

# **Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.**

Alex Bateman\*, Ewan Birney, Richard Durbin, Sean R. Eddy<sup>1</sup>, Robert D. Finn and Erik L. L. Sonnhammer<sup>2</sup>.

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, England,

<sup>1</sup>Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA,

<sup>2</sup>Center for Genomics Research, Karolinska Institutet, S-171 77 Stockholm, Sweden.

## ***ABSTRACT***

Pfam is a collection of multiple alignments and profile hidden Markov models of protein domain families. Release 3.1 is a major update of the Pfam database and contains 1313 families which are available on the World Wide Web in Europe at <http://www.sanger.ac.uk/Software/Pfam/> and <http://www.cgr.ki.se/Pfam/>, and in the US at <http://pfam.wustl.edu/>. Over 54% of proteins in SWISS-PROT-35 and SP-TrEMBL-5 match a Pfam family. The primary changes of Pfam since release 2.1 are that we now use the more advanced version 2 of the HMMER software, which is more sensitive and provides expectation values for matches, and that it now includes proteins from both SP-TrEMBL and SWISS-PROT.

## ***INTRODUCTION***

Pfam is a database of protein families that is designed to be both accurate and comprehensive (1,2). Pfam is composed of two parts; the first part, Pfam-A, contains curated families each with an associated profile hidden Markov model (profile HMM) (3,4) that can be used for alignment and database searching. The second part of Pfam is Pfam-B, in which sequence segments that are not included in Pfam-A are clustered automatically, allowing Pfam to be comprehensive.

Each Pfam-A family consists of four elements, 1) annotation, 2) a *seed* alignment, 3) a profile HMM and 4) a *full* alignment. The annotation contains several compulsory fields that indicate the source used to make a family, how the alignment was made, thresholds for the profile HMM and details of the profile HMM construction. An example of a Pfam entry can be seen in Figure 1. The optional annotation contains references to the literature, World Wide Web URLs, cross-links to other databases and comment fields containing functional information. The *seed* alignment is a curated alignment that contains representative members of the family which are judged to be well aligned. The seed alignment contains minimal redundancy and is meant to change infrequently only to improve the alignment or extend the scope of the family. The profile HMM is constructed from the seed alignment using the HMMER 2 software. This profile HMM can then be used to search a sequence database for matches to the family. For each release of Pfam the profile HMMs are used to search a protein database. From each database search, sequences scoring above the family specific threshold are aligned to the profile HMM automatically to make a *full* alignment. The full alignment should contain all the known members of the protein family in the database.

To make Pfam comprehensive all the sequence segments that are not in Pfam-A are clustered together. This automatic clustering uses the Domainer algorithm, which was the basis for early versions of the ProDom database (5). This can produce poor alignments, but these are useful as a guide to relationships among families that are not yet included in Pfam-A.

## **USING PFAM**

The Pfam web sites allow the database to be queried in one of three ways: first, a user may have a new sequence for which they know nothing. In such a case, this sequence can be searched against the current collection of Pfam profile HMMs to locate regions of the sequence that belong to known domain families. Protein matches to Pfam-A profile HMMs are now displayed using a graphical representation. An example of this is shown in Figure 2. Second, if the user already has a SWISS-PROT or SP-TrEMBL identifier for the sequence they can access precalculated matches using the Swisspfam resource. In such cases the regions of the target sequence belonging to Pfam-A and Pfam-B are displayed. Finally, users can browse the information in Pfam by family or use a text search of Pfam and related PROSITE annotation to find families of interest.

Multiple alignments are a central feature of Pfam. The web sites provide access to the seed and full alignments in a variety of formats, allowing users to input Pfam data into their own software. Alignments are best viewed with specialised programs that can highlight similar regions and carry out manipulations of the alignment. We provide three viewers, Belvu (Unix only), java alignment viewer and jalview (See URL <http://circinus.ebi.ac.uk:6543/jalview/contents.html>), which can be automatically launched from the web sites.

