Pfam Documentation

Pfam Team

Nov 30, 2023

CONTENTS

Pfam is a large collection of protein families, each represented by multiple sequence aligments and profile hidden Markov models (HMMs).

CHAPTER

ONE

CONTENTS:

1.1 Summary

Proteins are generally comprised of one or more functional regions, commonly termed domains. The presence of different domains in varying combinations in different proteins gives rise to the diverse repertoire of proteins found in nature. Identifying the domains present in a protein can provide insights into the function of that protein.

The Pfam database is a large collection of protein domain families. Each family is represented by multiple sequence alignments and a profile hidden Markov model (HMM).

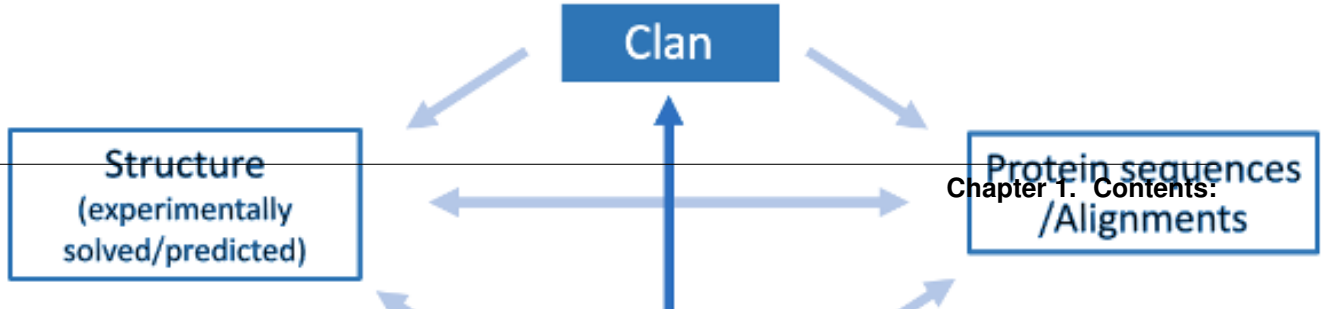
Each Pfam family, usually referred to as a Pfam-A entry, consists of a curated seed alignment containing a small set of representative members of the family, profile HMMs built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences belonging to the family, as defined by profile HMM searches of primary sequence databases.

Pfam entries are classified in one of six types:

cient to model an entire, diverse, structural superfamily and related Pfam entries are sometimes grouped together into clans; the relationship may be defined by:

ture or profile HMM.

resource by browsing by member database and choosing Pfam. For more information about InterPro you can have a look at its documentation.



the InterPro consortium.

1.3 Searching Pfam

There are multiple ways to look for information in Pfam by using the IntePro website.

1.3.1 Searching a specific Pfam entry

Users can navigate to specific Pfam entry pages by entering the Pfam identifier or accession number or a keyword that form part of its name via three different **Search boxes**:

1. When selecting the Browse + By member database option, the search box is located in the header of the results table.
2. After selecting Search + By text, a larger text box is shown in the center of the page.

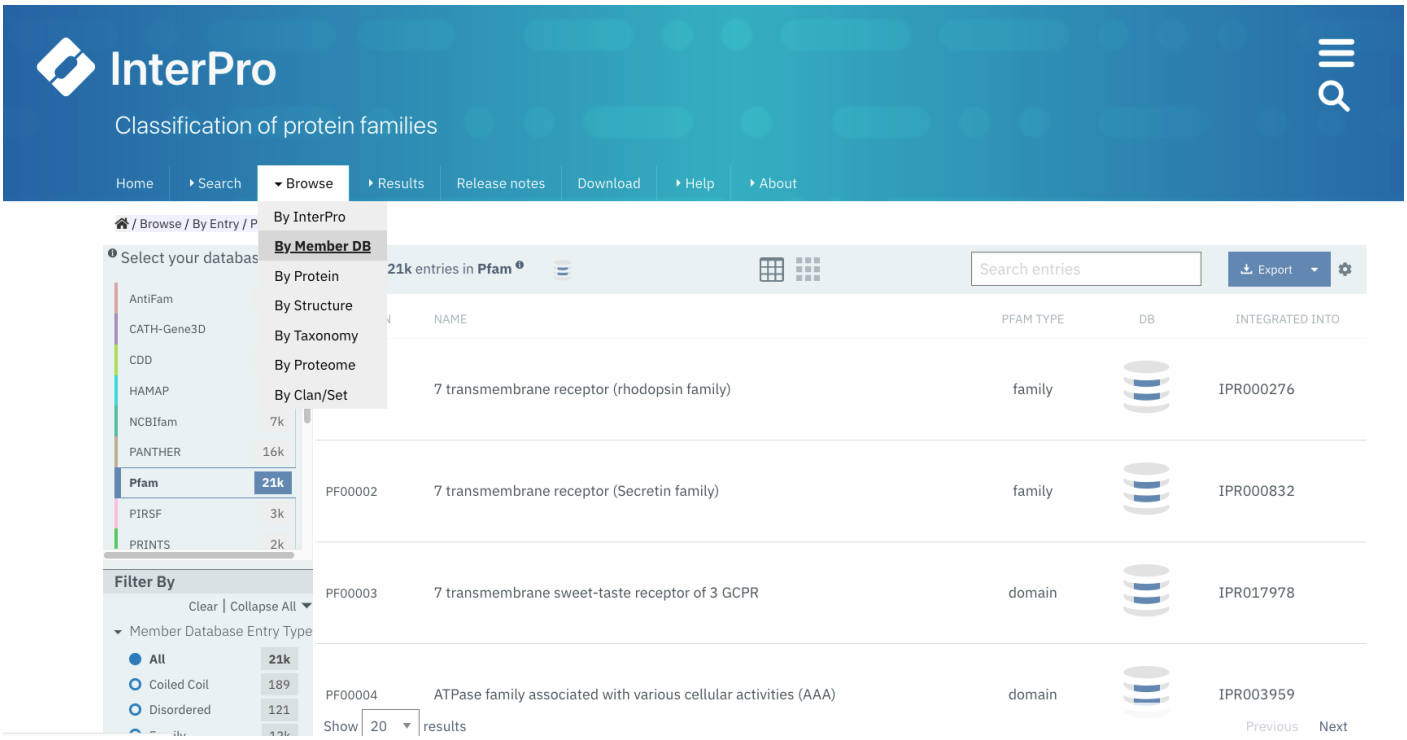


Fig. 3: Example of browsing the Pfam database. A paginated list of all available Pfam entries is displayed. A **Search box** appears on top of this list.

