D138–D141 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh121

The Pfam protein families database

Alex Bateman*, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich[1], Sam Griffiths-Jones, Ajay Khanna[2], Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer[1], David J. Studholme, Corin Yeats and Sean R. Eddy[2]

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [1]Center for Genomics and Bioinformatics, Karolinska Institutet, S-171 77 Stockholm, Sweden and [2]Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

Received September 17, 2003; Accepted October 20, 2003

ABSTRACT

Pfam is a large collection of protein families and domains. Over the past 2 years the number of families in Pfam has doubled and now stands at 6190 (version 10.0). Methodology improvements for searching the Pfam collection locally as well as via the web are described. Other recent innovations include modelling of discontinuous domains allowing Pfam domain definitions to be closer to those found in structure databases. Pfam is available on the web in the UK (http://www.sanger.ac.uk/Software/Pfam/), the USA (http://pfam.wustl.edu/), France (http://pfam.jouy.inra.fr/) and Sweden (http://Pfam.cgb.ki.se/).

INTRODUCTION

Pfam is a comprehensive collection of protein domains and families, with a range of well-established uses including genome annotation. Each family in Pfam is represented by two multiple sequence alignments and two profile-Hidden Markov Models (profile-HMMs). The functionality, use and philosophy of Pfam have been discussed in previous publications (1,2) and will not be discussed at length here. In the following sections we describe the most important improvements that have been made to the database in the past 2 years.

