This quick tour provides a brief introduction to the protein families database Pfam, based at EMBL-EBI. Each protein family is represented by multiple sequence alignments and a profile hidden Markov model, which is used to create a probabilistic model that represents the family.

Learning objectives:

- Provide a basic description of what Pfam provides.
- Search and discover protein-related information with Pfam.
- Describe where to find out more about Pfam.

What is the Pfam resource?

What is Pfam?

Proteins generally have one or more functional regions, which are commonly termed 'domains'. The presence of different domains in varying combinations on different proteins gives rise to the diverse functional repertoire found in nature. Identifying the domain(s) present in a protein can provide insight into the function of that protein. Pfam is a database of these conserved evolutionary units.

Each Pfam entry is represented by a set of aligned sequences with their probabilistic representation - called a profile hidden Markov model [5] (HMM). The profile HMM is trained on a small representative set of aligned sequences that are known to belong to the family (the 'seed' alignment). This model is then used to search exhaustively against a large sequence database (e.g. UniProtKB [6]) to find all homologous [7] sequences. Those sequences that are significantly similar to the model are aligned to the profile HMM in order to provide the full alignment.

Related Pfam entries may be grouped into sets, labelled as 'Clans'. These are typically large and divergent superfamilies, where a single profile HMM is insufficient to capture all members of a sequence.

Why do we need Pfam?

Our ability to generate sequence data far exceeds the rate at which we can functionally characterise sequences
experimentally. Therefore, computational methods are needed to help identify regions of similarity between sequences. Matching sequences to a Pfam entry allows us to transfer the functional information from an experimentally characterised sequence to uncharacterised sequences in the same entry. Pfam then provides comprehensive annotation \[8\] for each entry.

**What can I do with Pfam?**

With Pfam you can:

- search your protein or DNA sequence against our models;
- browse our families and clans;
- retrieve text annotation about any given family/entry;
- view multiple sequence alignments of a family or clan;
- view relationships between families in a clan;
- see protein structure information in the context of a family;
- view families according to their taxonomic spread;
- search the database by keywords

To access the Pfam database go to: http://pfam.xfam.org/ \[9\]

**More on Pfam families and clans**

**Pfam entries are classified into six different categories**

Depending on the length and nature of the sequence regions included in the entry, they are classified into one of the following six categories:

1) **Domain:**
A collection of related sequence regions that form a distinct structural unit.

2) **Family:**
A collection of related sequence regions that may contain one or more domains, but where there is insufficient evidence to support subdivision.

3) **Repeat:**
A short unit which is unstable in isolation but forms a stable structure when multiple copies are present.

4) **Motif:**
A short unit that carries a distinct role, for example in metal binding.

5) **Coiled-coil:**
Regions of a protein that form alpha-helices that align against each other to form a distinctive structure called a coiled-coil.

6) **Disordered:**
Regions on proteins that are inherently disordered but have sequence conservation.

**What is a Pfam clan?**
Related Pfam entries are grouped together into clans. The relationship between families in a clan may be defined by:

- sequence similarity (whilst still originating from a common ancestor [10]);
- similarity of known three-dimensional structures;
- functional similarity;
- similarity between their profile HMMs, as determined by algorithms such as HHsearch.

### Getting started with Pfam

Pfam [9] features separate pages for each of the following data types:

- family;
- clan;
- sequence;
- structure;

You can query each of the above pages using the accession [12] or name of the entity, or by searching for related keywords.

There are a series of ‘quick links’ available on the Pfam home page that allow you to query the database by a number of different methods. Clicking on any of the quick links (see the red arrow in Figure 1, below) will open a dialogue box for that particular search type. These include: a protein name or accession (e.g. VAV_HUMAN); a Pfam family name or accession (e.g. PF00571); a clan name or accession (e.g. ENTH_VHS); a PDB [13] accession (e.g. 2abl); or keyword(s) (e.g. ‘RNA binding’). Additional examples are given under each heading on the website.

There is an option to browse the contents of the database using the ‘browse’ tab at the top of the page (circled in red, Figure 1). Within this page, the data is arranged alphabetically by ‘family’, ‘clan’ or ‘proteome’.

There is also the option to search by keywords using the keyword search box, which can be found at the top right-hand corner of every page (see Figure 1).

