276–280 Nucleic Acids Research, 2002, Vol. 30, No. 1

The Pfam Protein Families Database

Alex Bateman*, Ewan Birney[1], Lorenzo Cerruti[2], Richard Durbin, Laurence Etwiller[1], Sean R. Eddy[3], Sam Griffiths-Jones, Kevin L. Howe, Mhairi Marshall and Erik L. L. Sonnhammer[4]

Wellcome Trust Sanger Institute and [1]The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [2]SIB, ISREC, 155, ch. des Boveresses, CH-1066 Epalinges s/Lausanne, Switzerland, [3]Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA and [4]Center for Genomics and Bioinformatics, Karolinska Institutet, S-171 77 Stockholm, Sweden

Received September 19, 2001; Accepted September 25, 2001

ABSTRACT

Pfam is a large collection of protein multiple sequence alignments and profile hidden Markov models. Pfam is available on the World Wide Web in the UK at http://www.sanger.ac.uk/Software/Pfam/, in Sweden at http://www.cgb.ki.se/Pfam/, in France at http://pfam.jouy.inra.fr/ and in the US at http://pfam.wustl.edu/. The latest version (6.6) of Pfam contains 3071 families, which match 69% of proteins in SWISS-PROT 39 and TrEMBL 14. Structural data, where available, have been utilised to ensure that Pfam families correspond with structural domains, and to improve domain-based annotation. Predictions of non-domain regions are now also included. In addition to secondary structure, Pfam multiple sequence alignments now contain active site residue mark-up. New search tools, including taxonomy search and domain query, greatly add to the functionality and usability of the Pfam resource.

INTRODUCTION

Pfam is a manually curated collection of protein families available via the web and in flat file form (1). Genome projects, including both the human and fly, have used Pfam extensively for large scale functional annotation of genomic data (2,3). The multiple sequence alignments around which Pfam families are built are important tools for understanding protein structure and function, and form the basis for techniques such as secondary structure prediction, fold recognition, phylogenetic analysis and mutation design. The latest version of Pfam (6.6) contains 3071 families that have matches to 69% of sequences and cover 49% of residues in the sequence database.

Each curated family in Pfam is represented by a seed and full alignment. The seed contains representative members of the family, while the full alignment contains all members of the family as detected with a profile hidden Markov model (HMM) constructed from the seed alignment using the HMMER2 software (http://hmmer.wustl.edu/). Full alignments can be large with the top 20 families now each containing over 2500 sequences. The majority of known protein sequences come from just a few thousand protein families. However, in an effort to be comprehensive, the curated families in Pfam-A are augmented by Pfam-B, an automatically generated supplement derived from the PRODOM database (4).

Pfam is available at four locations around the world, each providing a core set of functionality for accessing each family. Pfam is available in Europe on the World Wide Web at http://www.sanger.ac.uk/Software/Pfam/ (UK), http://www.cgb.ki.se/Pfam/ (Sweden) and http://pfam.jouy.inra.fr/ (France), and in the US at http://pfam.wustl.edu/. Documentation on the content and use of Pfam is available via the web. The web sites described above contain documentation on Pfam alignments, mark-up and family annotation. The alignments in Pfam are in Stockholm format, which is described in detail at http://www.cgb.ki.se/cgb/groups/sonnhammer/Stockholm.html, and the HMMER software is documented at http://hmmer.wustl.edu/.

Pfam ANNOTATION

Pfam contains annotation of each family in the form of textual descriptions, links to other resources and literature references. Pfam is a member of the InterPro consortium (5) and has, like the other member databases, contributed annotation and families to the InterPro project. InterPro aims to provide an integrated view of the diverse protein family databases and one of its strengths is that a comprehensive set of annotations has been created through the merging of information from each member. The InterPro annotation is often more comprehensive than the Pfam annotation, and so is imported into the Pfam web pages and can be accessed by following links to InterPro. Further improvements in the quality of Pfam family annotation are outlined in the following sections.

