Broadening Pfam Sequence Annotations

Jaina Mistry, Penny Coggill, Sean Eddy[§], Rob Finn, John Tate, Alex Bateman

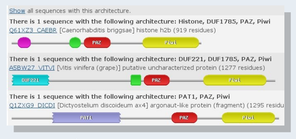
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK

[§]Janelia Farm Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Asburn, VA, US

What is Pfam?

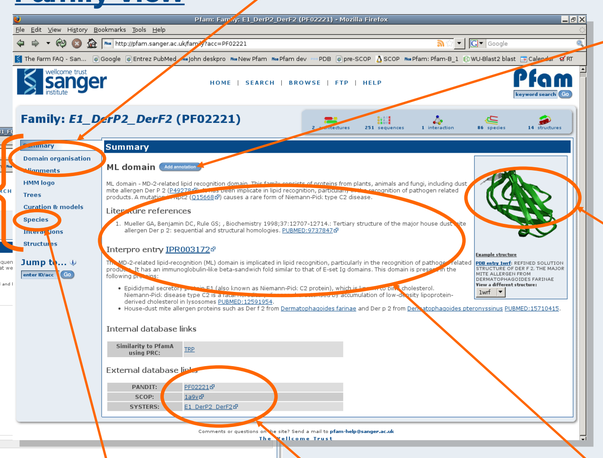
Pfam is a database, of conserved protein families or domains, commonly used for proteome annotation and sequence classification. It comprises two parts: (1) **Pfam-A** families, which are **manually annotated**, and consist of a representative seed alignment, **hidden Markov models** (HMMs), and a full alignment of all sequences that score above the curated threshold; and (2) **Pfam-B** families, **automatically** generated clusters of similar sequence regions not matched by Pfam-A that often indicate the presence of a domain. Many of the Pfam-A families are arranged into a hierarchical classification, termed clans. You can access and download the Pfam data via the website at **http://pfam.sanger.ac.uk**

Pfam website



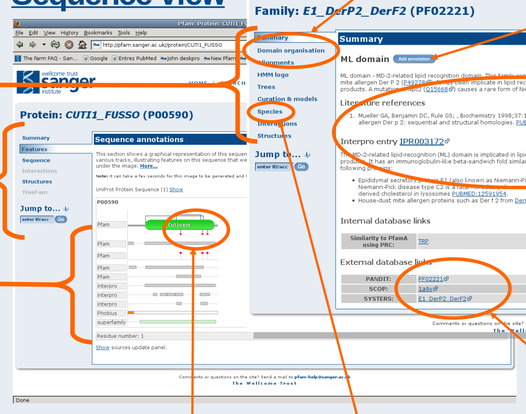
Different combinations (' architectures') of domains

Family view



Annotation submission button

Sequence view



Each tab shows a different view of the data

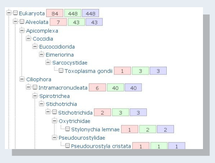
Interactive **structural** view

External protein annotation via **DAS**

Pfam and InterPro **annotation**

Pfam conservatively transfers known **active sites** between sequences in the same Pfam family. In Pfam release 23.0, over 1 million active site residues were predicted.

Method described in Mistry, Bateman, Finn, BMC Bioinformatics. 9;8:298 (2007)



Species distribution of the Pfam family

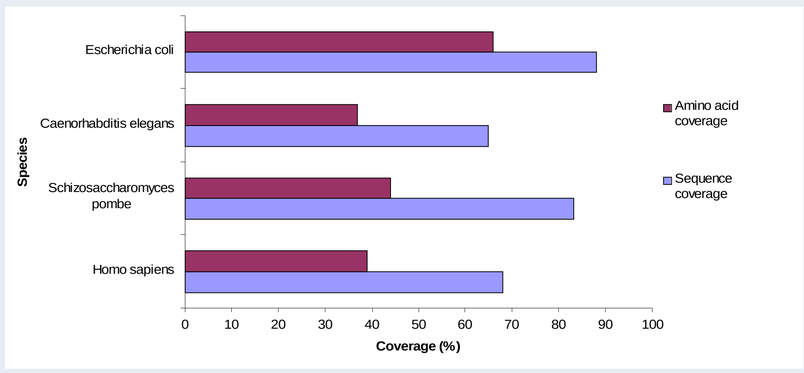
Links to other databases

Pfam coverage of proteomes

The proteome coverage of Pfam varies between species. Coverage is typically measured in the following ways:

