Nucleic Acids Research, 1998, Vol. 26, No. 1 323–326

The ProDom database of protein domain families

Florence Corpet, Jérôme Gouzy[1] and Daniel Kahn[1,*]

Laboratoire de Génétique Cellulaire and [1]Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes, INRA/CNRS, BP 27, F-31326 Castanet-Tolosan Cedex, France

Received September 19, 1997; Revised and Accepted October 15, 1997

ABSTRACT

The ProDom database contains protein domain families generated from the SWISS-PROT database by automated sequence comparisons. It can be searched on the World Wide Web (http://protein.toulouse.inra.fr/prodom.html) or by E-mail (prodom@toulouse.inra.fr) to study domain arrangements within known families or new proteins. Strong emphasis has been put on the graphical user interface which allows for interactive analysis of protein homology relationships. Recent improvements to the server include: ProDom search by keyword; links to PROSITE and PDB entries; more sensitive ProDom similarity search with BLAST or WU-BLAST; alignments of query sequences with homologous ProDom domain families; and links to the SWISS-MODEL server (http://www.expasy.ch/swissmod/SWISS-MODEL.html) for homology based 3-D domain modelling where possible.

INTRODUCTION

The rapid growth of primary sequence databases makes it more and more difficult to comprehend the ever increasing diversity of known proteins. One major underlying difficulty is that many proteins exhibit a combinatorial arrangement of domains, which makes it desirable to develop databases and tools to describe proteins at an intermediary level of structure, in terms of domain arrangements. The ProDom database was designed with this explicit purpose, with particular emphasis on the user interface. Domains are detected in an automatic process that uses sequence similarities between homologous domains of SWISS-PROT sequences. ProDom 'domains' thus essentially reflect protein subsequences conserved in various proteins. For each domain family a multiple alignment and a consensus sequence are computed, as well as links to PROSITE and PDB where relevant.

We have set up a World Wide Web server (http://protein.toulouse.inra.fr/prodom.html) which provides graphical access to ProDom. It allows the user to get a schematic visualisation of all proteins sharing a given homologous domain, or all proteins sharing a homologous domain with a given protein. Hypertext links give access to multiple alignments, consensus sequences and PROSITE and PDB links for each domain family. Any query sequence can be compared against ProDom using the BLAST or the WU-BLAST algorithm with a graphical output: a possible decomposition of new protein sequences into domains is quickly visualised.

BUILDING THE DATABASE

Method

The ProDom source database is SWISS-PROT (1). All sequences are compared with BLASTP (2), providing a list of homologous segment pairs that are grouped into homologous segment sets by transitive closure using the MKDOM program (3,4). These sets are further processed automatically in order to infer domain boundaries, either at the ends of bona fide sequences, at the ends of tandem repeats, or where sequence shuffling is detected. Multiple alignments are then systematically generated for all families using the MultAlin program (5), and a consensus sequence is calculated as the best weighted average sequence for each multiple alignment. Details of the method are provided elsewhere (3,4).

ProDom 'domains' are inferred on the basis of conserved subsequences as found in various proteins. Such a conservation corresponds frequently, though not always, to genuine structural domains: therefore domain boundaries should be treated with caution. For some domain families experts have been asked to correct domain boundaries on the basis of both sequence and structural information. This expertise will complement the automated process and improve the quality of ProDom domain families.

Database format

ProDom is built as two text files, 'prodom.mul' and 'prodom'. Each entry is a domain family with an automatically generated comment and a multiple domain alignment in the 'prodom.mul' file, or a consensus sequence in the 'prodom' file. We also provide a tool (FETCHDOM) to retrieve the domain decomposition of any protein that is present in ProDom, or to fetch multiple alignments of ProDom domain families.