STRUCTURAL DATA IMPROVES DOMAIN BOUNDARIES AND ANNOTATION

Domains are the structural and functional building blocks of proteins, and so where the data are available, structural information has been used to ensure that Pfam families correspond to single structural domains. The domain boundaries used are currently those defined by the SCOP database (6) and a new web-based tool allows direct cross-linking from domains on the SCOP web site to the corresponding Pfam families. This matching of

*To whom correspondence should be addressed. Tel: +44 1223 494950; Fax: +44 1223 494919; Email: agb@sanger.ac.uk

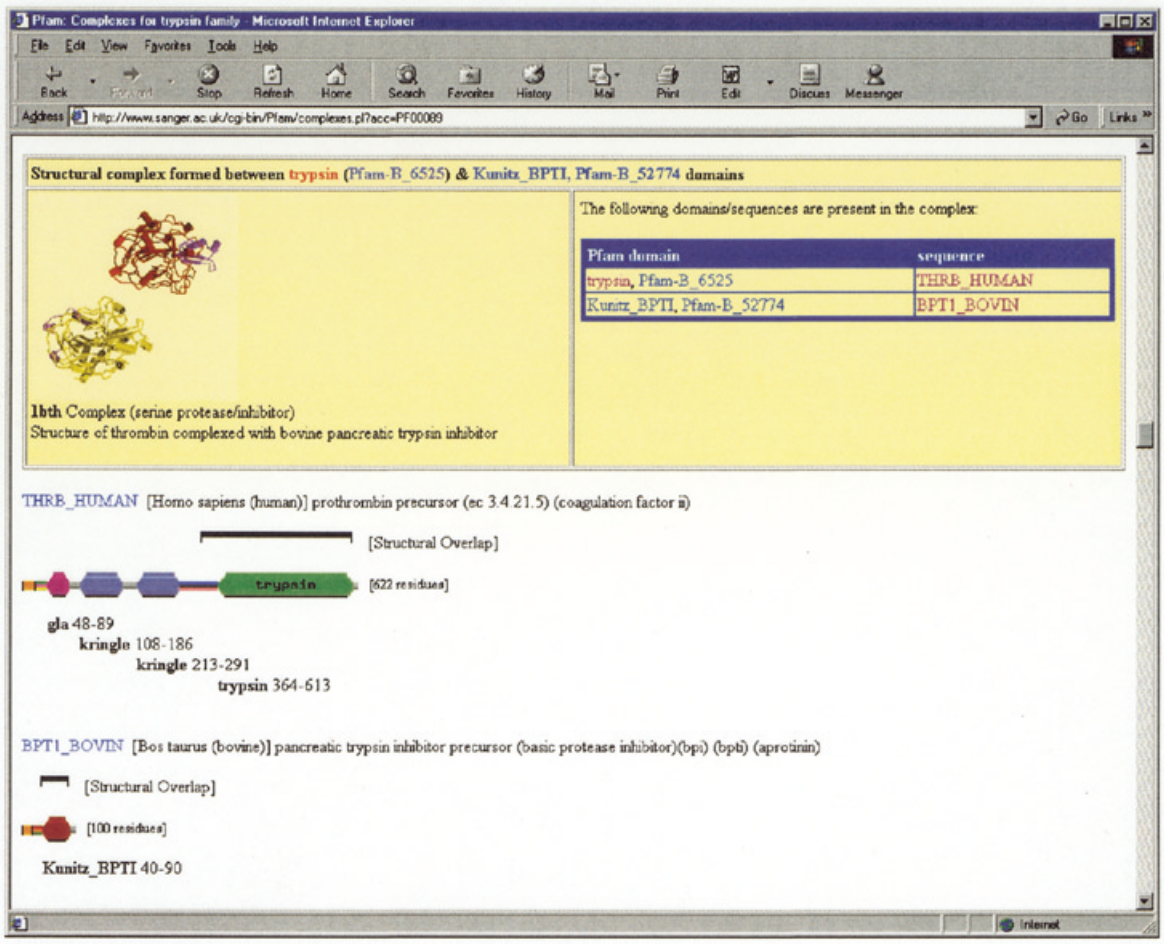

Figure 1. The web view of the structural complex of thrombin with bovine pancreatic trypsin inhibitor. The web page shows the Pfam domain structure of the two proteins in the complex; the regions of the proteins represented in the PDB structure are highlighted by the square bracket above each protein. The protein structure picture has been kindly provided by the PDBsum resource (17).