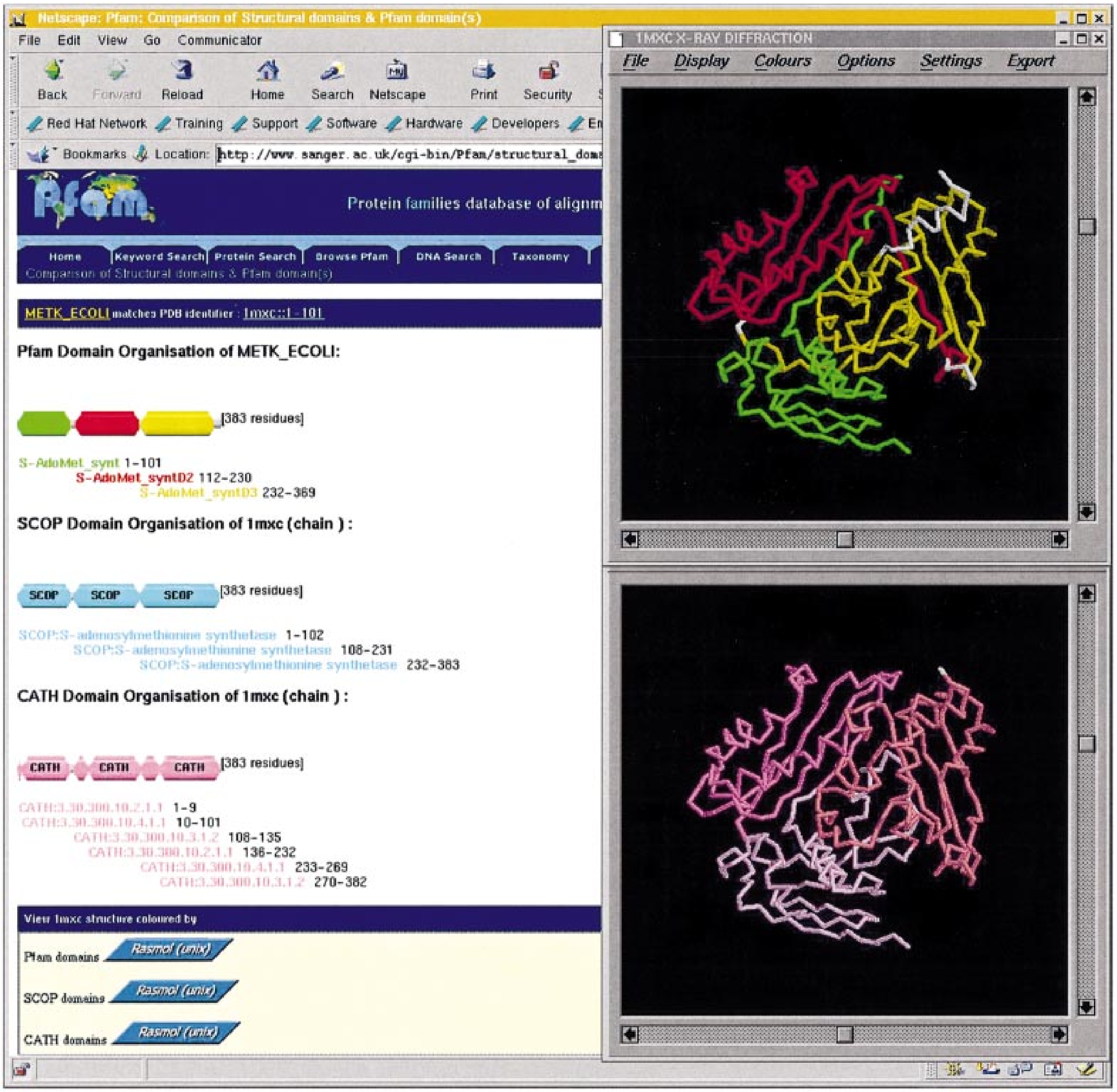
GROWTH OF PFAM

As of release 10.0, Pfam contains 6190 Pfam families. Pfam families match 75% of protein sequences in Swiss-Prot and TrEMBL (3) (and 53% of all residues). This compares with 3071 families and 69% coverage at release 6.6, 2 years ago (2). For those protein sequences that do not belong to any Pfam family, we derive automatically generated Pfam-B families. The Pfam-B families are derived from ProDom (4), a comprehensive set of protein domain families automatically generated from the Swiss-Prot and TrEMBL sequence databases. Many multi-domain protein sequences contain (non-overlapping) matches to both Pfam and Pfam-B families. The combination of Pfam and Pfam-B covers 82% of protein sequences in Swiss-Prot and TrEMBL. Every Pfam release is now built on the latest versions of Swiss-Prot and TrEMBL minimizing problems with out-of-date sequence entries.

Pfam has two large series of functionally uncharacterized families, known as Domains of Unknown Function (DUFs) and Uncharacterized Protein Families (UPFs). DUFs are families that have been created by Pfam whereas UPFs are those created by Swiss-Prot and added to Pfam. Tracking the number of DUFs and UPFs gives us some idea of how many families in Pfam are uncharacterized and how this number has changed over time. As of Pfam 6.6 there were 272 DUF and UPF families out of a total of 3071. Pfam release 10.0 contains 1004 DUF and UPF families out of 6190. Eighty-nine of the original 272 have been annotated. Of these, 20 were merged with other families and 69 were annotated with a function. Hence, on average, around 37 new domains of unknown function are added to Pfam every month and six are annotated with a function. The proportion of DUF and UPF families in the Pfam database has increased from 9% to 16%. However, the number of DUF and UPF matches to Swiss-Prot compared with the number of hits from annotated families has increased only marginally over this period. This reflects the increasing tendency for completely undescribed families to be small and specific to a few genomes.

To cope with the increased computational burden that doubling the number of families and therefore profile-HMMs creates we have two innovations to aid users. First, HMMER, the freely available profile-HMM software, used to construct and search Pfam, has been upgraded to a version 2.3 lineage (the current release is 2.3.1, see http://hmmer.wustl.edu). The principal difference between HMMER 2.3 and previous versions is a 2- to 3-fold speed-up on most platforms because of performance optimizations, and ~8-fold on Mac OS/X thanks to code contributed by Erik Lindahl at Stanford University. These performance improvements accelerate Pfam searches, and help keep pace with the growing size of the database. Secondly, users can now carry out batch searches of up to 1000 sequences at a time on the UK web server, with results being returned by email. This service means that users with moderate requirements do not need to install a local copy of Pfam and HMMER.