---

**Figure 1** Pfam database homepage showing quick links, browse and keyword search options.
Searching and visualising data in Pfam: sequence

Searching Pfam with a protein sequence

Clicking on the ‘search’ tab at the top of the home page (circled in red, Figure 2) opens up the ‘Search Pfam’ page, with various search options listed on the left hand side. One of the most valuable uses of Pfam is to find which family (or families) a sequence matches, and thus what its potential function might be.

You can do this by pasting a FASTA [14] formatted sequence (protein or DNA) into the sequence dialogue box (indicated by top red arrow in Figure 2), and clicking ‘submit’ (circled in red at the bottom of the page in Figure 2). This compares the submitted sequence against the Pfam database using the default \texttt{E-value} [15] shown (see the bottom red arrow in Figure 2).

An example set of results are shown on the next page.
Pfam curators set a statistical cut-off, known as a gathering threshold (GA) for an entry. Sequences failing to make a statistical match above this threshold are not reported as hits. The threshold is usually conservative, so that no known false positives appear in the family. Setting your own E-value threshold is possible, and may allow the presence of more distantly related domains to be found on your sequence; however, interpreting such hits requires caution.

Viewing the results

An example of the results is shown below in Figure 3. As demonstrated below, this sequence has significant matches with six different domains/families along its length.

The Family column shows which Pfam families are found on the queried sequence. The E-values and Bit scores indicate the significance of the results. Details of the residues that fall into each family are available under the 'show' option which displays the alignment between the query and the matching HMM.
What does the graphical representation mean?

Each Pfam family is displayed with a colour along the queried sequence. The name of the domain is shown within the graphical representation. If the queried sequences matches the full domain alignment, the full domain graphic is shown with round edges. However, if the queried sequence does not match the full length of the domain HMM model, the graphical representation of the domain is shown with a jagged edge within 5 positions of the first or last position of the profile HMM. For example, the match for C2 domain highlighted in the table (Figure 3) does not fully align with the queried sequence, hence the right-hand edge of the domain image is jagged.

Comparing your sequence to family models

The alignment views show you how closely your sequence matches the family HMMI models.

In the family column:

- #HMM represents the consensus sequence (most probably of that family model);
- #MATCH shows whether a residue from your sequence is accepted or not;
- #PP (posterior probability) gives a score for the position from 1-10, where *=10 is the best;
- #SEQ shows the residues in your input sequence.
Figure 4 Alignment views, showing the matches of your sequence to the family models.

You can click on the names or the family/domain icons to redirect to the family page. On the family page, information on structure, function and relation to other families are described. Literature references are also provided at the bottom of each section, see example in the next section.

## Searching and visualising data in Pfam: families

### Searching by Pfam entry name

Family pages contain all the data relating to any particular entry. Note, we refer to this as the "family page", but this could be any entry type, including domain [4], motif [16], repeat and so on.

Here we will go through an example entry – Piwi.

On the homepage (Figure 5, below) you can enter your query (protein name, uniprot [17] ID, etc…) in the 'Jump-To box' or in the box under ‘View a Pfam [18] entry’.

Here we have used the example entry ‘Piwi’ which directs us to the family page.
Figure 5 By using the 'Jump to' box you can search Pfam for “Piwi”.

Clicking ‘Go’ takes you to the Family page.

Navigating a Pfam entry page

The family summary page is the default display. There are tabs for Wikipedia entries related to the family and to InterPro [19]. If there is a Wikipedia page for the family, it will be the default display instead of the Pfam summary page. This is because Wikipedia pages are maintained by a wider community, and consequently, tend to contain richer media, are more extensive and often more up-to-date.

As Piwi has been extensively studied, this Wikipedia entry is particularly informative (Figure 6).
Figure 6 Entry page for Piwi, showing the summary data bar and Wikipedia content.

At the top right-hand side of the family pages (see red arrow in Figure 6), there are five navigation icons representing an overview of the data that is available for the entry. The icons function as ‘quick links’ to the relevant subpages, which are also linked from the left-hand column of the page.

**Domain architectures in a Pfam entry**

Clicking on the ‘architectures’ icon at the top right, will open the ‘domain organisation’ page listed in the column on the left. This page shows the various domain architectures of the sequences included in the entry, ordered by the number of times that architecture is seen. Identifying the different domains present in proteins is crucial to understanding how they function.

Looking at the domain organisation for Piwi, it can be seen that it is almost always associated with the PAZ domain (which recognises and binds an overhang-specific small interfering RNA), and Argonaute (ArgoN, ArgoL, ArgoMid) domains immediately upstream. Argonaute proteins are described in the Wikipedia article for this entry.
**Species distribution in a Pfam entry**

Clicking on the ‘species’ icon at the top right, will open the ‘species’ page listed in the column on the left. This opens a page showing a sunburst chart of all the protein species that carry a Piwi domain [4] (see Figure 8).

