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

Alex Bateman*, Ewan Birney, Richard Durbin, Sean R. Eddy¹, Robert D. Finn and Erik L. L. Sonnhammer²

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, ¹Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA and ²Center for Genomics Research, Karolinska Institutet, S-171 77 Stockholm, Sweden

Received October 2, 1998; Accepted October 8, 1998

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, (i) annotation, (ii) a seed alignment, (iii) a profile HMM and (iv) 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.

*To whom correspondence should be addressed. Tel: +44 1223 494 991; Fax: +44 1223 494 919; Email: agb@sanger.ac.uk
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.

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

Downloaded from https://academic.oup.com/nar/article-abstract/27/1/260/1236951 by guest on 31 March 2019
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 the pfamseq database. The top 1000 families match over 50% of the proteins from 114 750 sequences covering 27 573 470 residues. 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.

REFERENCES