Pfam: multiple sequence alignments and HMM-profiles of protein domains

Erik L. L. Sonnhammer*, Sean R. Eddy1, Ewan Birney2, Alex Bateman2 and Richard Durbin2

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 via the World Wide Web in the UK at http://www.sanger.ac.uk/Pfam/ and in the US at 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 Caenorhabditis 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 facilitates updating the database; the seed alignments are stable, 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 superfamilies. A Prosite pattern or Prints fingerprint may recognise a highly conserved motif shared amongst an otherwise highly diverged superfamly that Pfam splits into several families; conversely, Pfam may recognise a superfamly 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

*To whom correspondence should be addressed. Tel: +1 301 435 5930; Fax: +1 301 480 9241; Email: sonnhammer@ncbi.nlm.nih.gov
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 an 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 39 113 sequence segments and 6.8 million residues in the full alignments. All sequences were taken from Swissprot 34 (10). The alignments are on average ∼75 members per family in full alignments, and ∼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 13 289 clusters, 62 611 sub-sequences, and 8.2 million residues. On average, alignments are 146 residues wide (including gaps) and contain five members.

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.

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 domains.
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, number of sequences (and last line before the alignment starts). The alignment is in a simple format (Fig. 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**