There are seven individual nodes that are derived from the taxonomic lineage of each sequence, ranging from superkingdom down to species. For each node in the species tree there is a separate arc - and each arc is arranged radially, with the superkingdoms at the centre and the species around the outermost ring. The length of each arc is proportional to the number of sequences found within that species.

Segments of the tree are coloured according to their superkingdom, as explained in the controls box. Mousing over any part of the tree shows the taxonomic name and level, with both the number of sequences and the number species found at that level. See Figure 13 in ‘Getting alignment data from Pfam [18]’ for details on how to capture subsets of sequences from this tree.

Clicking on ‘tree’ tab at the top left allows the data to be viewed as a tree format instead.
In this example, the arrow shows the segment relating to human sequences in the family.

**Interactions between Pfam entries**

Information about the interactions of one Pfam entry with any others is found by clicking on the 'Interactions icon' or the left-hand tab.

Here we can see the interacting domains for Piwi (Figure 9). Interactions can be intra-molecular (with the same protein) as well as inter-molecular (between two proteins).
Searching and visualising data in Pfam: structures

Identifying structures in a Pfam entry

Viewing the structures of domains and proteins helps to understand what their function might be, and how individual residues are arranged in three-dimensional space. Often, two residues which seem distant along the linear protein sequence can be very close in the folded protein.

Clicking the ‘structures’ tab reveals the details of all the PDB [13] structures that have been determined for sequences in that family (Figure 10). There are links to other databases and to two different structure viewers. These currently use applets and you may need to modify your browser settings to allow them to work. The structure visualisation tools show the different domains within the structure.
Figure 10 Structures page showing the PDB structures that match this entry.

Click a PDB identifier to view a page that summarises all of the domains found within the structure. Links to similar PDB viewers and cross-references to other structural databases are provided also.

**Detailed structure matches for a Pfam entry**

To see the position of Piwi and the other domains within the overall structure, click on 'OpenAstexViewer' next to AGO1_HUMAN 4KRE from within the structures page.

This Astexview of PDB structure 4KRE (Figure 11) shows exactly where each domain maps to the structure that has been solved for the human argonaute protein: AGO1_HUMAN.
Figure 11 Astexview showing the mappings of domains to PDB:4KRE on AGO1-HUMAN protein.

To check the details of each domain match, click on the UniProt ID for AGO1_HUMAN, which will take you to the Protein page for that sequence.

**Individual proteins in a Pfam entry**

Each protein in the Pfam database has an individual entry page that shows all matches (Figure 12). This page gives an indication of the extent (and the scores) of the hits to each profile HMM along the protein length.

The E-values provide a good indication of how likely a match is to have occurred by chance – the closer to zero, the less likely the match is to have arisen accidentally from a search of the database.

We also show the presence of other features on the sequence, if any. These include low complexity regions, signal peptides, disulfide bridges, active-sites, transmembrane regions and disordered regions.
Figure 12 Protein page for the AGO1_HUMAN protein.

In this example the sequence link displays the sequence in FASTA format, and the structures link gives all the PDB structures solved for this protein with their mappings.

Getting data from Pfam

Getting alignment data from Pfam
Clicking on the ‘alignments’ tab in the left hand column (circled in red in Figure 13) shows the various options for viewing, formatting and downloading alignments for each Pfam entry. The various alignments available include the ‘seed’ and ‘full’ alignments, as well as alignments against representative proteomes and sequence databases. The 'seed' alignments are based on UniProtKB [6] reference proteomes.

The top panel shows which pre-calculated alignments are available to view by which methods (Jalview [20], HTML [21] or PP/heatmap [22]).

The second panel allows you to select an alignment and then format it to suit your requirements, based on the various options provided. This formatted alignment can then be either downloaded or viewed within the browser window.

The third panel shows which pre-calculated alignments are available to download, either as raw stockholm files or as a compressed Gzipped file.
**Figure 13** How to generate and download an alignment from Pfam.

It is also possible to download the raw HMM from the “curation & model” link in the left-hand column (as shown in Figure 14).