Content of the current release

Release 34.1 of ProDom (June 1997) contains 53 597 domain families. These are sorted by decreasing number of protein sequences in the families. Each non-fragmentary sequence from SWISS-PROT release 34 is treated in ProDom 34.1. More recent sequences (up to 21 May 1997) and fragmentary ones have been added if they share similarity with a ProDom domain family. The database requires ~56 Mb of disk storage space. The present

*To whom correspondence should be addressed. Tel: +33 561 28 53 29; Fax: +33 561 28 50 61; Email: dkahn@toulouse.inra.fr

distribution frequency is one main release for each SWISS-PROT release (with the same release number).

SEARCHING ProDom WITH BLAST TOOLS

ProDom is a useful tool to help determine domain arrangements of protein sequences. If the sequence is present in the corresponding SWISS-PROT release, the FETCHDOM program directly provides a proposed domain arrangement. If the sequence is new, a ProDom similarity search should be performed using, for instance, the BLASTP (2,6) or the WU-BLAST (7) program and ProDom consensus sequences as a target database: this search is less sensitive than a search of the primary sequence database but it is faster and redundant information is avoided. A more sensitive search can be performed using ProDom as a multiple domain database: all domain sequences present in ProDom domain families are searched for similarity with the query sequence, and results are filtered to retain only the best hit for each domain family at any given position within the query. This search is as fast and sensitive as a direct search of the primary sequence database, but in addition the filtered output directly provides a possible domain arrangement. Note that the philosophy underlying homology search with ProDom differs from the one proposed with BLOCKS (8) and PRINTS (9) in that ProDom multiple alignments are exhaustively searched for domains most similar to the query, whereas BLOCKS and PRINTS use position-specific scoring matrices and patterns, respectively.

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://protein.toulouse.inra.fr/prodom.html . The ProDom server displays multiple alignments of ProDom domains and relevant links, it provides a graphical representation of protein domain arrangements, and it allows for ProDom similarity searches with graphical outputs.

Multiple alignments of ProDom domain families (Fig. 1)

ProDom domain families can be accessed through ProDom numerical IDs, through keywords, or through relevant PROSITE (10) or PDB (11) entries. PROSITE and PDB links are calculated