3. In the top right corner of any InterPro page, next to the magnifying glass.

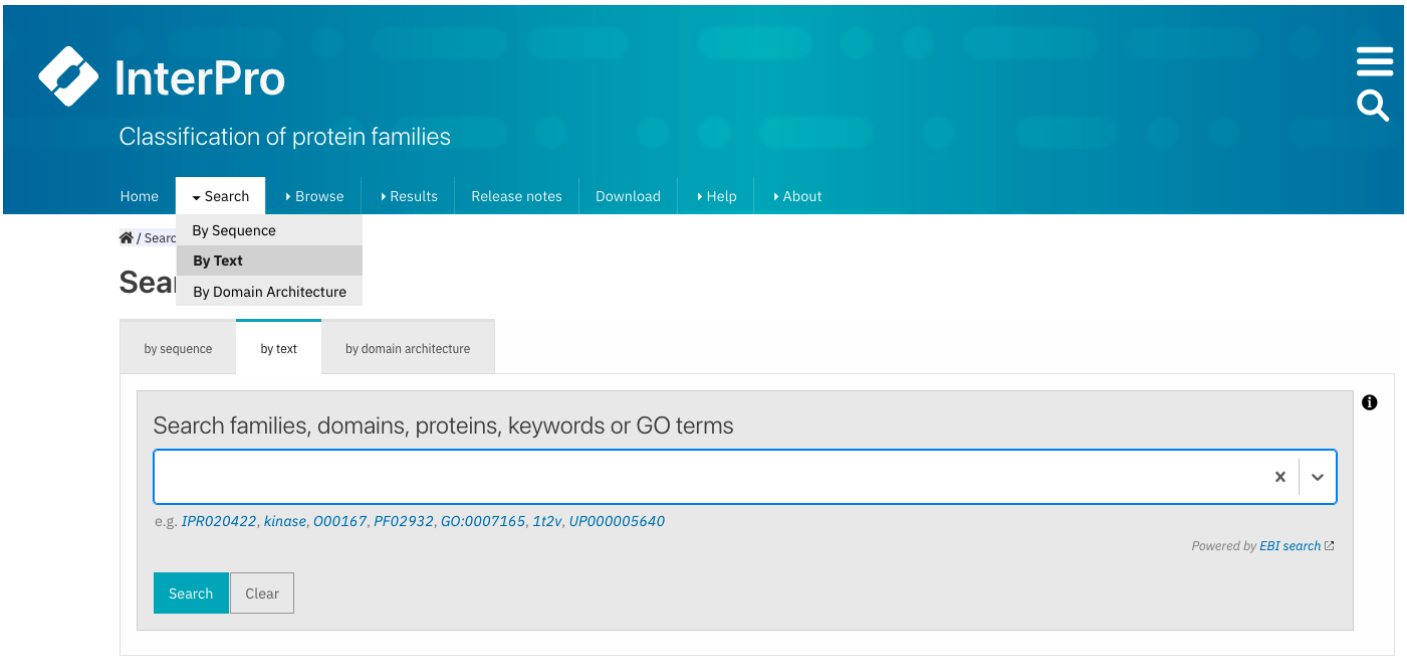


Fig. 4: Example of searching specific Pfam entry pages by entering the Pfam identifier or accession number or a keyword.

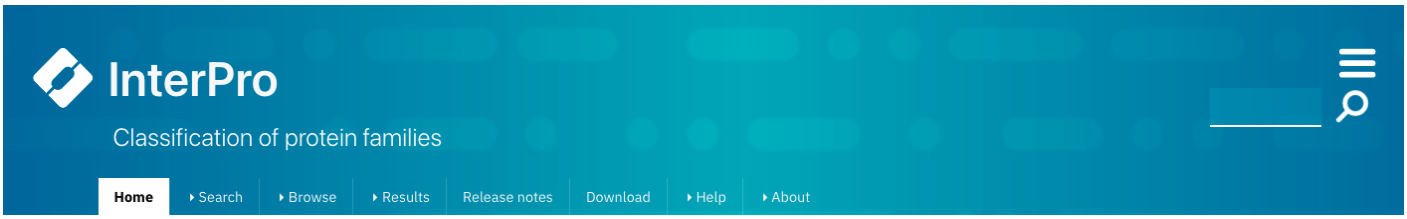


Fig. 5: On the InterPro website header, a search box appears when hovering the mouse next to the magnifying glass on the right; it can be used to search for Pfam information.

relevant page in the InterPro site, by using:

Search	Find
Pfam accession number	Pfam entry page
Pfam identifier or name	Pfam entry page
Clan identifier	Pfam Clan page
UniProt accession	IntePro protein page, which includes Pfam matches (with coordinates)
Gene names	IntePro protein page, which includes Pfam matches (with coordinates)
PDB identifier	IntePro structure page, which includes a 3D visualisation of Pfam matches
Proteomes	If it is a reference proteome, the InterPro proteome page will be displayed
Keywords, free text	List of possible matches

1.3.2 Searching a protein sequence against Pfam

Searching a protein sequence against the Pfam library of HMMs will enable you to find out the domain architecture of the protein, and thus what its potential function might be. If your protein is present UniProt version used to make the current release of InterPro, we have already calculated its domain architecture. You can access this by entering the Uniprot sequence identifier in any of the Search boxes mentioned above (see *Searching a specific Pfam entry*).

Using the InterPro online sequence search

If your sequence is not in the InterPro database, you could perform a single-sequence or a batch search against the Pfam database on the InterPro website. This search uses the web based InterProScan tool, which allows you to scan up 100 sequences at a time with a maximum length of 40,000 amino acids. To run any online search you can follow these steps:

1. Click the **Search + By Sequence** in the InterPro website menu. This opens the InterPro sequence search page.
2. Provide the FASTA formatted protein sequence(s) of interest by pasting them into the text box or by importing them from a file.

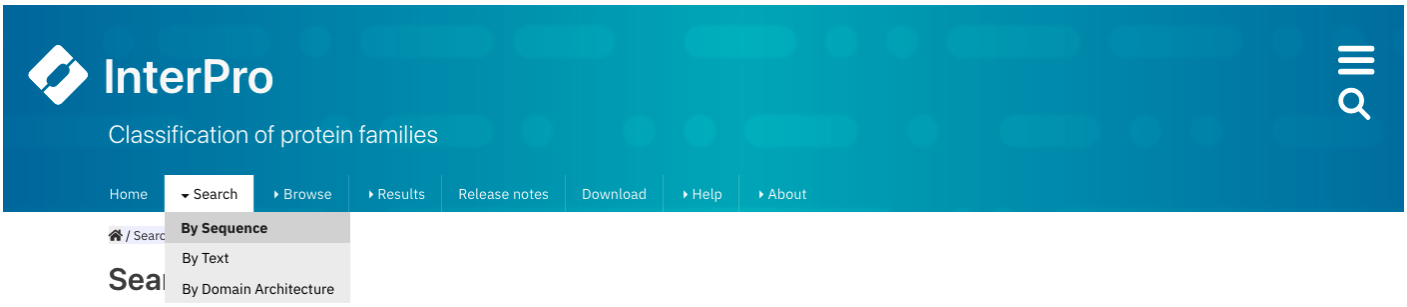


Fig. 6: Selecting **Search + By Sequence** in the InterPro website menu.