**Figure 14** How to download an HMM from Pfam.

It is possible to retrieve sequences for a particular species or subset of the taxonomy, within a Pfam entry, from the species link in the left-hand column. By selecting the required species or level in the diagram, the subset of sequences is defined. It is then possible to generate a FASTA [14] file of the sequences, or align this subset to the HMM for the entry (see red arrows at bottom right-hand side of Figure 15). These selections open in a new window.
Figure 15  How to retrieve a set of sequences from a particular taxonomic level.

Fetching larger volumes of data

- You can retrieve data from any of our webpages via the RESTful interface [23] on the help page [24].
- As an advanced user, you can view the database-schema [25] on the help page to determine which tables to download from the ftp site [26], in order to assemble a local database to construct your own MySQL [27] queries.
- For further queries you can email our pfam-help [at] ebi.ac.uk (helpdesk).

Submitting information to Pfam

We welcome feedback and suggestions on any aspect of the Pfam resource.

- To update an annotation for which a Wikipedia article exists, click the 'Edit Wikipedia article' button (Figure 16).

- If there is no Wikipedia article and you wish to update an annotation, or make other comments on an existing entry, use the 'Provide feedback' button, or send an email to pfam-help [at] ebi.ac.uk

- If you wish to submit a new family, a new alignment for an existing family or a new profile HMM please email pfam-help [at] ebi.ac.uk (the Pfam helpdesk)

with the data. There is a link to this email address at the bottom of every page.
Get help and support on Pfam

Support

- You can contact us via the pfam [18]-help [at] ebi.ac.uk (Pfam helpdesk).
- Our help [24] page can be accessed from the menu found at the top of every page.
- Also, Pfam is supported by community help platforms such as biostars [28].

References


Funding

- European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI).
- Howard Hughes Medical Institute.
- BBSRC [BB/L024136/1].
- Wellcome Trust [WT077044/Z/05/Z].
The Pfam team

For a list of the current Pfam team members, see our Pfam consortium page [30].

Collaborators

- Sean Eddy, Howard Hughes Medical Institute.
- Erik Sonnhammer, Stockholm Bioinformatics Center, Albanova, Stockholm University.

Contributors

Sara El-Gebali [1]

EMBL-EBI
Scientific Database Curator

Dr. Sara El-Gebali is a Curator for Pfam Database at the European Bioinformatics Institute (EMBL-EBI). She did her undergraduate studies in Molecular Biology at the University of Lund, Sweden before moving to Queen Mary School of Medicine and Dentistry to gain an MSc degree in Molecular Pathology and Genomics. After working in the Medical Research Institute (MRI) laboratories at the University of Dundee and at the Institute of Oncology Research in Southern Switzerland (IOSI), she joined the University of Bern, Switzerland where she completed her PhD in Biochemistry and Molecular Medicine. She worked as a Database Curator for EMBO in Heidelberg, Germany before joining the Protein Families team at EMBL-EBI in 2016.

Lorna Richardson [2]

EMBL-EBI
Scientific Curator
Lorna is a scientific curator in the sequence families team at EMBL-EBI. She is responsible for InterPro, Pfam and Genome Properties curation, as well as outreach and training. Prior to joining the sequence families team at the EBI, she worked with the eMouseAtlas project at the University of Edinburgh, curating 3D anatomy and gene expression data. She has been working in the field of biological database curation since 2001.

Penelope Coggill [31]
EMBL-EBI
Pfam annotator

Penny Coggill has been working as an annotator, creating Pfam entries for the Pfam database since 2007, when it was housed at the Sanger Institute. She gained a first degree from Cambridge University in Natural Sciences, a Masters from Aberdeen University in Animal Nutrition and has been working in the science of DNA sequencing and then the genetics of the MHC region in humans and primates since first joining the Sanger Centre in 1999.

Rob Finn [3]
EMBL-EBI
Team Leader, Sequence Families

Dr Rob Finn leads EMBL-EBI’s Sequence Families team, which is responsible for the InterPro, Pfam and Rfam data resources as well as the HMMER web searches. Rob joined EMBL-EBI in 2014 from the HHMI Janelia Research Campus in the US, where he led a group that designed fast, web-based, interactive protein sequence searches and annotations. Between 2001-2010, he was the project leader for Pfam at the Wellcome Trust Sanger Institute in the UK. Rob’s academic background is in microbiology and he holds a PhD in biochemistry from Imperial College, London.

Source URL: https://www.ebi.ac.uk/training/online/course/pfam-quick-tour

Links
[1] https://www.ebi.ac.uk/training/online/trainers/selgebali_9606
[2] https://www.ebi.ac.uk/training/online/trainers/lornar_8729