Pfam: Multiple Sequence Alignments and HMM-Profiles of Protein Domains

Erik L.L. Sonnhammer[1*], Sean R. Eddy[2], Ewan Birney[3], Alex Bateman[3], Richard Durbin[3]

[1]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, Building 38A, Room 8N805 National Institutes of Health, Bethesda, MD 20894, USA; [2]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA, [3]Sanger Centre, Hinxton Hall, Cambridge CB10 1SA, UK

*) Corresponding author

Fax: +1-301-480-9241
Email: sonnhammer@ncbi.nlm.nih.gov

Abstract

Pfam contains multiple alignments and hidden Markov model based profiles (HMM-profiles) of complete protein domains. The definition of domain boundaries, family members, and alignment is done semi-automatically based on expert knowledge, sequence similarity, other protein family databases, and the ability of HMM-profiles to correctly identify and align the members. Release 2.0 of Pfam contains 527 manually verified families which are available for browsing and on-line searching on the World Wide Web at http://www.sanger.ac.uk/Pfam/ and http://genome.wustl.edu/Pfam/. Pfam 2.0 matches one or more domains in 50% of Swissprot-34 sequences, and 25% of a large sample of predicted proteins from the *C. elegans* genome.

Introduction

A relatively small number of structural and functional domains are used in a large number of different proteins. Particularly for protein analysis and annotation in large-scale sequencing projects, there is a growing need for easily interpretable and sensitive detection of common protein domains. A protein containing one or more common domains can produce a morass of hundreds or thousands of BLAST hits when searching single sequence databases (e.g. Genbank, Swissprot, PIR). Although searches can be augmented by tools that condense and summarise results[1], satisfactory annotation of such proteins often becomes a time-consuming and error-prone process. Instead, a search of an organised database of protein domain families can produce more concise results which simplify annotation, domain parsing, and functional prediction for a query sequence[2-5]. Protein family databases are typically based on multiple sequence alignments of known family members. Conserved features can be recognised in the alignment and given higher weight in searches, which for distant similarities can often render the comparison more sensitive than pairwise alignment approaches.

We present here Pfam[6] release 2.0. Pfam was developed in order to use HMM-profile analysis to complement BLAST analysis in the *C. elegans* genome project. The main distinction between Pfam and other protein family databases is that for all of Pfam, both the family definition and the search method span entire domains, including not only conserved motifs but also less-conserved regions, insertions, and deletions. HMM-profile methods allow variable conservation and insertions/deletions to be dealt with in a fairly robust way[7,8]. Modelling of complete domains should facilitate more biologically meaningful sequence annotation, and, in some cases, more sensitive detection.

Description of the data

Contents of Pfam

For each protein domain family in Pfam, there are three important files. The *seed alignment* is a manually verified multiple alignment of a representative set of sequences (figure 1). An *HMM-profile* is built from the seed alignment for database searching and alignment purposes. A *full alignment* is generated automatically from the seed HMM-profile by searching Swissprot for all detectable members and aligning them to the HMM-profile. The distinction between seed and full alignments facilitates updating the database; the seed alignments are stable resources, whereas full alignments and HMM-profiles can be generated automatically for any new Swissprot (or other sequence database) release.

Each family has a name, a permanent accession number, and a record of the methods used to identify the family members and create the alignments. There is also either a brief description of the usual function and structure of the domain, or (more often) links to other on-line documentation resources such as Prosite and Prints.

Both the seed and the full alignments are subjected to a small array of ‘quality control’ procedures, to verify that the alignments are sensible, that the HMM-detected sequences in the full alignment include all presumed members of the family in Swissprot and no other sequences, and that the family does not overlap with other Pfam families. The process of generating the Pfam family is iterated, if necessary, until all quality requirements are met.