## **CHANGES TO PFAM**

In earlier releases of the Pfam database only proteins from the SWISS-PROT database (6) were included in the full alignment. To give a more comprehensive coverage of known protein sequences we now also include proteins from SP-TrEMBL(6). The current release is built from searches of a fixed database called *pfamseq* that is composed of all proteins from SP-TREMBL-5 and SWISS-PROT-35. The Pfamseq database currently contains 67,193,197 residues in 209,668 proteins. The Pfamseq database is available from the Pfam FTP sites (see below).

Pfam 3.1 uses the new HMMER 2 package for all profile HMM construction, database searching and construction of full alignments. HMMER 2 format profile HMMs are not compatible with the HMMER 1 software. Pfam 3.1 contains the exact method used to construct the HMMs in the BM (build method) field of the annotation file. HMMER 2 contains two database searching programs: *hmmsearch* that replaces all previous HMMER 1 programs for searching a single profile HMM against a database, and the *hmmpfam* program which allows searching a collection of profile HMMs, such as Pfam with a single sequence. The new database searching programs, *hmmsearch* and *hmmpfam* give two types of score for each sequence. The first is like the scores from HMMER 1, which gives the score per domain match in bits. The second type of score gives the profile HMM combined score of all domain matches in a sequence to a profile HMM. This is particularly useful for proteins with tandem domains and can allow the sensitive detection of sequences belonging to a family even when the match to each individual domain is weak. Except in certain cases involving score adjustments due to repetitive sequence masking in HMMER, the combined score is the sum of domain scores. HMMER 2 contains estimates of statistical significance using extreme value distribution fitting of randomly generated sequences to the profile HMM using the *hmmcalibrate* program. This greatly increases the

sensitivity of database searching for many families. In some families, negative bit scores are shown to have significant expectation scores with the extreme value statistics.

## **PFAM STATISTICS**

Pfam release 3.1 contains 1313 families, which includes matches from 114,750 sequences covering 27,573,470 residues. This represents 54% of sequences in pfamseq and 41% of residues in pfamseq. The top 1000 families match over 50% of the proteins in the pfamseq database. To see how Pfam has progressed we compared the coverage of Pfam release 3.1 with the previous public release of Pfam, version 2.1. For Pfam release 2.1 48% of SWISS-PROT sequences had at least one match to a Pfam-A family and for Pfam release 3.1 61% of SWISS-PROT sequences had at least one match to a Pfam-A family.

Pfam can be used to estimate what proportion of the sequence databases can be structurally modeled using standard comparative modeling techniques. 516 families have a link to the SCOP database (7), which accounts for 32% of the sequences and 22% of the residues in pfamseq. Interestingly these figures are very similar to those quoted for the complete genome of *Mycoplasma genitalium* (8). This is an underestimate because there are structures solved whose sequence does not belong to a Pfam family.

## **QUALITY CONTROL**

There are a number of quality issues in maintaining Pfam. The principle quality control comes from the manual curation of the *seed* alignments, their resulting profile HMMs and the associated annotation of the family. When problems in the database are noticed by individuals, they can be resolved by modifying the seed alignment. The seed alignments are never automatically rebuilt, even between releases, making the changes to them a permanent feature of the database. A second, automatic quality control feature is provided by the fact that we do not permit any overlap between families in either the *seed* or the *full* alignments. This represents the biological constraint that structural domains do not overlap each other. When erroneous alignments are made, they often overlap with a number of other families, immediately highlighting the error to the curator. As the coverage of Pfam is over 50%, this non-overlapping criterion is a powerful quality control measure. Finally, the integrity of

the database is checked with respect to the underlying protein database (SWISS-PROT and SP-TrEMBL) and the annotation files have strict format definitions. We base Pfam on a fixed release of SWISS-PROT and SP-TrEMBL, but when we update Pfam to a new fixed release we manually check the changes to the *seed* alignments to ensure that they continue to represent their families. These checks are designed to ensure the integrity of the database is maintained.