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


Netscape: Pfam: Comparison of Structural domains & Pfam domain(s)
File Edit View Go Communicator
Back Forward Reload Home Search Netscape Print Security
Red Hat Network Training Support Software Hardware Developers
Bookmarks Location: http://www.sanger.ac.uk/cgi-bin/Pfam/structural_dom
Pfam
Protein families database of alignm
Home Keyword Search Protein Search Browse Pfam DNA Search Taxonomy
Comparison of Structural domains & Pfam domain(s)
1MXC X-RAY DIFFRACTION
File Display Colours Options Settings Export
SCOP SCOP SCOP
CATH CATH CATH
Rasmol (unix)
Rasmol (unix)
Rasmol (unix)

METK_ECOLI matches PDB identifier : 1mxc::1-101

Pfam Domain Organisation of METK_ECOLI:

[383 residues]

S-AdoMet_synt 1-101
S-AdoMet_syntD2 112-230
S-AdoMet_syntD3 232-369

SCOP Domain Organisation of 1mxc (chain) :

[383 residues]

SCOP:S-adenosylmethionine synthetase 1-102
SCOP:S-adenosylmethionine synthetase 108-231
SCOP:S-adenosylmethionine synthetase 232-383

CATH Domain Organisation of 1mxc (chain) :

[383 residues]

CATH:3.30.300.10.2.1.1 1-9
CATH:3.30.300.10.4.1.1 10-101
CATH:3.30.300.10.3.1.2 108-135
CATH:3.30.300.10.2.1.1 136-232
CATH:3.30.300.10.4.1.1 233-269
CATH:3.30.300.10.3.1.2 270-382

View 1mxc structure coloured by

Pfam domains
SCOP domains
CATH domains

Figure 1. Comparison of Pfam, SCOP and CATH domain definitions for *S*-adenosylmethionine synthetase. The definitions of SCOP and Pfam are very similar, but rather different from the stricter structural definition of CATH. The definitions can be compared at the level of structure for each database. The two Rasmol windows on the right show the Pfam definition above the CATH definition.

IMPROVED MODELLING OF DOMAINS IN PFAM

Pfam aims to be a database of accurate protein domain definitions. In the past 2 years we have split many existing families into structural domains. Collaborations with the structural protein domain databases SCOP (5) and CATH (6) have enabled the development of a domain comparison tool to aid this process (see Fig. 1). This tool allows the relationship of the structural domain architecture defined by CATH and SCOP to be compared with each other and Pfam. Such comparisons help ensure consistency of domain definitions in the three independent databases and facilitate their linking at a common level. The domain comparison tool uses web services to retrieve the domain boundaries from CATH and SCOP on user request. These web services are maintained by each database ensuring up to date data and minimizing discrepancies between database versions. Web users can view structures marked up according to domain boundaries with Rasmol or RasWin (7).

An area of significant difference in domain definitions between Pfam and the structural databases is due to discontinuous domains defined in SCOP and CATH. A discontinuous domain is one where the linear sequence of the domain is interrupted by another inserted domain. For example, the IMPDH domain (Pfam accession PF00478) is found as a continuous domain in the GuaB protein, and with a pair of inserted CBS domains (PF00571) in inosine monophosphate dehydrogenases. Currently, there are 29 examples of discontinuous domains in Pfam. Modelling of discontinuous domains is achieved by forcing the profile-HMM to allow the inserted domain as a long insertion. We do this by using the –hand option in the HMMER software along with a # = GC RF line. For clarity the sequence of the inserted domain is also masked with X characters and the presence of a nested domain is indicated in the flat files by an NE tag. This improved modelling has allowed more accurate description of discontinuous domains as we see them in protein structures, and leads to increased search sensitivity.

IMPROVED FAMILY MEMBERSHIP

To provide users with a more unified view of protein domains, we have implemented two web-based innovations. First, the SMART database of protein domains (8) and Pfam each contain many entries that are not available in the other, and in other cases the family memberships differ markedly. The two databases exchange lists of matches and present these matches on the Pfam and SMART websites.