Most Pfam families are based on, and cross-referenced to, corresponding Prosite or Prints entries. In many cases, however, the definition of which sequences belong to a family differs between the databases. This is a pragmatic consequence of the different search methods used. Prosite and Prints detection relies primarily on short conserved patterns corresponding to superfamily motifs. A Prosite pattern or Prints fingerprint may recognise a highly conserved motif shared amongst an otherwise highly diverged superfamily that Pfam splits into several families; conversely, Pfam may recognise a superfamily that Prosite and Prints classify into several distinct families with distinct motif signatures. For some protein domain families, there may be no motif sufficiently conserved to make a discriminative pattern or fingerprint. (Prosite is increasingly incorporating profiles for these families; these Prosite profiles are very similar to Pfam models.) Only the largest (>15 members) Prosite families were systematically used to construct Pfam entries. For smaller families, constructing a HMM-profile is of less value since the sensitivity is unlikely to improve relative to single-sequence searching, and because a small sample is often non-representative. Of the 71 Pfam families with no corresponding Prosite or Prints entry, 55 were 'discovered' as large clusters in Pfam-B (see below). 24 Pfam families contain links to other World Wide Web (WWW) protein family documentation resources, some of which were gleaned from the ProWeb server[9].

Pfam 2.0 contains 527 families, comprising 39113 sequence segments and 6.8 million residues in the full alignments. All sequences were taken from Swissprot 34[10]. The alignments are on average 275 residues wide, including gaps. There are on average about 75 members per family in full alignments, and about 22 in seed alignments.

Pfam-B. For comprehensiveness, all Swissprot sequences not in Pfam are clustered automatically by the program Domainer[2], which also constructs multiple alignments automatically and is the basis for the ProDom protein family database. The quality of these alignments tends to be low, but domain-based automated clustering is a convenient method of identifying large obvious families that need to be targeted for Pfam model construction. Although we do not stably maintain, annotate, or produce HMM-profiles of these clusters, we make them available as Pfam-B. Pfam-B 2.0 contains 13289 clusters, 62611 subsequences, and 8.2 million residues. On average, alignments are 146 residues wide (including gaps) and contain 5 members.

Sequence database coverage. As shown in figure 2, 48% of the sequences and 32% of the residues in Swissprot 34 are included in annotated Pfam alignments. If unannotated Pfam-B clusters are also taken into account, 81% of sequences and 71% of residues in Swissprot 34 are included in Pfam. In searches of a large and presumably unbiased set of predicted protein sequences from the *C. elegans* genome, 25% of sequences and 13% of residues show significant hits to Pfam HMM-profiles. The numbers are slightly lower for prokaryotic genomes.

Searching Pfam

The US and UK Pfam WWW servers provide users the ability to search query protein sequences against one, all, or a few Pfam HMMs. Results are returned in tabular format, and both GIF- and Java-based graphical representations are available optionally. An example of the results from such a search is shown in figure 3. Here, the *C. elegans* Kin-11 gene product (E01H11.1) is shown to possess a duplicated phorbol esters/diacylglycerol binding domain (DAG/PE-bind), a C2 domain, a protein kinase catalytic domain (pkinase), and a duplicated domain frequently associated C-terminally to protein kinase domains (pkinase_C).

Users can also use Pfam HMM-profiles to search protein sequences locally using the freely available HMMER software package at http://genome.wustl.edu/eddy/hmmer.html#hmmer. For comparing genomic and EST data to Pfam HMM-profiles, the programs GeneWise and ESTWise[11] are available at http://www.sanger.ac.uk/Software/Wise2/.

World Wide Web servers, FTP access, and format

The Pfam home pages are http://www.sanger.ac.uk/Pfam/ at the Sanger Centre in the UK and http://genome.wustl.edu/Pfam/ at Washington University in the USA. The two servers are separately maintained and differ slightly in their services and capabilities, but are based on the same underlying Pfam database. Both servers support HMM searching, browsing of the family alignments and documentation, and lookup of the domain organisation of proteins in Swissprot.

The entire database, including accessory data files such as Pfam schematics for Swissprot proteins, is also available as flat file format ASCII files by anonymous FTP at ftp.sanger.ac.uk and genome.wustl.edu in /pub/databases/Pfam/.