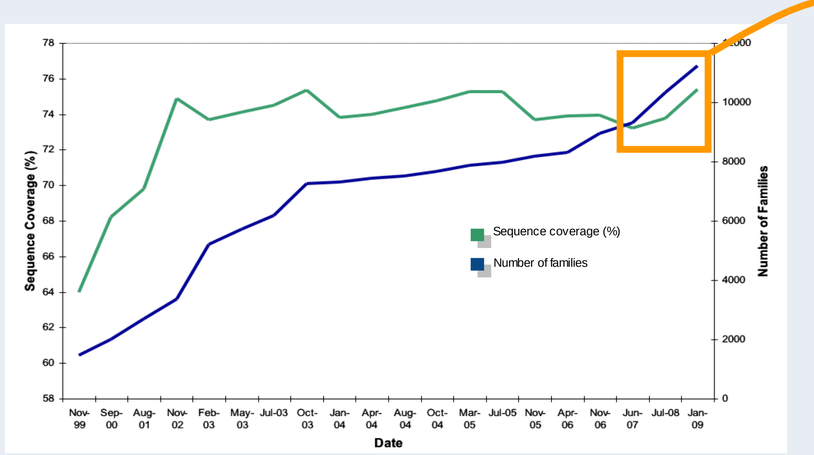
Sequence coverage is defined as the proportion of sequences that have a match to at least one Pfam-A family

Amino acid coverage is defined as the proportion of amino acids that belong to a Pfam-A family.



The coverage of a few **model organisms** is shown above. We achieve a much higher sequence coverage than amino acid coverage, and our coverage of **bacterial** proteomes is **better** than for other species.

Towards a complete classification of protein space

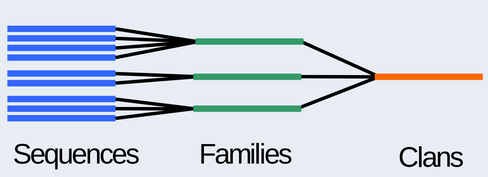


Adapted from Sammut, Finn, Bateman Brief. Bioinformatics 9:210 (2008)

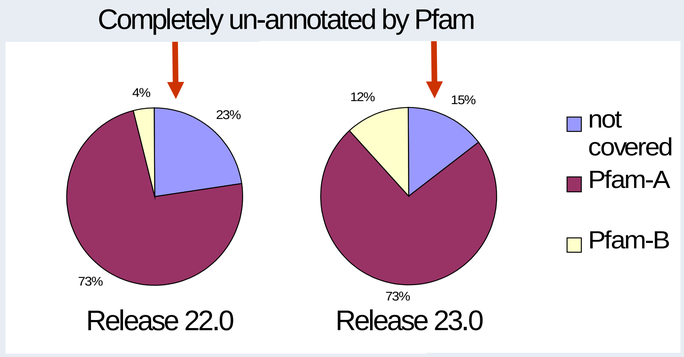
As the protein sequence databases continue to grow, Pfam **maintains** its coverage at **~75%** by adding to the existing families.