families and domains enables enhanced understanding of the function of multi-domain proteins. For example, the OTCace family contains two related enzymes, aspartate carbamoyl transferase and ornithine carbamoyl transferase. Structural data have shown that these approximately 300 amino acid proteins consist of two structurally similar domains, the N-terminal domain binds carbamoyl phosphate and the C-terminal domain binds aspartate/ornithine. Each domain is now represented by a well annotated Pfam family. These two activities are also found at the C-terminus of glutamate-dependent carbamoyl phosphate synthase, a large multi-domain protein whose Pfam-based annotation also now clearly describes ATP-binding and oligomerisation domains among others. In some cases, the action of chopping a single family into two or more structural domains also enables the elucidation of increased instances of the particular domain, sometimes in novel protein contexts. For example, the cytochrome reductase family has been split into its constituent FAD and NAD binding domains, which are found more generally in a number of oxidoreductases. In all, approximately 300 Pfam families have been split into two or more domains, with the domain boundaries of many more refined to better match the available structural data.

To help clarify these changes, we have introduced a new annotation field 'type' (TP). At present, a Pfam family can be classified as a family, domain, repeat or motif. Family type is the default class which simply states that the members are related. A domain is defined as an autonomous structural unit, or a reusable sequence unit that may be found in multiple protein contexts. In contrast, a repeat is not usually stable in isolation; rather, multiple tandem repeats are usually required to form a globular domain or extended structure. Motifs generally describe shorter sequence units found outside globular domains. Pfam release 6.6 contains 2032 families, 980 domains, 54 repeats and 5 motifs.

Protein–protein interaction data provide an important source of information for studying protein families and their cellular roles. We have used data from known three-dimensional protein complexes in the PDB (7) to infer protein–protein interactions between Pfam domains. NCBI BLAST2 (8) was used to find the correspondence between known structures (PDB chains) and sequences in the sequence databases. These data were used to analyse structural complexes between Pfam domains. An example of the graphical interface to this data provided on the UK web site is shown in Figure 1.

NON-DOMAIN ANNOTATION

Although Pfam attempts to classify proteins into domains where possible, some regions of proteins are not expected to form stable globular domains. These include regions of biased amino acid composition [termed low sequence complexity regions (9)], coiled-coils, transmembrane regions and signal peptides. However, these regions are of considerable interest and so predictions are reported on the UK web site. These predictions are pre-computed over the sequence database by the following third party programs: TMHMM (10) (transmembrane regions), SignalP (11) (signal peptide regions), ncoils (12) (coiled-coil regions) and SEG (9) (low complexity regions). The regions and associated scores are stored in the Pfam relational database (see below).

Non-Pfam regions require a different web-based graphical representation. In contrast with Pfam-A and Pfam-B regions, non-Pfam regions can overlap with each other and with Pfam regions. Overlapping regions are resolved for the graphical display by a hierarchical approach. The default hierarchy (signal peptide > Pfam-A > transmembrane > Pfam-B > low complexity > coiled-coil) is easily changed by the user, to enable the visualisation of different features.

ACTIVE SITE INFORMATION

When viewing multiple sequence alignments it is useful to be able to see the sequence location of features of interest. Structural features have previously been incorporated into Pfam alignments, and more recently we have included active site residues. We have used the ACT_SITE feature table lines from SWISS-PROT as the data source. The alignments with added mark-up clearly show whether active site residues are conserved in all members of a family. The most frequent active site residues in SWISS-PROT are C, D, E, H, K, R, S and Y (Fig. 2). Other non-polar residues do occur, but at a much lower frequency. The glycine residues are found to be reactive bonds in trypsin inhibitors, which are not true active site residues. We can gain information about the nature of active site residue substitutions by examining the distribution of amino acids within columns that correspond to active site residues, also shown in Figure 2.