The format of the Pfam alignment flat files is based on the EMBL/Swissprot two character field labels. The following Pfam-specific labels are used: AL, alignment method of seed members; AM, alignment method of full alignment; AU, Author responsible for the alignments; GA, Gathering method/search program and cutoffs used to build full alignment; SE, Source suggesting the seed members belong to the same family; SQ, Sequence number (and last line before the alignment starts). The alignment is in a simple format (see figure 1) which consists of one line per subsequence containing the Swissprot sequence ID, start and end of the segment, and the aligned subsequence itself (no length limit). In the Pfam flat file, the corresponding Swissprot accession number is added to the right of each alignment line. Users of the Pfam database or WWW servers should cite this article as the appropriate reference.

Acknowledgements

We thank Robert Finn for preparing most of the new families for Pfam 2.0, and Jose Aguilar for writing and maintaining the Washington University Pfam server. Pfam development in SRE's group is supported by grant R01-HG01363 from the NIH National Human Genome Research Institute. Pfam development at the Sanger Centre is supported by the Wellcome Trust.

References

1. Sonnhammer,E.L.L. and Durbin,R. (1994) *Comput. Appl. Biosci.*, **10**, 301-307.

2. Sonnhammer,E.L.L. and Kahn,D. (1994) *Protein Sci.*, **3**, 482-492.

3. Attwood,T.K., Beck,M.E., Bleasby,A.J., Degtyarenko,K., Michie,A.D. and Parry-Smith,D.J. (1997) *Nucleic Acids Res.*, **25**, 212-217.

4. Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217-221.

5. Henikoff,J.G., Pietrokovski,S. and Henikoff,S. (1997) *Nucleic Acids Res.*, **25**, 222-226.

6. Sonnhammer,E.L.L., Eddy,S.R. and Durbin,R. (1997) *Proteins*, **28**, 405-420.

7. Krogh,A., Brown,M., Mian,I.S., Sjoelander,K. and Haussler,D. (1994) *J. Mol. Biol.*, **235**, 1501-1531.

8. Eddy,S.R. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361-365.

9. Henikoff,S., Endow,S.A. and Greene,E.A. (1996) *Trends Biochem. Sci.*, **21**, 444-445.

10. Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31-36.

11. Birney,E. and Durbin,R. (1997) In *ISMB-97; Proceedings Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, pp. 56-64.

Figures

Figure 1. Example of a typical Pfam entry, the SH2 family. Shown is the flat file record including a reduced version of the seed alignment.

Figure 2. Pfam 2.0 contains domains from nearly half of all Swissprot 34 proteins. The automatic clusters in Pfam-B 2.0 contain domains from 33% of the Swissprot proteins that do not contain Pfam domains. When counting residue-by-residue, roughly a third of Swissprot is covered by Pfam and Pfam-B each. Pfam-B does not include proteins known to be fragments or segments shorter than 30 residues; the figures for unique sequences are therefore overestimated.