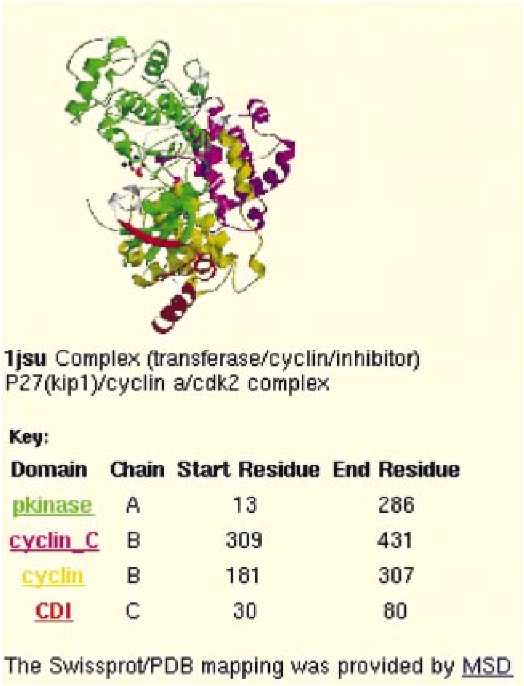
Secondly, we have applied language modelling techniques from speech recognition to identify weak domain hits (9). When the HMMER score is below the curated threshold, there is often sufficient contextual information contained in the other domain hits to the protein to increase this score above the threshold. Language modelling is applied as a post-processing step to the set of all HMMER identified matches with an E-value of <1000. A dynamic programming algorithm is used to find the highest scoring domain architecture for a protein, taking into account both HMMER and context scores. We find 32 587 additional domain occurrences in this way, accounting for an additional 0.5% residue coverage in Pfam release 10.0.

IMPROVED STRUCTURE IMAGES

Despite the large increase in the number of Pfam entries, just over one-third of entries contain at least one protein of known 3D structure. Previously, Pfam used structure images kindly provided by the PDBsum database (10). To make the images more informative with respect to Pfam we now colour the structures by Pfam domain. This domain mark-up of structures was greatly aided by the mapping of PDB sequences to Swiss-Prot sequences provided by the EBI Macromolecular Structure Database (E-MSD) (11). The static images are generated using Molscript (12) and rendered using Raster3D (13). Each image is accompanied by a brief description of the structure, followed by the domain mark-up key, which contains links to the family pages for all the domains in the structure (Fig. 2).

AVAILABILITY

The Pfam database is freely available on the web in the UK (http://www.sanger.ac.uk/Software/Pfam/), the USA (http://pfam.wustl.edu/), France (http://pfam.jouy.inra.fr/) and Sweden (http://Pfam.cgb.ki.se/). All data are available for download in flat file form from the FTP sites linked from each Pfam website, and also as a set of MySQL relational database files.



1jsu Complex (transferase/cyclin/inhibitor)
P27(kip1)/cyclin a/cdk2 complex

Key:

Domain	Chain	Start Residue	End Residue
pkinase	A	13	286
cyclin_C	B	309	431
cyclin	B	181	307
CDI	C	30	80

The Swissprot/PDB mapping was provided by MSD

Figure 2. Structure image coloured by Pfam domain, including a coloured domain key below. This structure shows the complex of P27, cyclin and CDK2. Clicking on the image links to the PDBsum resource (10).

ACKNOWLEDGEMENTS

We would like to thank William Mifsud, Nicola Kerrison, David Waterfield and Ben Vella Briffa for adding many of the new families to Pfam. We are grateful to Kevin Howe for useful discussions and advice. We would also like to thank Timo Lassmann and Markus Wistrand for help maintaining the Sweden Pfam website and Lorenzo Cerutti for maintaining the French Pfam website. This work was funded by the The Wellcome Trust and and an MRC (UK) E-science grant.

REFERENCES

1. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
2. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
3. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
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. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
6. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
7. Sayle,R. and Milner-White,E. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–374.

8. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
9. Coin,L., Bateman,A. and Durbin,R. (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl Acad. Sci. USA*, **100**, 4516–4520.
10. Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
11. Boutselakis,H., Dimitropoulos,D., Fillon,J., Golovin,A., Henrick,K., Hussain,A., Ionides,J., John,M., Keller,P.A., Krissinel,E. *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, **31**, 458–462.
12. Kraulis,P. (1991) MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
13. Bacon,D. and Anderson,W. (1988) A fast algorithm for rendering space-filling molecule pictures. *J. Mol. Graph.*, **6**, 219–220.