TAXONOMY

The 'taxonomy search' tool (UK web site), allows the user to find Pfam entries specific to a group of organisms using a taxonomy query language. Complex queries are possible by using logical operators (AND, OR, NOT) and parentheses. The taxonomic information for each protein match is extracted from the SWISS-PROT/TrEMBL databases (13).

One use of this tool is to aid identification of putative drug targets. For example, as part of a screen for possible drug targets unique to the malaria parasite, one might want to identify all Pfam domains present in *Plasmodium falciparum* but not in the vertebrate host. The taxonomic query '*Plasmodium falciparum* AND NOT *Vertebrata*' returns 26 Pfam domains, 10 of which have already been postulated as drug targets against *P.falciparum.*

Using the taxonomy search software we have evaluated how the four major kingdoms (eukaryota, bacteria, archaea and viruses) are represented in the Pfam collection. The results are shown in Table 1. The data clearly show a bias towards eukaryotes, with over two-thirds of Pfam families containing a eukaryotic representative. A large number of these families are specific to eukaryotes, perhaps reflecting the invention of novel proteins in this kingdom, or possibly simply the biases in known protein sequence databases. Archaeabacterial proteins occur in just over one-third of Pfam families, reflecting the relatively small number of sequences, and only 49 families are restricted solely to archaea. Viral sequences are found in 571 Pfam families.

Table 1. The taxonomic distribution of Pfam families

	Eukaryota	Bacteria	Archaea	Viruses
Total number of Pfam entries	2155	1737	1030	571
Number of unique Pfam entries	824	356	49	225

The first row shows the number of Pfam entries in each kingdom. The second row shows the number of Pfam entries specific for each kingdom.

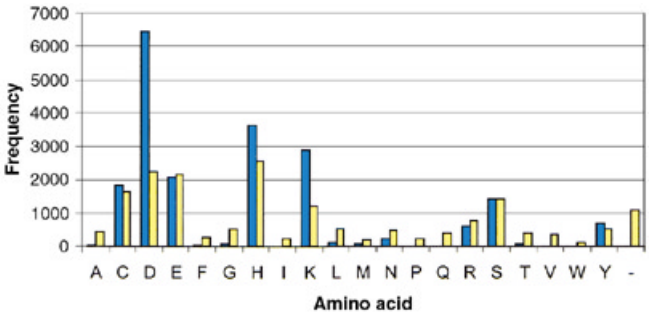

Figure 2. Distribution of active site residues in SWISS-PROT and Pfam seed alignments. Histogram showing the frequency of amino acids found in active sites from SWISS-PROT (blue) and the frequency of amino acids aligned to active site residues (yellow).

ANALYSIS OF DOMAIN ARCHITECTURE EVOLUTION

Pfam is an excellent resource for studying the evolution of domain architecture in proteins. To make such analyses possible even by the casual user, we have equipped the Pfam web servers with a number of tools. NIFAS allows visual inspection of domain architectures in an evolutionary tree, and has been described previously (14). Two new tools have been developed and are described below.

Similar domain organisation

One of Pfam's main uses is to return the domain organisation of a protein of interest. This will inform the user which domain families it belongs to, as a valuable complement to traditional similarity searching. Another way to analyse sequence similarity is to look for proteins that share the same overall domain organisation, although these may not be the most sequence-similar proteins. This search functionality is now available on the Sweden web server. There is no obviously correct way to assign a score to similarity in domain organisation, so the proteins are heuristically ranked by the number of domains in common, from identical domain architectures, through re-ordered

combinations, to smaller numbers of common domains. All proteins are listed as schematic graphics of the domain architectures, and their functional description may be shown.