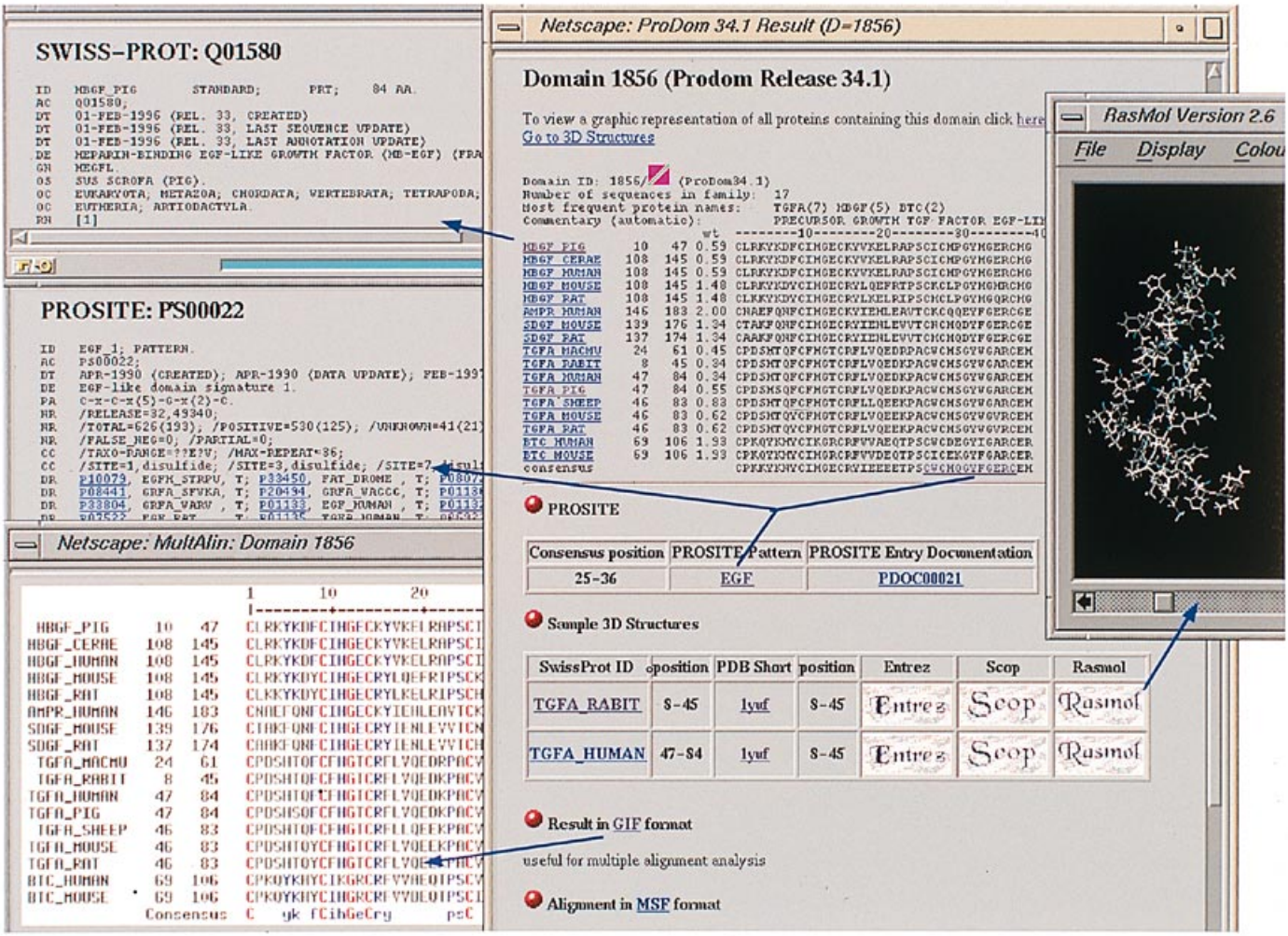

Figure 1. Example of ProDom WWW server usage. This page displays ProDom 34.1 domain family 1856, corresponding to EGF-like growth factors. Clicking on a protein name displays the corresponding SWISS-PROT entry on the ExPASy server (top left, ref. 17). The PROSITE pattern in the consensus sequence is hypertext-linked to the PROSITE entry on ExPASy (middle left). Clicking on the RasMol field loads the structure into the RasMol molecular viewer (right, ref. 15). The GIF image on the lower left highlights conserved positions in this family as calculated with MultAlin (5).

by direct comparison with ProDom domains using the LASSAP program (12). A ProDom domain family is displayed as a multiple alignment of homologous domains and is linked to several other databases (Fig. 1 and Table 1). The header includes the ProDom family number with its pattern for graphical representation, the number of sequences in the family, the most frequent protein names and a comment line containing the most frequent words in the SWISS-PROT description field (DE).

Each protein domain is defined by a SWISS-PROT ID, by positions for domain start and end, by its weight in the multiple alignment and by sequence. The ID is hypertext linked to the corresponding SWISS-PROT entry. The last line in the alignment provides a consensus sequence for the domain family, with highlighted PROSITE patterns if relevant.

The alignment is followed by a list of appropriate PROSITE patterns and PDB structures with hypertext links. When a PDB structure is found to correspond with a domain of the family, it is linked to PDB (11), Entrez (13) and SCOP (14) entries; a click on the 'RasMol' field results in the structure being automatically loaded into the RasMol molecular viewer (15). It is also possible to retrieve the alignment in MSF format (Multiple Sequence File) to print it or to import it into a multiple alignment software. The multiple alignment can also be viewed as a GIF image with conserved positions highlighted, which facilitates its interpretation (Fig. 1).