3. Expand the **Advanced options**, click on **Unselect all** protein sequence applications and select **Pfam**.

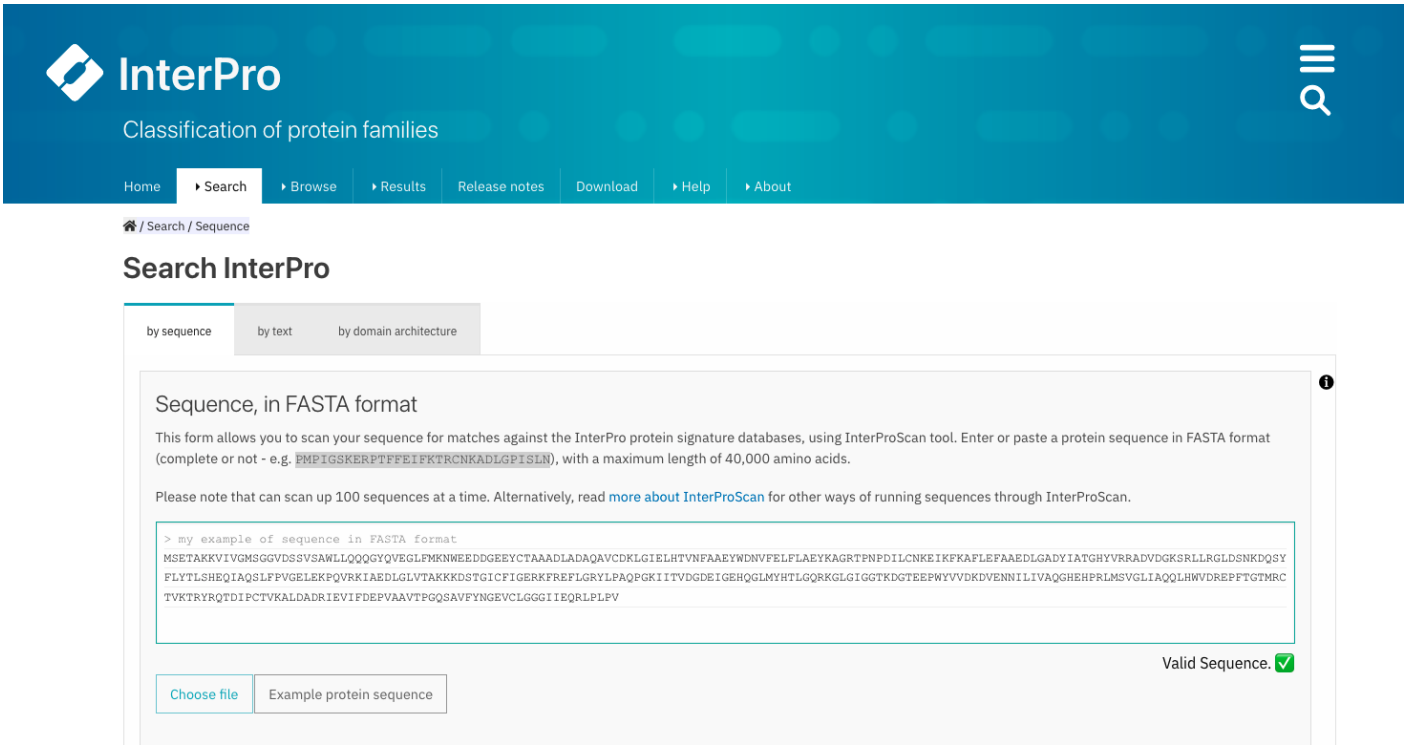
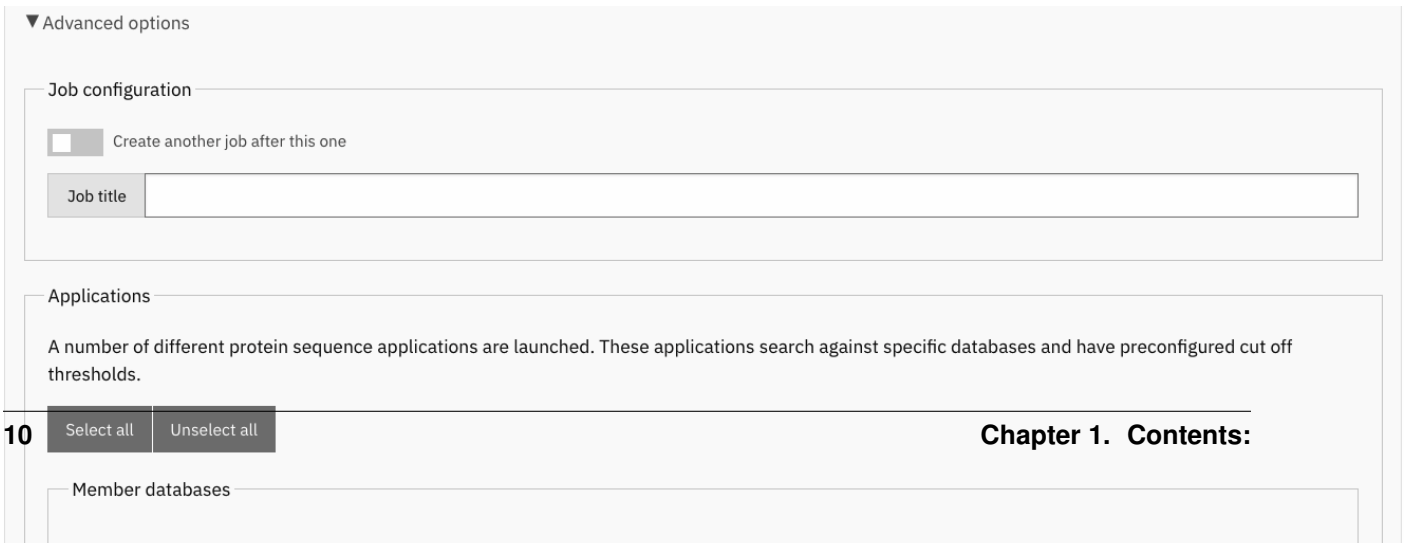


Fig. 7: Example of protein sequence in FASTA format in the text box.

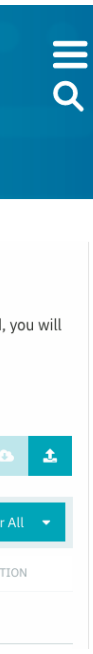
4. Click on the **Search** button.



tions and will get a pop-up notification when the job has been completed (this requires the browser notifications to be enabled).

The results of the submitted job are accessible by selecting Results + Your InterProScan Searches in the InterPro website menu.

Interpreting the protein viewer



the protein sequence viewer. The Pfam and InterPro entries are grouped by type (family, domain, repeat, site). The coloured bars indicate the location of entry matches on the protein sequence. Each matched InterPro entry is displayed on a separate line, with the Pfam entries integrated in it displayed below where relevant. The Pfam entries that remain unintegrated in InterPro entries are displayed separately in the *Unintegrated* category.

in and out of the protein sequence. The **Options** button offers the possibility to personalise the display by changing the colour code of the entries, the labels (accession number, short name and/or description can be displayed on the right-hand side of the viewer), collapsing the visualisation to show InterPro entries only or to display also the contributing entries from the member databases. The tooltip should be kept active to see a pop-up box with the accession number, description and amino acid coordinates of the match of an entry when hovering the mouse over it. Snapshots of the results can be taken in PNG format.

Local protein search

site menu. Pfam entries that the proteins should or should not contain can be included or excluded from the domain architecture. The **Order of domain matters** option offers the possibility to arrange the domains in a particular order. The **Exact match** option fine tunes the search to find only proteins containing the selected domains (no extra domain in the proteins). Domains can be selected by entering a domain name, Pfam accession or InterPro accession.

as shown in the figure below.

can be found in *Summary*. Usually, a curated description of the entry is displayed below, with the relevant literature references.

by clicking on the title.

the species name; and a small-size protein viewer displays the location of the Pfam entry in the protein. The coordinates of the match can be shown by hovering the mouse over it. You can also export this data in different formats, by clicking on the **Export** button, and customise the page settings, by clicking on the wheel icon.

architecture is seen. Identifying the different domains present in proteins is crucial to understand how they function.

in the protein. When hovering over a domain, more details are shown in a tooltip, including the domain's position.

right-hand side of the viewer. The list of proteins with this architecture is available by clicking on the protein number.

belong to.