Domain query tool

To ask other questions about the presence or absence of certain domain architecture features, a general purpose tool has been installed on the Sweden web server. A menu-driven interface allows the user to specify a query consisting of a set of Pfam domains, with or without ordering or gap constraints, similar to regular expressions. The user can retrieve a list of all proteins with a certain domain combination motif, e.g. all proteins with an Fz, a kringle and a protein kinase domain. It is also possible to perform negative queries, e.g. retrieve all proteins with an Fz and protein kinase domain that do not have a kringle domain in between. The results are ordered with the same graphical schematics as the previous tool.

CHANGES TO Pfam SEARCHING

Previously, Pfam families were based on hits to either global (ls) or fragment (fs) model HMMs. The latter does not penalise long gaps, thus allowing partial matches to the HMM to be found. The decision on which model to use for a specific family was largely arbitrary, but influenced by membership criteria. For example, families such as the REV family of viral anti-repression trans-activator proteins contain many proteins annotated as fragments in SWISS-PROT/TrEMBL, many of which are missed by an HMM search using the ls model. However, with increased emphasis on domain families, it seems more intuitive to base families on the global model to match whole domains where possible. To solve this problem, we have recently rebuilt all Pfam families using both ls and fs model HMMs, and calculated membership from the global model, but adding hits to the fs model which were not considered significant matches to the ls model. This approach has lead to a substantial increase in the number of protein matches to many families and also in coverage at the residue level.

A number of small format changes have been necessary as a result of this global change. Each model requires separate gathering thresholds (GA), and each has associated trusted (TC) and noise (NC) cutoffs. These numbers are all specified in the family annotation. Web-based searches now provide the option to search using global or fragment models.

As well as providing searches of the Pfam HMMs, the UK web site now offers the option to search against SMART (15) and TIGRFAM (16) HMM collections. Pfam, SMART and TIGRFAM domains may overlap so a tool has been provided to allow the display priority to be altered.

THE Pfam RELATIONAL DATABASE

The traditional implementation of Pfam, as a directory-structure of text files, one directory for each family, has proved to be extremely stable and robust. The revision control system has been used to provide an update history for the database, and allows us to re-create any release of the database. However, the text file based implementation is not well suited to performing cross-family queries on the live database, for example querying for all Pfam domains lying on a specific protein sequence. This kind of query is performed extensively in Pfam to enforce one of the key quality controls, the overlap criterion, which states that no residue of any protein can belong to more than one family. In the past, the only way to perform queries of this nature has been to search through the alignment files for every family, looking for occurrences of the sequence of interest. This is slow, and becomes slower as the number of families increases.

PfamRDB is a mySQL relational database consisting of approximately 10 tables adhering to a tight relational schema. It is updated in-phase with the live Pfam database to maintain absolute consistency. Some data (for example HMMs and alignments) are not currently stored in PfamRDB. PfamRDB also contains additional information, for example non-domain mark-up of sequences (low-complexity, coiled-coil, trans-membrane and signal peptide, as described above), and also projections of Pfam domains onto solved structures in the PDB.

ACKNOWLEDGEMENTS

We are grateful to the many people who have submitted data to Pfam. In particular, William Mifsud, Matthew Bashton and Nina Mian have added many of the new families in Pfam. We thank Christian Storm and Volker Hollich for implementing the NIFAS and Domain Query tools. We are also grateful to Roman Laskowski for allowing us to incorporate protein structure pictures from the PDBsum resource (17) and to Rob Finn for helpful comments.

REFERENCES

1. Sonnhammer,E.L.L., Eddy,S.R. and Durbin,R. (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
2. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
3. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. 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.*, **28**, 267–269.
5. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2000) InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.
6. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
7. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macro-molecular structure. *J. Mol. Biol.*, **112**, 535–542.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Wootton,J.C. (1994) Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.*, **4**, 413–421.
10. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
11. Nielsen,H., Brunak,S. and von Heijne,G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.

12. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
13. Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.
14. Storm,C.E. and Sonnhammer,E.L.L. (2001) NIFAS: visual analysis of domain evolution in proteins. *Bioinformatics*, **17**, 343–348.
15. Ponting,C.P., Schultz,J., Milpetz,F. and Bork,P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 242–244.
16. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.
17. Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.