Figure 3. Tabular output (a) and schematic output (b) from a Pfam search with the *C. elegans* protein E01H11.1 as query. Both pictures were taken from the Washington University WWW server.

```
ID   SH2
AC   PF00017
DE   Src homology domain 2
AU   Sonnhammer ELL
AL   Clustalw
AM   hmma -qR
SE   Swissprot_feature_table
GA   Bic_raw 25 hmmfs 20
DR   PROSITE; PDOC50001;
DR   SCOP; 1sha; sf;
SQ   58
ABL1_CAEEL/179-254  WYHGKISRSDSEAILGS..GITGSFLVRESETSIG...QYTISVRHDG......RVFHYRINVDNTE..KMFITQEVKFRTLGELVHHH
BLK_MOUSE/117-198   WFFRTISRKDAERQLLAPMNKAGSFLIRESESNKG...AFSLSVKDIT..TQGEVVKHYKIRSLDNG..GYYISPRITFPTLQALVQHY
BTK_HUMAN/281-362   WYSKHMTRSQAEQLLKQE.GKEGGFIVRDS.SKAG...KYTVSVFAKSTGDPQGVIRHYVVCSTPQS..QYYLAEKHLFSTIPELINYH
CSW_DROME/111-186   WFHGNLSGKEAEKLILERGK.NGSFLVRESQSKPG...DFVLSVRTDD......KVTHVMIRWQDK...KYDVGGGESFGTLSELIDHY
CSW_DROME/6-81      WFHPTISGIEAEKLLQEQGF.DGSFLARLSSSNPG...AFTLSVRRGN......EVTHIKIQNNGD...FFDLYGGEKFATLPELVQYY
CTK_HUMAN/122-196   WFHGKISGQEAVQQLQPP..EDGLFLVRESARHPG...DYVLCVSFGR......DVIHYRVLHRDG...HLTIDEAVFFCNLMDMVEHY
DRK_DROME/60-134    WYYGRITRADAEKLLSN..KHEGAFLIRISESSPG...DFSLSVKCPD......GVQHFKVLRDAQS..K.FFLWVVKFNSLNELVEYH
FER_HUMAN/460-531   WYHGAIPRIEAQELLKK....QGDFLVRESHGKPG...EYVLSVYSDG......QRRHFIIQYVDN...MYRFEG.TGFSNIPQLIDHH
FPS_DROME/438-510   WFHGVLPREEVVRLLNN....DGDFLVRETIRNEES..QIVLSVCWNG......H.KHFIVQTTGEG..NFRFEG.PPFASIQELIMHQ
FPS_FUJSV/511-581   WYHGAIPRSEVQELLKY....SGDFLVRESQGKQ....EYVLSVLWDG......QPRHFIIQAADN...LYRLED.DGLPTIPLLIDHL
FRK_HUMAN/116-193   WFFGAIGRSDAEKQLLYSENKTGSFLIRESESQKG...EFSLSVLDGA......VVKHYRIKRLDEG..GFFLTRRRIFSTLNEFVSHY
GTPA_HUMAN/181-256  WYHGKLDRTIAEERLRQAGK.SGSYLIRESDRRPG...SFVLSFLSQMN.....VVNHFRIIAMCG...DYYIGG.RRFSSLSDLIGYY
GTPA_HUMAN/351-426  WFHGKISKQEAYNLLMTVG.QVCSFLVRPSDNTPG...DYSLYFRTNEN......IQRFKICPTPNN...QFMMGGRYYNSIGDIIDHY
NCK_HUMAN/282-356   WYYGKVTRHQAEMALNER.GHEGDFLIRDSESSPN...DFSVSLKAQG......KNKHFKVQLKET....VYCIGQRKFSTMEELVEHY
P85A_HUMAN/624-698  WNVGSSNRNKAENLLRG..KRDGTFLVRES.SKQG...CYACSVVVDG......EVKHCVINKTATG..YGFAEPYNLYSSLKELVLHY
P85B_BOVIN/618-692  WYVGKINRTQAEEMLSG..KRDGTFLIRES.SQRG...CYACSVVVDG......DTKHCVIYRTATG..FGFAEPYNLYGSLKELVLHY
PIP4_RAT/668-741    WYHASLTRAQAEHMLMRVPR.DGAFLVRKR.NEPN...SYAISFRAEG......KIKHCRVQQEGQ....TVMLGNSEFDSLVDLISYY
SEM5_CAEEL/60-136   WYLGKITRNDAEVLLKKPTVRDGHFLVRQCESSPG...EFSISVRFQD......SVQHFKVLRDQNG..K.YYLWAVKFNSLNELVAYH
SHC_HUMAN/378-449   WFHGKLSRREAEALLQLN....GDFLVRESTTTPG...QYVLTGLQSG......QPKHLLLVDPEG....VVRTKDHRFESVSHLISYH
SRC1_DROME/162-244  WFFENVLRKEADKLLLAEENPRGTFLVRPSEHNPN...GYSLSVKDWED.GRGYHVKHYRIKPLDNG..GYYIATNQTFPSLQALVMAY
SRC2_DROME/214-292  WYVGYMSRQRAESLLKQG.DKEGCFVVRKS.STKG...LYTLSLHTKVP...QSHVKHYHIKQNARC..EYYLSEKHCCETIPDLINYH
SRK1_SPOLA/122-199  WFLGKIKRVEAEKMLNQSFNQVGSFLIRDSETTPG...DFSLSVKDQD......RVRHYRVRRLEDG..SLFVTRRSTFQILHELVDHY
SRK4_SPOLA/122-199  WFFGQVKRVDAEKQLMMPFNNLGSFLIRDSDTTPG...DFSLSVRDID......RVRHYRIKKLENG..TYFVTRRLTFQSIQELVAYY
STK_HYDAT/126-203   WYFGDVKRAEAEKRLMVRGLPSGTFLIRKAETAVG...NFSLSVRDGD......SVKHYRVRKLDTG..GYFITTRAPFNSLYELVQHY
SYK_HUMAN/15-92     FFFGNITREEAEDYLVQGGMSDGLYLLRQSRNYLG...GFALSVAHGR......KAHHYTIERELNG..TYAIAGGRTHASPADLCHYH
SYK_PIG/163-238     WFHGKISRDESEQIVLIGSKTNGKFLIRAR..DNG...SYALGLLHEG......KVLHYRIDKDKTG..KLSIPGGKNFDTLWQLVEHY
TEC_MOUSE/246-329   WYCRNTNRSKAEQLLRTE.DKEGGFMVRDS.SQPG...LYTVSLYTKFGGEGSSGFRHYHIKETATSPKKYYLAEKHAFGSIPEIIEYH
TXK_HUMAN/150-231   WYHRNITRNQAEHLLRQE.SKEGAFIVRDS.RHLG...SYTISVFMGARRSTEAAIKHYQIKKNDSG..QWYVAERHAFQSIPELIWYH
VAV_MOUSE/671-745   WYAGPMERAGAEGILTN..RSDGTYLVRQRVKDTA...EFAISIKYNV......EVKHIKIMTSEG...LYRITEKKAFRGLLELVEFY
YES_XIPHE/159-241   WYFGKLSRKDTERLLLLPGNERGTFLIRESETTKG...AYSLSLRDWDE.TKGDNCKHYKIRKLDNG..GYYITTRTQFMSLQMLVKHY
YKF1_CAEEL/20-101   YFHGLIQREDVFQLLDN....NGDYVVRLSDPKPGEPRSYILSVMFNNKLDENSSVKHFVINSVEN...KYFVNNNMSFNTIQQMLSHY
ZA70_HUMAN/163-239  WYHSSLTREEAERKLYSGAQTDGKFLLRPRK.EQG...TYALSLIYGK......TVYHYLISQDKAG..KYCIPEGTKFDTLWQLVEYL
ZA70_MOUSE/10-87    FFYGSISRAEAEEHLKLAGMADGLFLLRQCLRSLG...GYVLSLVHDV......RFHHFPIERQLNG..TYAIAGGKAHCGPAELCQFY
```