In a further drive to increase coverage, over the last year we have used the following methods

- Accelerated **building** of ~1000 new families from Pfam-B and structures
- **Expanding** the **diversity** of sequences in seed alignments of older families to reflect the contents of the current sequence database
- Moving to using the **ADDA** database for making **Pfam-B** families as it is more comprehensive than PRODOM, used in previous releases



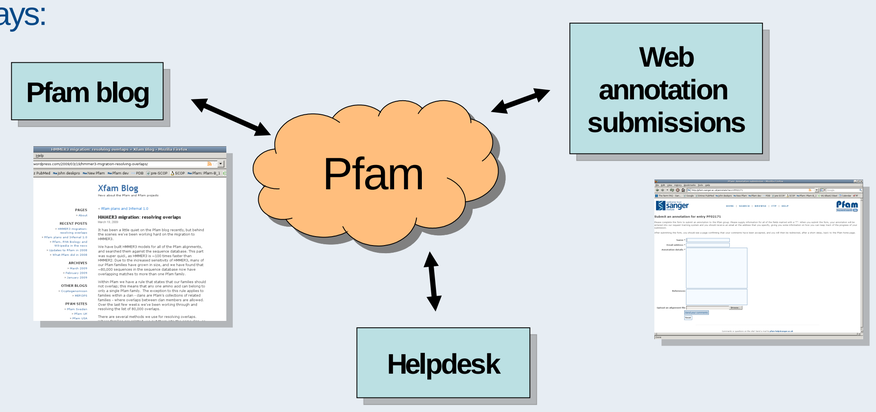
- 25% of families are now classified into 400 clans; this allows transfer of annotation between families and identification of remote structural homologues.



Each pie chart represents the total number of sequences in UniProtKB

Interacting with our community

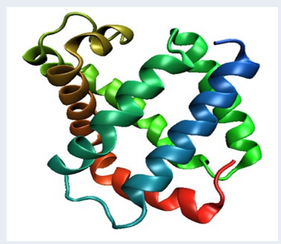
We support our user-community and receive feedback in the following ways:



We **welcome** receipt of alignments, annotation and references for **new families**, and annotation-updates on existing families. All incoming queries to our helpdesk **pfam-help@sanger.ac.uk** are tracked.

Our **blog** informs users about Pfam news and future plans. It is linked from the Pfam website, or you can visit it at **http://xfam.wordpress.com**

Improved speed and sensitivity with HMMER3

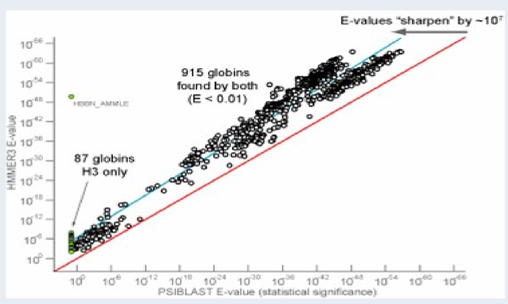


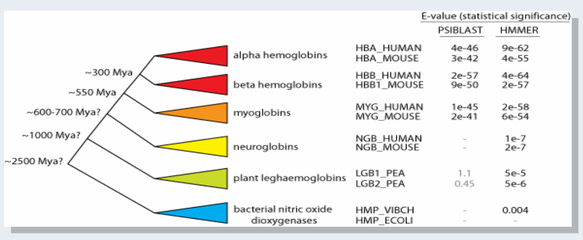
Aplysia myoglobin (PDB 1mba) Wikipedia

An initial profile-HMM was made from three vertebrate hemoglobins and one myoglobin using **HMMER3** hmmbuild.

The HMM was searched with **HMMER3** hmmsearch against Uniprot 7.0 (207K seqs, containing ~ 1060 known globins). The results were **compared** with a **PSI-BLAST** search, starting with the same four sequences.

With a cut-off at E <= 0.01:
PSI-BLAST finds **915** globins (in 9 sec)
HMMER3 finds **1002** globins (in 10 sec)





HMMER3 is **more sensitive** than PSI-BLAST in finding more distant relatives, and is 100 times **faster** than HMMER2.

Nature Precedings : doi:10.1038/npre.2009.3194.1 : Posted 28 Apr 2009