## **AVAILABILITY OF PFAM**

Pfam is available for browsing and interactive searching on the World Wide Web in Europe at <http://www.sanger.ac.uk/Pfam/> and <http://www.cgr.ki.se/Pfam/>, and in the US at <http://pfam.wustl.edu/>. The Pfam database is available as several data files: the file *Pfam* contains the profile HMMs; *Pfam-A.seed* contains annotation and seed alignments; *Pfam-A.full* contains annotation and the full alignments; *Pfam-B* contains the alignments from Pfam-B; *pfamseq* contains the sequences from the current *pfamseq* database in fasta format. The files are available by anonymous FTP in Europe from <ftp.sanger.ac.uk> in `/pub/databases/Pfam/` and from <ftp.cgr.ki.se> in `/pub/data/Pfam`, and in the US from <ftp.genetics.wustl.edu> in `/pub/Pfam/`.

## **ACKNOWLEDGEMENTS**

We are grateful to Sarah Teichmann, Arne Elofsson, Chris Ponting, Joerg Schulz and Peer Bork for providing new families, and to Mats Jonsson for writing the graphical representation software.

## **FIGURE CAPTIONS**

Figure 1. An example of a Pfam entry from the flatfile release, for the DnaJ domain. The Pfam entry is composed of three sections: a section of compulsory fields, optional family specific annotation and the alignment. The Pfam database format is based on EMBL/SWISS-PROT field labels. The following Pfam specific labels are used: AU, author of the Pfam entry; SE, source suggesting members of the seed are related; AL, alignment method of seed members; BM, the building method for the profile HMM; GA, gathering method/ search program and cutoffs used to build full alignment; TC, lowest sequence and lowest domain bits score found in a member of the full alignment; NC, highest sequence and highest domain bits score of matches not included in the full alignment; SQ, number of sequences in alignment. The alignment format used in Pfam is a single line per subsequence: the first

column has the sequence identifier followed by the start and end points in the sequence, the second column contains the alignment, and the final column contains the accession number.

Figure 2. An example of a Pfam query with the sequence `GTPA_HUMAN`. The matches are listed in text form at the top, followed by the graphical representation of the protein. The first column shows a !! mark if the match is above the trusted cutoff for the family. The second column gives the domain family name followed by the start and end points in the query sequence. The fourth column gives the score of the match in bits, the fifth column gives the E-value of the match. Each match to a Pfam domain is shown in graphical form as a single colour and is hyperlinked to the family specific page.

## REFERENCES

1. Sonnhammer, E. L. L., Eddy, S. R. and Durbin, R. (1997) *Proteins*, **28**, 405-420.
2. Sonnhammer, E. L. L., Eddy, S. R., Birney, E., Bateman, A. and Durbin, R. (1998) *Nucleic Acids Res.*, **26**, 320-322.
3. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994) *J. Mol. Biol.*, **235**, 1501-1531.
4. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361-365.
5. Sonnhammer, E. and Kahn, D. (1994) *Protein Sci.*, **3**, 482-492.
6. Bairoch, A. and Apweiler, R. (1998) *Nucleic Acids Res.*, **26**, 38-42.
7. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.*, **247**, 536-540.
8. Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y. and Bork, P. (1998) *J. Mol. Biol.*, **280**, 323-326.