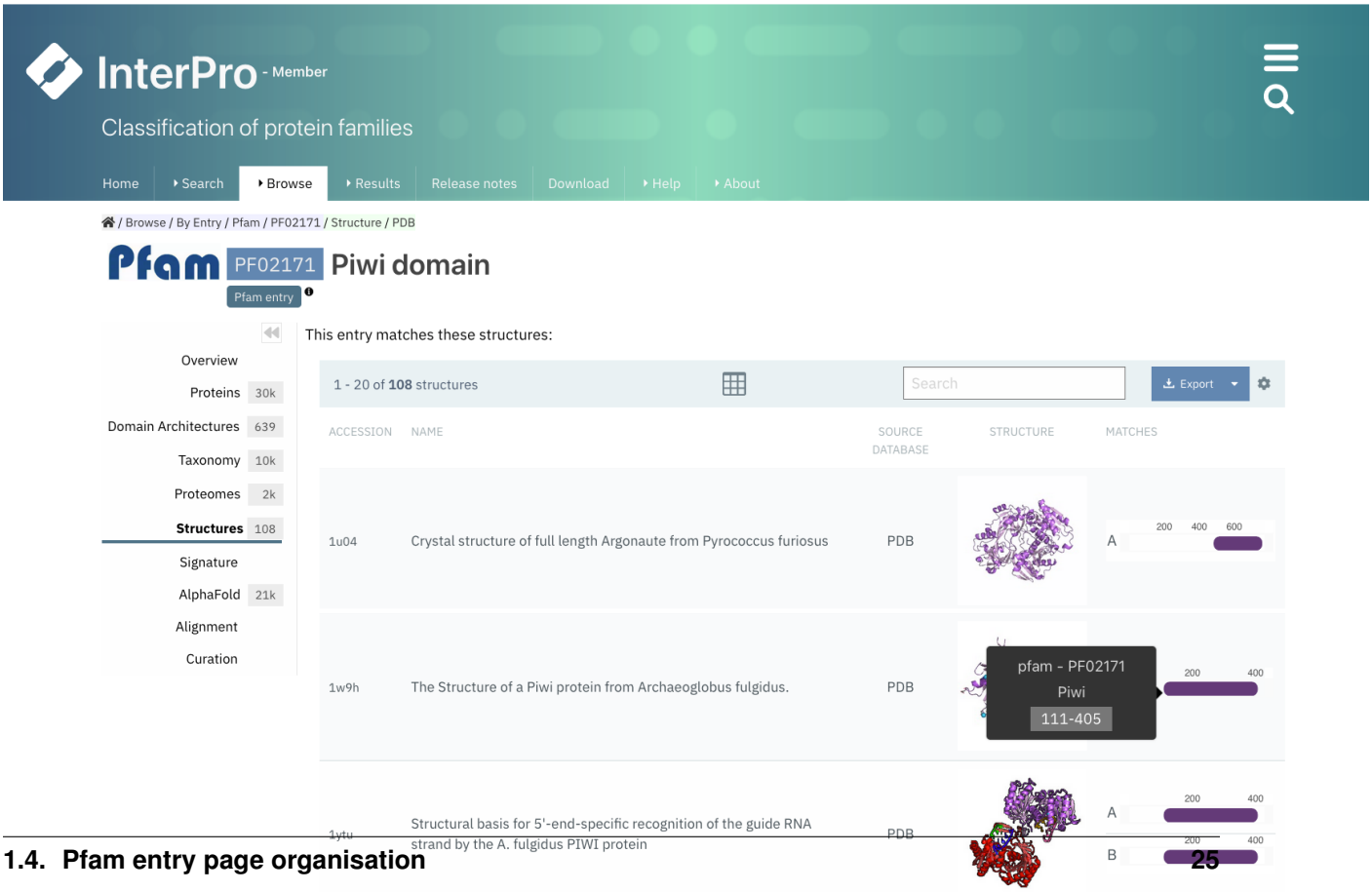
sequence, ranging from superkingdoms down to species, are displayed. For each node in the taxonomy tree there is a separate ring - and each ring is arranged radially, with the superkingdoms at the centre and the species around the outermost ring. The length of each ring is proportional to the number of proteins found within each taxon. You can choose how many rings you want to see from the options on the right-hand side of the page.

section. Mousing over any part of the sunburst chart shows the taxonomic name and level, with both the number of sequences and the number of species found at that level shown on the right-hand side.

choose to see only data from key species instead. These visualisation options can be chosen from the icon panel above the sunburst. All this information can be downloaded in different formats.

this list shows the Proteome ID (which is a link to the Proteome page in InterPro), the name of the species carrying this proteome and the number of proteins in this proteome that match the entry. From the **Actions** column, users can also see a list of these proteins by clicking the first icon (**View matching proteins**), download the data in different formats or **View proteome information**.

For each structure, you can see the PDB accession, the name of the structure in PDB, and a small-sized protein sequence viewer displaying the location of the Pfam entry in the protein structure chain.



tant along the linear protein sequence can be very close in the folded protein.

By clicking on a PDB accession, name or small image of the structure, a view of the corresponding InterPro structure page that summarises all of the entries of Pfam and other databases and resources for each chain of the structure will be displayed in a protein sequence viewer.

The position of each entry within the overall 3D structure can be visualised by choosing the Pfam entry of interest in the drop-down list **Highlight Entry in the 3D structure** or by clicking on the bar corresponding to the entry match in the protein sequence viewer. Additionally, links to similar PDB viewers and cross-references to other structural databases are provided in the **External links** section.

1.4.7 Signature

This tab shows the HMM logo of the Pfam model, visualised using Skylign. HMM logos are one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form.

The visualisation displays the amino acid conservation for each residue in the model. The rendered area can be dragged to a desired position to navigate large logos. Alternatively, a specific residue number can be written in the **Model column** text box. When selecting a particular residue in the logo, the probabilities of each amino acid are displayed in the bottom part.

dicted structures available in AlphaFoldDB for the proteins belonging to this entry is displayed in this tab. For each protein in the list, its Uniprot accession, name, the species it belongs to, its length, and a button that allows you to show the predicted structure of this protein in the structure viewer are displayed.

tab, where the position of the different entries in the 3D structure viewer are displayed by clicking on the bar corresponding to the entry match in the protein sequence viewer.

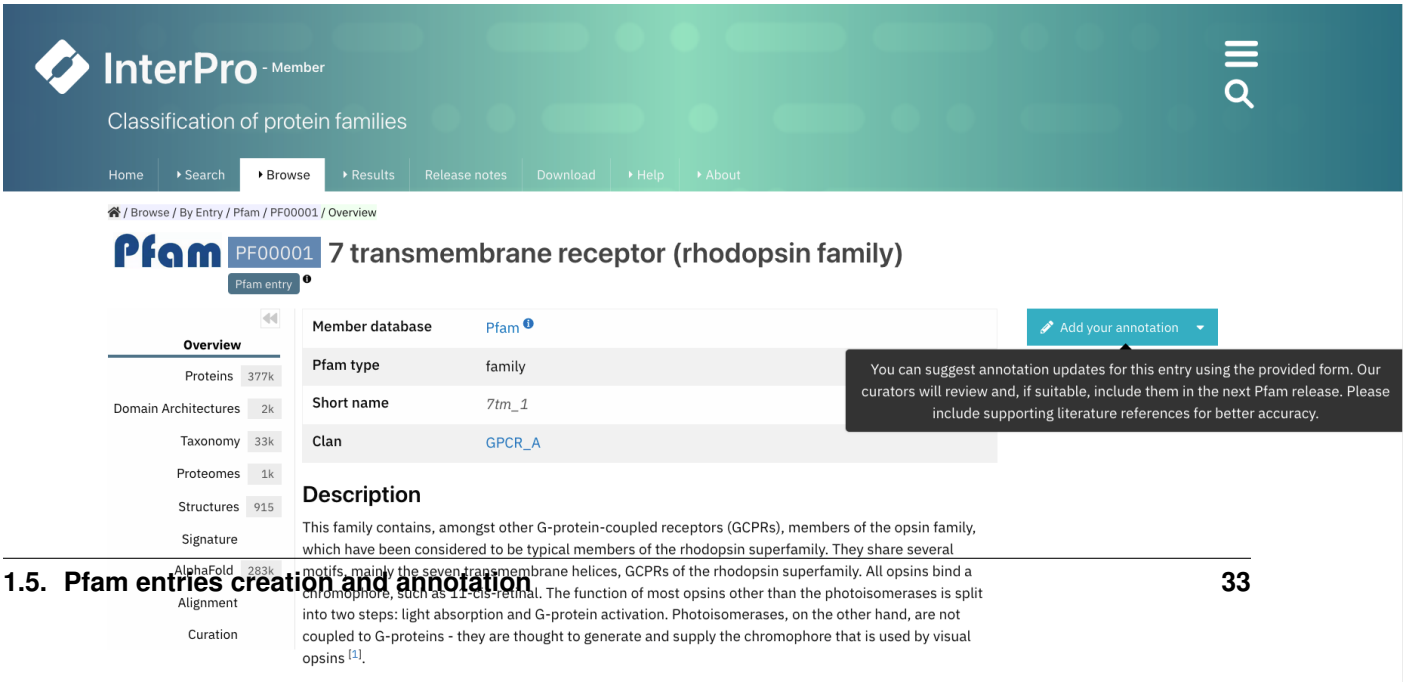
it normally has a relatively short number of protein sequences (from the Uniprot Reference proteomes).

ing the Pfam entry and offers the possibility to download it.

ing to the UniProt Reference Proteomes. Subsequently, 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 quite conservative, to minimise false positives (although they are unavoidable sometimes). The Pfam model is then run against the whole UniProtKB database before every InterPro release and these are the matches shown in the *Proteins* tab on the Pfam entry page.

mation. Many of them have a description created by Pfam curators. Anyone can contribute to this annotation by contacting directly the curators through the **Add your annotation** toolbox located on the right-hand side of the **Overview** tab.

to the Pfam helpdesk and we will endeavour to build a Pfam entry for it. We ask that you supply us with a multiple sequence alignment of the domain (please send the alignment file as a text file (e.g. .txt) and not in the format of a specific application such as Microsoft Word (e.g. a .doc) file), and associated literature evidence if available.