Table 1. ProDom links

Link	Source	URL	Ref.
Sequence ID	SWISSPROT	http://www.expasy.ch	17
Sequence pattern	PROSITE	http://www.expasy.ch	17
3D structure	PDB	http://www.pdb.bnl.gov	11
3D structure	NCBI Entrez	http://www3.ncbi.nlm.nih.gov	13
3D structure	SCOP	http://scop.mrc-lmb.cam.ac.uk	14

Graphical view of domain arrangements in protein families (Fig. 2)

A graphical view presents domain arrangements for a given protein, for proteins sharing a given ProDom domain, or for all proteins sharing homology with a given SWISS-PROT entry (Fig. 2). Each protein is shown on a single line, starting with its name, followed by the schematic display of its domains as boxes drawn to scale. Each domain family has a unique representation among 12 720 different combinations of 16×15 colours and 53 motifs; small domain families starting with ProDom family 12 721 upwards are represented by a numbered box, while one-membered families are shown with a narrow, empty box. Protein names are hypertext-linked to SWISS-PROT, and domain schemes are hypertext-linked to corresponding multiple alignments. Proteins are sorted by domain composition so that proteins with similar domain composition appear clustered. A simplification option generates only one display for each different type of domain arrangement, which appears useful to interpret large protein families. This graphical output can be saved as a GIF image for printing or for exporting into an image or a text editor.

Exploiting ProDom similarities on the World Wide Web

ProDom can be searched for similarity with a query sequence using BLAST tools (BLASTP, BLASTX or WU-BLAST). If the query sequence shares homology with at least one ProDom family, BLAST results are followed by a graphical representation of its proposed domain arrangement, similar to displays in Figure