Figure 1

```

ID      DnaJ
AC      PF00226
DE      DnaJ domain
AU      Finn RD, Birney E
SE      Prosite
AL      Clustalw
BM      hmmbuild HMM SEED
BM      hmmscalibrate --seed 0 HMM
GA      hmmssearch -T 10 --domT 10
TC      10.50 10.50
NC      7.90 7.90
RN      [1]
RM      94287451
RT      DnaJ-like proteins: molecular chaperones and specific regulators of Hsp70.
RA      Cyr DM, Langer T, Douglas MG;
RL      Trends Biochem Sci 1994;19:176-181.
RN      [2]
RM      97415577
RT      Inactivation of pRB-related proteins p130 and p107 mediated by the J domain
RT      of simian virus 40 large T antigen.
RA      Stubdal H, Zalvide J, Campbell KS, Schweitzer C, Roberts TM, DeCaprio JA;
RL      Mol Cell Biol 1997;17:4979-4990.
CC      The structure of the DnaJ domain by NMR.
RN      [3]
RM      96291434
RT      NMR structure of the J-domain and the Gly/Phe-rich region of the
RT      Escherichia coli DnaJ chaperone.
RA      Pellicchia M, Szyperki T, Wall D, Georgopoulos C, Wuthrich K;
RL      J Mol Biol 1996;260:236-250.
DR      SCOP: lxbl; sf;
DR      PROSITE; PDOC00553;
CC
CC      DnaJ domains (J-domains) are associated with hsp70
CC      heatshock system and it is thought that this domain
CC      mediates the interaction. DnaJ-domain is therefore
CC      part of a chaperone (protein folding) system
CC
CC      The T-antigens, although not in Prosite are
CC      confirmed as DnaJ containing domains from literature
CC      (see ref 2 above)
SQ      38
CAJ1_YEAST/6-71      EYYDILGIKP.....EATPTEIKKAYRRKAMETHPKHPD.....DPDAQAKFQAVGEAYQVLSDFGLRSKYDQFG P39101
CBPA_ECOLI/5-69      DYVAIMGVVKP.....TDDLKTIKTAYRRRLARKYHPDVSKP.....PDAAERFKEVAEAEWEVLSDEQRRRAEYDQMW P36659
CSP_RAT/15-80        SLYHVLGLDK.....NATSDDIKKSYRKLALKYHPDKNPD.....NPEAADKFKEINNAHAILTDTAKRNIYDKYG P54101
DNAJ_ERYRH/6-70      DFYEILGVSK.....SATDAEIKKAYRQLAKKYHPDINKE.....DGAAEKFEVQEAYEVLSDSQKRANYDQFG Q05646
DNAJ_HAEDU/5-70      DYVEVLGLQK.....GATEKDIKRAYRKLAAKYHPDKNQG.....SKDSEEFKQITEAYEILTDQKRAAYDQYG P48208
DNJ1_HUMAN/4-68      DYYQTLGLAR.....GASDEEIKRAYRRQALRYHPDKNKE.....PGAEKFEKIEAEAYDVLSDPRKREIFDRYG P25685
DNJ2_ALLPO/13-74     KYEVLGVSK.....NATPEDLKKAYRKAIAIKNHPDKGGD.....PEKFKEIQQAYEVLNDPEKREIYDQYG P42824
DNJL_MYCGE/2-64      NLYDLELPT.....TASIKEIKIAYKRLAKRYHPDVNKL.....GSQTFVEINNAYSILSDPNQKKEYDSML P47248
DNJM_MYCGE/7-71      DYVEVLGITP.....DADQSEIKKAFRKLAKKYHPDRNNA.....PDAAKIFAEINEANDVLSNPKKRANYDKYG P47442
HLJ1_YEAST/21-85     EFYEILKVDR.....KATDSEIKKAYRKLAIKLHPDKNSH.....PKAGEAFKVINRAFEVLSNNEEKRSIYDRIG P48353
NPLL_YEAST/125-196   DPYEILGIST.....SASDRDIKSAYRKLVSFKFHPDKLAKGLT.....PDEKSMVEETYYVQITKAYESLTDDELVRQNYLKYG P14906