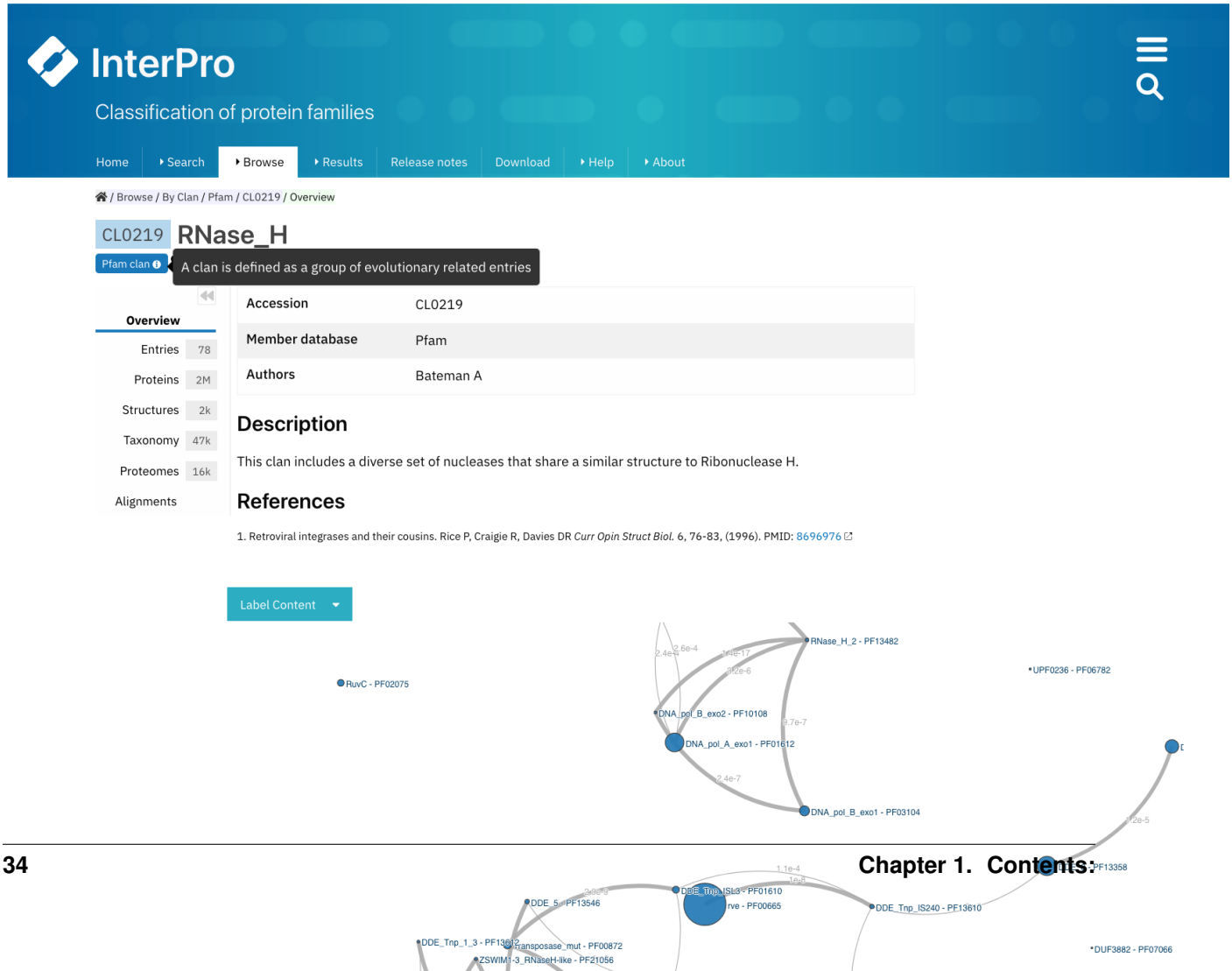
description of the Pfam entry, you may find the text from a Wikipedia article that we feel provides a good description of the Pfam family.

If a family does not yet have a Wikipedia article assigned to it, there are several ways for you to help us add one. You can find more information about the process in the *Wikipedia* section.

1.6 Clan page organisation

If a Pfam entry is included in a **Pfam clan** this information will be displayed in the **Overview** tab in the Pfam entry page, next to *Clan*, below the Pfam short name, with a link to the corresponding clan page. More information about how clans are defined can be found in *Summary*.

Additionally, it is possible to browse through the Pfam clans by selecting Browse + By Clan/Set in the InterPro website menu and select **Pfam** in the database section.



accession number, its short name and the author(s) are shown at the top. A description of the clan is displayed below, with the relevant literature references.

An interactive view of the Pfam entries included in the clan is also displayed, different label types can be chosen through the **Label Content** menu: Accession, Name and Short name.

1.6.2 Entries

The list of Pfam entries included in the clan is provided in this tab. For each entry, accession, name, short name and links to the entries SEED alignment and domain architectures pages are available.

Users can export this data in different formats, by clicking on the **Export** button, and customise the page settings, by clicking on the wheel icon.

view can be customised to show:

grated into UniProt).

the protein accession or name, and the InterPro taxonomy page can be accessed by clicking on the species name.

the page settings, by clicking on the wheel icon.

ing to the clan. For each structure, you can see the PDB accession and the name of the structure in PDB.

ing InterPro structure page that summarises all of the entries of Pfam and other databases and resources for each chain of the structure will be displayed in a protein sequence viewer.

entry of interest in the drop-down list **Highlight Entry in the 3D structure** or by clicking on the bar corresponding to the entry match in the protein sequence viewer. Additionally, links to similar PDB viewers and cross-references to other structural databases are provided in the **External links** section.

loaded in different formats.

teome page in InterPro), the name of the species carrying this proteome and the number of proteins in this proteome that match the entry are displayed. From the **Actions** column, users can also access a list of these proteins by clicking the first icon (**View matching proteins**), download the data in different formats or **View proteome information**.

each other. By clicking on each entry, users can see a small-size protein viewer showing the alignment of the related entries.

- Quick tour provides a brief introduction to the Pfam database and how to access its annotations through the InterPro website.

- Creating families provides a tutorial on how to create a Pfam entry.

- Repeats describes how repeats are represented in Pfam.

- Webinar explaining where to find Pfam annotations in the InterPro website.

- *What is Pfam?*
- *What is a Pfam entry page?*
- *What is a clan?*
- *What criteria do you use for adding families into clans?*
- *What is Pfam-N?*

den Markov models (HMMs). Each Pfam profile HMM represents a protein family or domain. By searching a protein sequence against the Pfam library of profile HMMs, you can determine which domains it carries i.e. its domain architecture. Pfam can also be used to analyse proteomes and questions of more complex domain architectures.

