domain proteins).

The choice of single domain proteins obviates the need to engineer specific domains and should make expression and purification easier to achieve. A set of 2,587 protein targets was thus selected from ProDom99.1 for structural genomics.[1]

Contribution to InterPro

The comprehensiveness of ProDom has been used to provide InterPro with a set of novel protein families. By novel, we mean that these families have not yet been annotated by the other databases involved in InterPro. The definition of quality criteria enables good candidate families to be extracted from ProDom for further annotation and integration into InterPro.

The candidate domain families were selected on the following criteria:

- they have at least two members, to deal only with true families;
- they are not already referenced in InterPro;
- they contain at least one single domain protein from SWISS-PROT shorter than 500 amino acids, to ensure ProDom domain span is correct;

- the diameter of the families lies between 10 and 500 PAM (percentage of acceptable point mutations), thus the families are homogenous but not overly redundant.

These criteria applied to ProDom 2001.3 provided a set of 2,454 domain families that are good candidates to be documented in InterPro. In the era of large sequencing projects, these candidate families are useful to help focus annotation effort. Thus the comprehensiveness of ProDom makes it a unique tool for speeding up the identification of novel domain families to be incorporated into InterPro.

Acknowledgements
The ProDom project is supported by the 'Programme de Bio-Informatique Inter-Organismes', the 'Réseau des Génopoles' and the European Union (QLRT-1999–30517).

References

1. Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000), 'ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons', *Nucleic Acids Res.*, Vol. 28(1), pp. 267–269.
2. URL: http://www.toulouse.inra.fr/prodom.html
3. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28(1), pp. 45–48.
4. URL: http://www.toulouse.inra.fr/prodomCG.html
5. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.
6. Gouzy, J., Corpet, F. and Kahn, D. (1999), 'Whole genome protein domain analysis using a new method for domain clustering', *Comput. Chem.*, Vol. 23(3–4), pp. 333–340.
7. URL: http://www.expasy.org/
8. URL: http://www.ebi.ac.uk/proteome/
9. Wootton, J. C. and Federhen, S. (1996), 'Analysis of compositionally biased regions in sequence databases', *Methods Enzymol.* Vol. 266, pp. 554–571.
10. Bateman, A., Birney, E., Cerruti, L. *et al.* (2002), 'The Pfam protein families database', *Nucleic Acids Res.*, Vol. 30(1), pp. 276–280.
11. Altschul, S. F., Gish, W., Miller, W. *et al.* (199), 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215(3), pp. 403–410.
12. Guex, N. and Peitsch, M. C. (1997), 'SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling', *Electrophoresis*, Vol. 18(15), pp. 2714–2723.
13. Apweiler, R., Attwood, T. K., Bairoch, A. *et al.* (2001), 'The InterPro database, an integrated documentation resource for protein families, domains and functional sites', *Nucleic Acids Res.*, Vol. 29(1), pp. 37–40.
14. Falquet, L., Pagni, M., Bucher, P. *et al.* (2002), 'The PROSITE database, its status in 2002', *Nucleic Acids Res.*, Vol. 30(1), pp. 235–238.
15. Sussman, J. L., Lin, D., Jiang, J. *et al.* (1998), 'Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules', *Acta Crystallogr. D. Biol. Crystallogr.*, Vol. 54(1 (Pt 6)), pp. 1078–1084.
16. Etzold, T. and Argos, P. (1993), 'SRS – an indexing and retrieval tool for flat file data libraries', *Comput. Appl. Biosci.*, Vol. 9(1), pp. 49–57.
17. Corpet, F., Gouzy, J. and Kahn, D. (1999), 'Browsing protein families via the 'Rich Family Description' format', *Bioinformatics*, Vol. 15(12), pp. 1020–1027.
18. Bernstein, H. J. (2000), 'Recent changes to RasMol, recombining the variants', *Trends Biochem. Sci.*, Vol. 25(9), pp. 453–455.
19. URL: http://www.umass.edu/microbio/chime/index.html
20. Kraulis, J. (1991), 'MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures', *J. Appl. Crystallogr.*, Vol. 24, pp. 946–950.
21. Kabsch, W. and Sander, C. (1983), 'How good are predictions of protein secondary structure?', *FEBS Lett.*, Vol. 155(2), pp. 179–182.
22. Marcotte, E. M., Pellegrini, M., Ng, H. L. *et al.* (1999), 'Detecting protein function and protein–protein interactions from genome sequences', *Science*, Vol. 285(5428), pp. 751–753.
23. Capela, D., Barloy-Hubler, F., Gouzy, J. *et al.* (2001), 'Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021', *Proc. Natl Acad. Sci. USA*, Vol. 98(17), pp. 9877–9882.