PSI_SCHPO/6-68       KLYDCLEVRP.....EASEAELKKAYRKLALKYHPDKNPN.....GEKKFKEISLAYEVLSDPQRRKLYDQYG Q09912
RESA_PLAFF/523-587   LYVDILGVGV.....NADMNEITERYFKLAENYYPYQKRS.....STVFNHFRKVNAYQVLDGIDKKRWYNYG P13830
TAMI_POVHA/12-75     ALISLLDLEPQ.....YWG DYGRMOKCYKCKLQQLHPDKGGN.....EELMQQLNTLWTKLKDGLYRVRLLG P03079
TASM_BFDV/6-67       RLTELLCLPV.....TATAADIKTAYRRRTALKYHPDKGGD.....EEKMKELNTLMEEFRETEGLRADETLE P13895
XDJ1_YEAST/9-77      RLYDVLGVTR.....DATVQEIKTAYRKLALKHHPDKYVDQD.....SKEVNEIKFKEITAAEYILSDPEKKSHYDLYG P39102
YD1J_SCHPO/32-110    TPYEILELPR.....TCTANDIKRKYIELVKKHHPDKMKNASQLAPTESPPEINKHNEEYFRLLLANALLSDKRRREYDRFG Q10247
YFHE_ECOLI/2-74      DYFTLFLGLPAR....YQLDQALSLRFQDLQRQYHPDKFASGSQ.....AEQLAAVQOSATINQAWQTLRHPLMRAEYLLSL P36540
YFL1_YEAST/44-108    NFYKFLKLPKL....QNSSTKEITKNRKLKSKKYHPDKNPK.....YRKLIERLNLATQILSNSSNRKIYDYDL P43613
YGB8_YEAST/13-82     TFYELFPKTFPKKLP...IWTIDQSRRLRKEYRQLQAQHPDMAQQ.....GSEQSSTLNQAYHTLKDPLRRSQYMLKL P53193
YGM8_YEAST/79-151    NLYDVLELPTPLDVHTIYDDLPLQIKRKYRRTLALKYHPDKHPD.....NPSIIHKFHLLSTATNLLTNADVRPHYDRWL P52868
YIS4_YEAST/6-71      EYYDILGVST.....TASSIEIKKAYRKKSIQEHHPDKNPN.....DPTATERFQAISEAYQVLDGDDDLRAKYDKYG P40564
YJ67_YEAST/8-76      THYEILRIPS.....DATQDEIKKAYRNRLNTHPKLKSISI.....HDTVSNVTINKIQDAYKILSNIKTRREYDRLLI P47138
YJH3_YEAST/585-655   DYYKILGVSP.....SASSKEIRKAYLNLTCKYHPDKIKANHN.....DKQESIHETMSQINEAYETLSDDDKRKEYDLRS P40358
YJQ2_YEAST/13-77     TYYSILGLTS.....NATSSEVHKSYLKLARLLHPDKTKS.....DKSEELFKAVVHAHSILTDQKLRDYDRDL P46997
YLW5_CAEEL/531-595   DYYKTLGVDK.....KSDAKAIKKAYFQLAKKYHPDVNKT.....KEAQTQFQEI SEAYEVLSDTKRQYDAYG P34408
YNW7_YEAST/4-70      CYEELLGVET.....HASDLELKKAYRKLALQYHPDKNPDN.....VEEATQKFAVIRAAYEVLSDPQERAWYDSHK P53863
YQ07_CAEEL/562-626   DAYSVFLGRS.....DCSDDDIKRNKYRKLAAALVSPDKCTI.....DAADQVYELVDVAFSAIGYKDSRSEYTLN P09446
ZRF1_MOUSE/88-161    DHYAVLGLGHVR....YTATQRQIKAAHKAMVLLKHPDKRKAAGE.....PIKEGDNDFYFTCIKRAYEMLSDPVKRRAPNSVD P54103
ZU01_YEAST/97-168    DLYAAMGLSKLR....FRATESIQIIKAHRKQVVKYHPDKQSAAG.....GSLDQDGFKKI IQKAFETLTDNSNKRQAYDSCD P32527

```

Figure 2

**Starting search. Estimated time: 33 seconds. Please wait...**

## Pfam HMM search results

Clicking on the model name takes you to the Pfam documentation for that protein family.

---

Model	Seq-from	Seq-to	Score	E-value	Description
!! SH2	181	256	112.1	<b>1.2e-37</b>	Src homology domain 2
!! SH3	282	339	44.6	<b>2.2e-09</b>	Src homology domain 3
!! SH2	351	426	116.8	<b>2.8e-39</b>	Src homology domain 2
!! PH	474	577	110.0	<b>1.4e-31</b>	PH (pleckstrin homology) domain
!! C2	596	675	20.5	<b>0.0018</b>	C2 domain
!! RasGAP	769	943	320.3	<b>2.2e-92</b>	GTPase-activator protein for Ras-like GTPases