Pfam entry for more information on how to access them.

tion, from annotation to structure predictions of the protein members. See *Pfam entry page organisation* for a detailed description on how this data is presented.

lection of entries that have arisen from a single evolutionary origin. Evidence of their evolutionary relationship can be in the form of similarity in tertiary structures, or, when structures are not available, from common sequence motifs.

the same clan, we only show one of those matches. If the sequence region is also in the seed alignment for an entry, only the match to that entry is shown. Otherwise we show the entry that corresponds to the match with the lowest E-value.

page, or alternatively they can be accessed by by selecting Browse + By Clan/Set in the InterPro website menu and select **Pfam** in the database section.

isation for more information.

imental and predicted structures to guide us and that is always the gold standard. We also intend to harmonise this organisation with the ECOD classification. In the absence of a structure we use:

same region of the sequence

dicate a relationship

decision about where families are related and we strive to find information in the literature that support the relationship, e.g. common function.

Google Research team using deep learning approaches. You can read more about it in this initial blog post and this update. The matches for Pfam-N are displayed under the '**Other features**' section in the protein sequence viewer.

information in the Pfam database can be accessed through the InterPro website, where it is hosted. See *Getting started* for more information.

as signatures, provided by several collaborating databases (referred to as member databases). One of it 13 member databases is Pfam. For further information you can explore the InterPro About pages.

in both directions to improve protein classification.

However, it can still provide very important information about a protein of interest.

make the searches feasible, we screen the DNA for potential domains using ncbi-blast and the Pfam-A.fasta as a target library. GeneWise is then used to calculate a subset of profile HMMs against the DNA. There is some down-weighting of the bits-per-position between H2 and H3 HMMs that the conversion does not account for, leading inevitably to some false negatives for some families/sequences. However, until GeneWise is patched to deal with HMMER3 models, this is the best course of action.

those in the most up-to-date versions of the sequence databases. If your sequence isn't in Pfam, you can still find out what domains it contains by pasting it into the sequence search box (see *InterPro online sequence search* for more information).

such cases the FASTA file with the full length sequences will contain multiple copies of the same sequence.

the domain (please send the alignment file as a text file (e.g. *.txt*) and not in the format of a specific application such as Microsoft Word (e.g. a *.doc*) file) or a list of Uniprot accessions, and associated literature evidence if available.

the sequence has used a delete state in the profile HMM to jump past a match state. This means that the sequence is missing a column that the profile HMM was expecting to be there. The '.' character is used to pad gaps where one sequence in the alignment has sequence from the profile HMMs insert state. See the alignment below where both characters are used. The profile HMM states emitting each column are shown. Note that residues emitted from the Insert (I) state are in lower case.

```
FBRL_XENLA/86-131          RKVIVEPHR.HEGIFICRGK.EDALVTKNLVPGESVYGEKRISVEDGE
Q9ZSE3_EUGGR/37-85         -AVVVEPHKvHAGIFVSRGKsEDSLATLNLVPGVSVYGEKRVQTETTD
FBRL_MOUSE/90-135          KNVMVEPHR.HEGVFICRGK.EDALVTKNLVPGESVYGEKRVSISEGD
FBRL_TETTH/64-108          KTIIVK-HR.LEGVFICKGQ.QEALVTKNFFPGESVYNEKRMSVEENG
HMM STATES                 MMMMMMMMMIMMMMMMMMMMIMMMMMMMMMMMMMMMMMMMMMMMMM
```

of interest in the drop-down list **Highlight Entry in the 3D structure**.

3D structure can be visualised by hovering the mouse over the coloured bar representing the Pfam match in the protein sequence viewer.

domain, but don't have a multiple sequence alignment, we still want to know, for simple families just one sequence is enough. Again contact the Pfam helpdesk.

bines information from several of them in a single searchable resource.

in a protein.

- *Alignment coordinates*
- *Architecture*
- *Clan*
- *Domain*
- *Domain score*
- *DUF*
- *Envelope coordinates*
- *Family*
- *Full alignment*
- *Gathering threshold (GA)*
- *HMMER*
- *Hidden Markov model (HMM)*
- *Motif*
- *Noise cutoff (NC)*
- *Pfam-A*
- *Pfam-B*
- *Posterior probability*
- *Repeat*
- *Seed alignment*
- *Sequence score*
- *Trusted cutoff (TC)*

cally determined to lie, whereas the alignment coordinates delineate the region over which HMMER is confident that the alignment of the sequence to the profile HMM is correct. Our full alignments contain the envelope coordinates

from HMMER3.

This is not quite true for HMMER3.

the minimum score a sequence must attain in order to belong to the full alignment of a Pfam entry. For each Pfam profile HMM we have two GA cutoff values, a sequence cutoff and a domain cutoff.

We search our profile HMMs against the UniProt protein database to find homologous sequences.

els against the UniProtKB database. All of the sequences which score above the threshold for a Pfam entry are included in the entry's full alignment.

to the match/insert state is likely to be correct, whereas a low posterior probability indicates that there is alignment uncertainty. This is indicated on a scale with '*' being 10, the highest certainty, down to 1 being complete uncertainty. Within Pfam we display this information as a heat map view, where green residues indicate high posterior probability, and red ones indicate a lower posterior probability.

single domain, the sequence and the domains score for the protein will be identical. We use the sequence score to determine whether a sequence belongs to the full alignment of a particular Pfam entry.

hits that would be expected to have a score equal to or better than this value by chance alone. A good E-value is much less than 1. A value of 1 is what would be expected just by chance. In principle, all you need to decide on the significance of a match is the E-value.

the size of the database searched. For each Pfam family, we set a bit score gathering (GA) threshold by hand, such that all sequences scoring at or above this threshold appear in the full alignment. It works out that a bit score of 24 equates to an E-value of approximately 0.1, and a score 27 of to approximately 0.01. From the gathering threshold both a "trusted cutoff" (TC) and a "noise cutoff" (NC) are recorded automatically. The TC is the score for the next highest scoring match above the GA, and the NC is the score for the sequence next below the GA, i.e. the highest scoring sequence not included in the full alignment.

quence score" is the total score of a sequence aligned to the model (the HMM); the "domain score" is the score for a single domain — these two scores are virtually identical where only one domain is present on a sequence. Where there are multiple occurrences of the domain on a sequence any individual match may be quite weak, but the sequence score is the sum of all the individual domain scores, since finding multiple instances of a domain increases our confidence that that sequence belongs to that protein family, i.e. truly matches the model.

that the match is twice as likely to have been emitted by the model than by the Null. A bit score of 2 means that the match is 4 times as likely to have been emitted by the model than by the Null. So, a bit score of 20 means that the match is 2 to the power 20 times as likely to have been emitted by the model than by the Null.

Pfam: The protein families database in 2021 J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman *Nucleic Acids Research* (2020) doi: 10.1093/nar/gkaa913

The Pfam protein families database in 2019: S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. Finn *Nucleic Acids Research* (2019) doi: 10.1093/nar/gky995

The Pfam protein families database: towards a more sustainable future: R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman *Nucleic Acids Research* (2016) Database Issue 44:D279-D285

The Pfam protein families database: R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta *Nucleic Acids Research* (2014) Database Issue 42:D222-D230

The Pfam protein families database: M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn *Nucleic Acids Research* (2012) Database Issue 40:D290-D301

The Pfam protein families database: R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman *Nucleic Acids Research* (2010) Database Issue 38:D211-D222

The Pfam protein families database: R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, J.S. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer and A. Bateman *Nucleic Acids Research* (2008) Database Issue 36:D281-D288

Pfam: clans, web tools and services: R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman *Nucleic Acids Research* (2006) Database Issue 34:D247-D51

Enhanced protein domain discovery by using language modeling techniques from speech recognition: L. Coin, A. Bateman and R. Durbin *Proc. Natl. Acad. Sci.* USA. (2003) 100(8):4516-20

The Pfam Protein Families Database: A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, D.J. Studholme, C. Yeats and S.R. Eddy *Nucleic Acids Research* (2004) 32:D138-D141