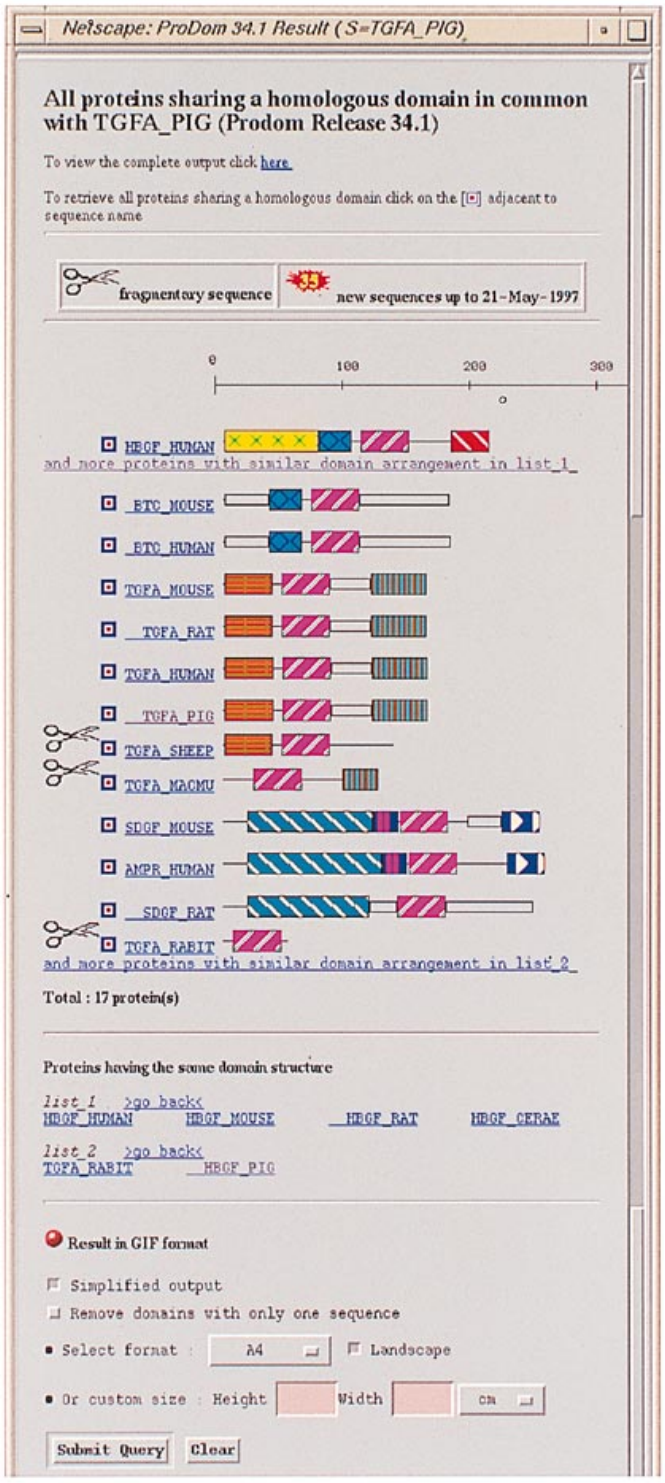

Figure 2. Graphical representation of protein domain arrangements. Automated schematic display of proteins sharing homology with pig TGF-α precursor (TGFA_PIG). Each domain type is represented by a specific colour pattern which is hypertext-linked to the corresponding multiple alignment (see Fig. 1). Protein names are linked to the corresponding SWISS-PROT entries on the ExPASy server (17). Clicking the coloured square to the left of a protein name displays all proteins sharing homology with this protein.

2. Each target ProDom domain can be further exploited, either to align the query with the ProDom domain family using MultAlin (5), or to generate 3-D models of domains on the basis of homology using the Swiss-Model server (16), where applicable.

ACCESS

Anonymous FTP site

ftp://ftp.toulouse.inra.fr/pub/prodom

Email server

prodom@toulouse.inra.fr . Send the word HELP as the only word in the message body

WWW server

http://protein.toulouse.inra.fr/prodom.html

ACKNOWLEDGEMENTS

We wish to thank Claude Chevalet, Amos Bairoch, Manuel Peitsch, Jean-Jacques Codani and Eric Glémet for stimulating discussions, and Eric Claverie for programming assistance. The ProDom project was supported by the Groupement de Recherches et d'Etudes sur les Génomes (contract 107/94) and by the Ministère de l'Education Nationale, de la Recherche et de la Technologie (ACC-SV13).

REFERENCES

1 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36. [See also this issue *Nucleic Acids Res.* (1998) **26**, 38–42.]
2 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
3 Sonnhammer,E.L.L. and Kahn,D. (1994) *Protein Sci.*, **3**, 482–492.
4 Gouzy,J., Eugène,P., Greene,E.A., Kahn,D. and Corpet,F. (1997) *Comput. Appl. Biosci.*, **13**, in press.
5 Corpet,F. (1988) *Nucleic Acids Res.*, **16**, 10881–10890. http://www.toulouse.inra.fr/multalin.html
6 Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
7 Gish,W. (1997) WU-BLAST, http://blast.wustl.edu/
8 Henikoff,J.G., Pietrokovski,S. and Henikoff,S. (1997) *Nucleic Acids Res.*, **25**, 222–225. See also this issue *Nucleic Acids Res.* (1998) **26**, 309–312.]
9 Atwood,T.K., Beck,M.E., Bleasby,A.J., Degtyarenko,K., Michie,A.D. and Parry-Smith,D.J. (1997) *Nucleic Acids Res.*, **25**, 212–216. [See also this issue *Nucleic Acids Res.* (1998) **26**, 168–169.]
10 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
11 Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds) *Crystallographic Databases-Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.
12 Glémet,E. and Codani,J.J. (1997) *Comput. Appl. Biosci.*, **13**, 137–143.
13 Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) *Methods Enzymol.*, **266**, 141–162.
14 Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res.*, **25**, 236–239.
15 Sayle,R.A. and Milnerwhite,E.J. (1995) *Trends Biochem. Sci.*, **20**, 374–376.
16 Peitsch,M.C. (1996) *Biochem. Soc. Trans*, **24**, 274–279. http://www.expasy.ch/swissmod/SWISS-MODEL.html
17 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, **19**, 258–260.