Figure 1.

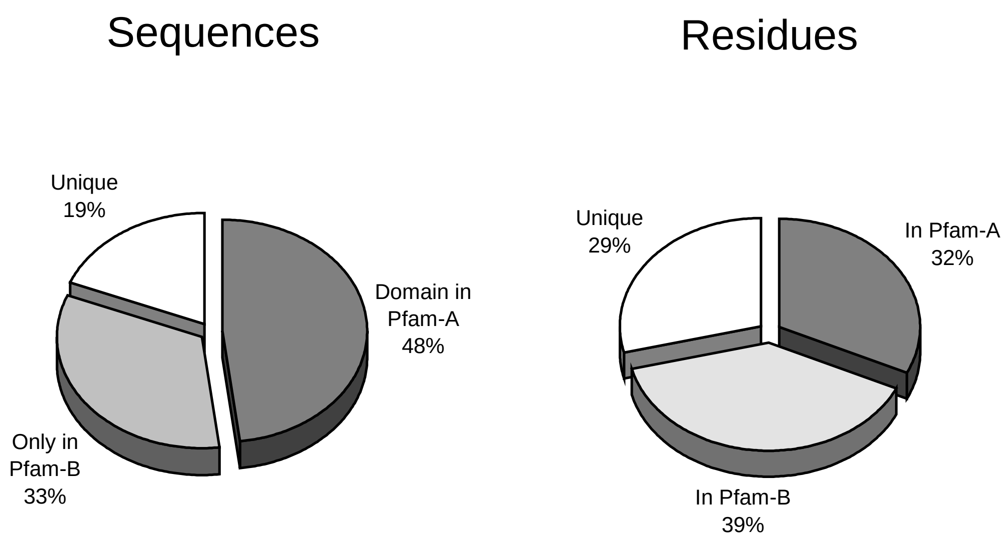
Sequences
Residues
Unique
19%
Domain in
Pfam-A
48%
Only in
Pfam-B
33%
Unique
29%
In Pfam-A
32%
In Pfam-B
39%