The Pfam Protein Families Database: A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall and E.L. Sonnhammer *Nucleic Acids Research* (2002) 30(1):276-280

The Pfam Protein Families Database: A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe and E.L. Sonnhammer *Nucleic Acids Research* (2000) 28:263-266

Pfam 3.1: 1313 multiple alignments match the majority of proteins: A. Bateman, E. Birney, R. Durbin, S.R. Eddy, R.D. Finn and E.L.L. Sonnhammer *Nucleic Acids Research* (1999) 27:260-262

Pfam: multiple sequence alignments and HMM-profiles of protein domains: E.L.L. Sonnhammer, S.R. Eddy, E. Birney, A. Bateman and R. Durbin *Nucleic Acids Research* (1998) 26:320-322

Pfam: a comprehensive database of protein families based on seed alignments: E.L.L. Sonnhammer, S.R. Eddy and R. Durbin *Proteins* (1997) 28:405-420

Homology-Based Annotation of Large Protein Datasets M. Punta, J. Mistry *Data Mining Techniques for the Life Sciences. Methods in Molecular Biology* vol 1415 (2016) doi: 10.1007/978-1-4939-3572-7_8

Identifying Protein Domains with the Pfam Database P. Coggill, R.D. Finn, A. Bateman *Current Protocols in Bioinformatics* Chapter 2, Unit 2.5 (2008) doi: 10.1002/0471250953.bi0205s23

Pfam: a domain-centric method for analysing proteins and proteomes J. Mistry and R.D. Finn *Comparative Genomics. Methods in Molecular Biology* vol 396 (2007) doi: 10.1007/978-1-59745-515-2_4

Pfam: the protein families database R.D. Finn (eds M.J. Dunn, L.B. Jorde, P.F.R. Little, S. Subramaniam) *Genetics, Genomics, Proteomics and Bioinformatics*, Section 6: Protein Families (2005) doi: 10.1002/047001153X.g306303

Identifying protein domains with the Pfam database R.D. Finn, A. Bateman and S. Griffiths-Jones *Current Protocols in Bioinformatics* (2003) doi: 10.1002/0471250953.bi0205s01

a good description of the Pfam entry.

it, below the traditional Pfam annotation created by curators. Click on the title of the Wikipedia article for the full article to open in a new tab.

be assigned an article that already exists. In some cases, however, no suitable article exists, and in that case we would encourage you to consider adding one to Wikipedia yourself.

Wikipedia articles linked to Pfam families. In addition, if you come across a family that does not yet have a Wikipedia article assigned to it, we would really like to add one. If you know of an article that would provide a useful description of a family, please let us know via our annotation submission form (click the **Add your annotation** button on the family page).

Before you edit for the first time

its and contributions are more likely to be accepted (and remain) if they are in accordance with this policy.

- Five pillars
- Policies and guidelines
- Wikipedia help contents
- Wikipedia Tips
- Editing help

How your contribution will be recorded

for this entry in Wikipedia. If you are a registered user and currently logged in, your changes will be recorded under your Wikipedia user name. However, if you are not a registered user or are not logged on, your changes will be

logged under your computer's IP address. This has two main implications. Firstly, as a registered Wikipedia user your edits are more likely seen as valuable contribution (although all edits are open to community scrutiny regardless). Secondly, if you edit under an IP address you may be sharing this IP address with other users. If your IP address has previously been blocked (due to being flagged as a source of 'vandalism') your edits will also be blocked. You can find more information on this and creating a user account in Wikipedia.

Does Pfam agree with the content of the Wikipedia entry?

ticle. The Wikipedia community does monitor edits to try to ensure that (a) the quality of article annotation increases, and (b) vandalism is very quickly dealt with. However, we would like to emphasise that Pfam does not curate the Wikipedia entries and we cannot guarantee the accuracy of the information on the Wikipedia page.

Contact us

meaning. This page gives an in-depth description of the elements of the library from the *Nightingale component* and the *Domain graphic tool*.

tion of protein features using Nightingale v4.

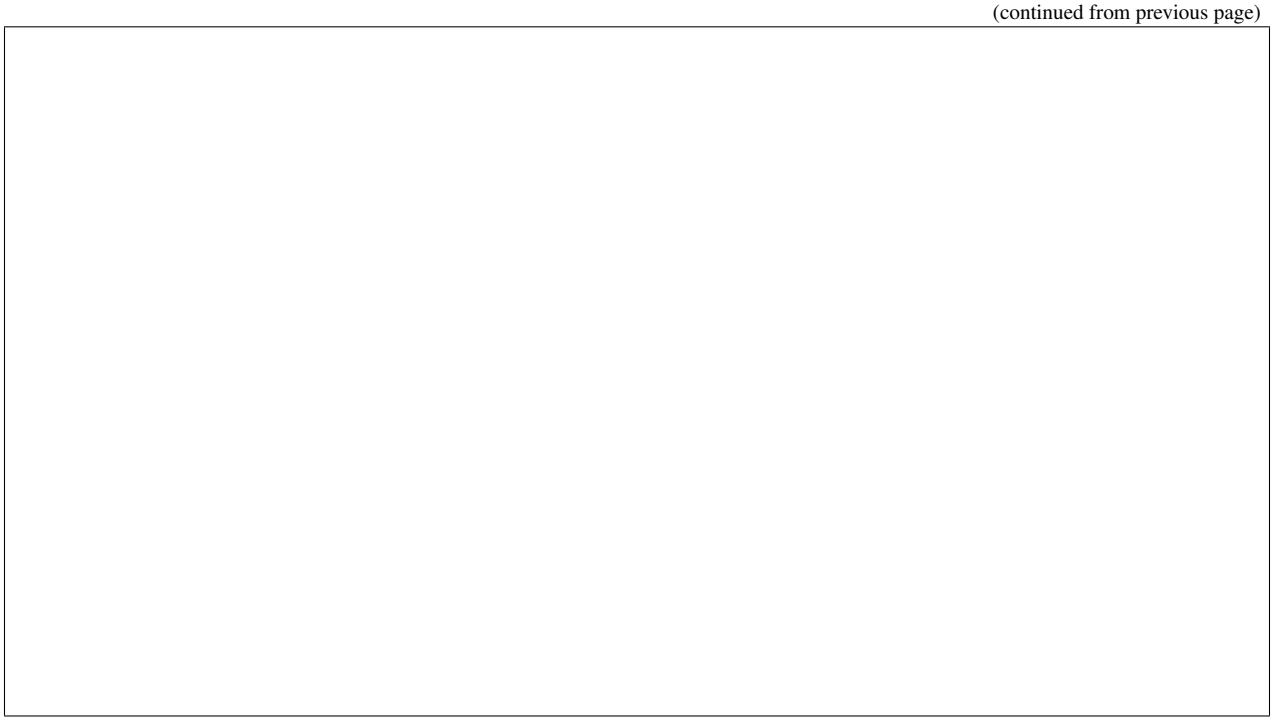
generated.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)