Figure 2.

Score	Query from	Query to	HMM from	HMM to	Pfam Family	Description
97.67	104	153	1	50	DAG_PE-bind	Phorbol esters / diacylglycerol binding domain
92.44	169	218	1	50	DAG_PE-bind	Phorbol esters / diacylglycerol binding domain
137.88	240	328	1	92	C2	C2 domain
276.16	413	674	1	247	pkinase	Eukaryotic protein kinase domain
84.44	675	741	1	69	pkinase_C	Protein kinase C terminal domain
70.99	807	857	17	69	pkinase_C	Protein kinase C terminal domain

Figure 3a

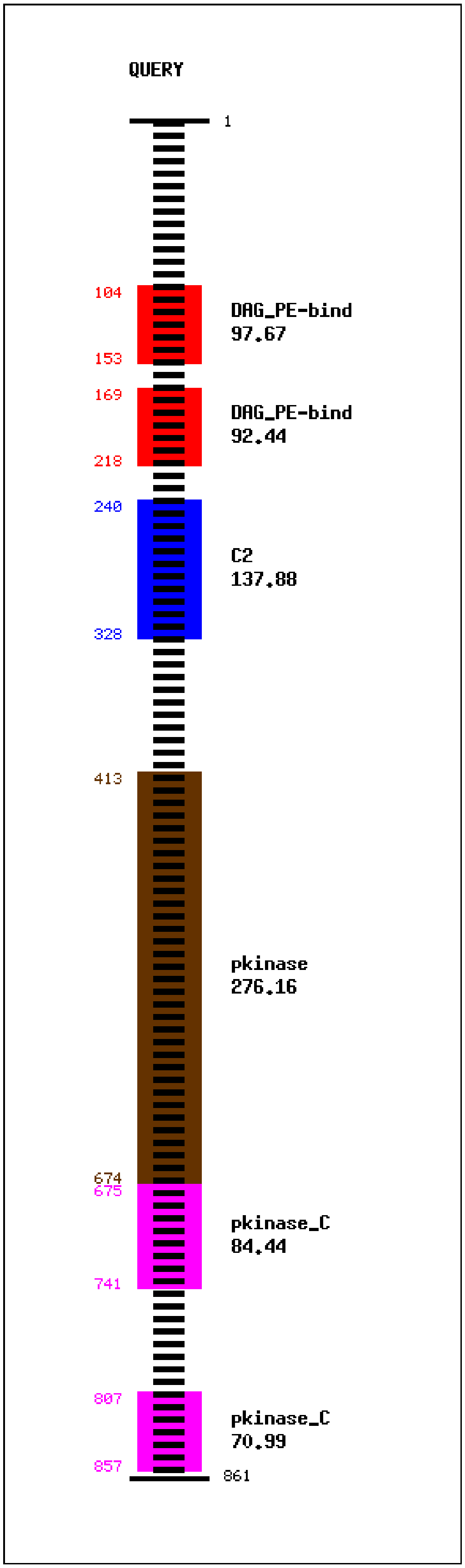
QUERY
1
104
DAG_PE-bind
97.67
153
169
DAG_PE-bind
92.44
218
240
C2
137.88
328
413
pkinase
276.16
674
675
pkinase_C
84.44
741
807
pkinase_C
70.99
857
861

Figure 3b