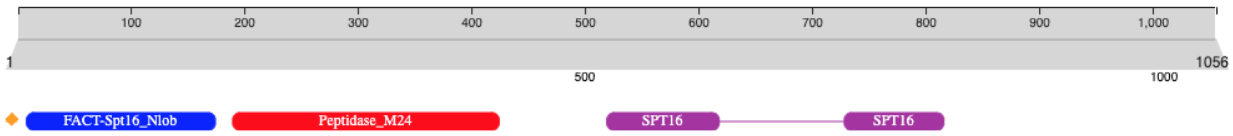
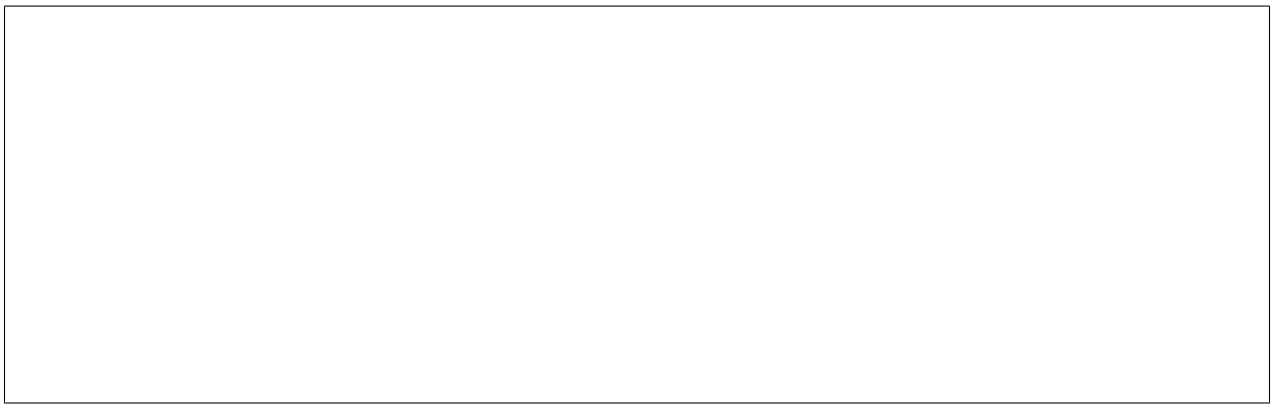


Fig. 33: Example of a domain visualisation using Nightingale v4.

ular meaning. This page gives an in-depth description of the elements of the Domain graphics library. Please note that we do not recommend to use this tool anymore, but to use the *Domain visualisation using Nightingale* instead.

The sequence

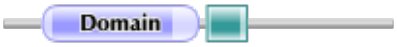


Pfam-A

cation types are rendered slightly differently.

Family/domain

enough, the domain name is shown within the domain itself. In most cases, you can click on the domains to visit the "family page" for that domain. Moving the mouse over the domain image should also display a tooltip showing the domain name, as well as the start and end positions of the domain.



(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

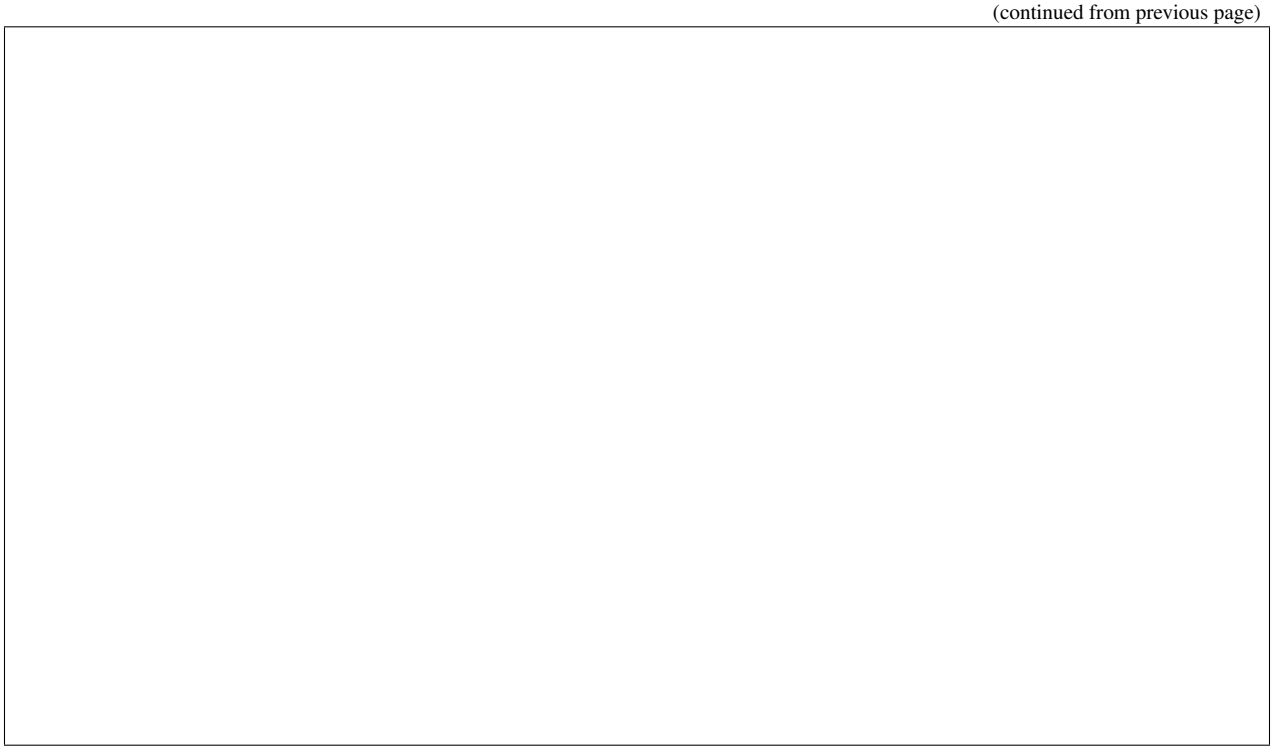
(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)



regions. The graphic above shows short envelope regions at the ends of both domains.

the domain graphic is drawn with a jagged edge instead of a curved edge. Similarly, when a sequence match does not pass through the last position of the HMM, the C-terminal side of the domain graphic is drawn with a jagged edge. In some rarer cases, the sequence match may not pass through either of the first or last positions of the HMM, in which case both sides are drawn with jagged edges. Examples of all three cases are shown below.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

Repeat/motif

gles with straight edges. As for families and domains, partial matches are represented with jagged edges.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

Discontinuous nested domains

serted or *nested* (both referring to the inner domain). For example, in many sequences containing an IMPDH domain (PF00478), the IMPDH domain is continuous along the primary sequence. However, in some cases the linear sequence of the IMPDH domain is broken by the insertion of a CBS domain (PF00571), as shown below.

nested domain is found inserted within a surface exposed loop, having little or no effect on the structure of the other domain. Such an arrangement explains why and how these nested domains can be functionally tolerated.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

Other sequence motifs

regions. These usually take lower priority than other regions that are drawn and they are therefore often obscured by, for example, a Pfam-A graphic being drawn over the top of them. An example of each motif is shown below.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

```
}
```

Signal peptides

is characterised by a short hydrophobic helix (approximately 7-15 residues). This helix is preceded by a slight positively charged region of highly variable length (approximately 1-12 residues). Between the hydrophobic helix and the cleavage site is a somewhat polar and uncharged region, of between 3 and 8 amino-acids. In InterPro, we use Phobius and SignalP for the prediction of signal peptides and they can be represented graphically by a small orange box.

Low complexity regions

are generally not well understood and are masked out to focus on globular domains within the protein.

Disordered regions

Coiled-coils

wide variety of proteins, many functionally very important. In InterPro they are obtained from COILS.

Transmembrane regions

20 amino-acids in length. Phobius and TMHMM are used for the annotation of transmembrane regions, which can be represented by a red rectangle.

Other Sequence features

above the sequence and the active site residues below the line.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

Disulphide bridges

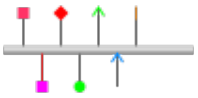
represented by a solid bridge-shaped line. When multiple disulphide bonds occur, the heights of the bridges are adjusted to avoid overlaps between them. Inter-protein disulphides are represented by single vertical lines. Moving the mouse over the "bridge graphic" shows the details of the bond in a tooltip.

Active site residues

tally determined, those that are predicted by UniProt and those predicted by Pfam. All three types can be represented by a "lollipop" with a diamond head. The head is coloured red, pink and purple for each of the three types respectively.

"Lollipops"

simple coloured bar, or as an arrow (pointing away from the sequence) or a “pointer” (an arrow pointing towards the sequence).



(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

Tooltips

necessary metadata for generating tooltips.

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

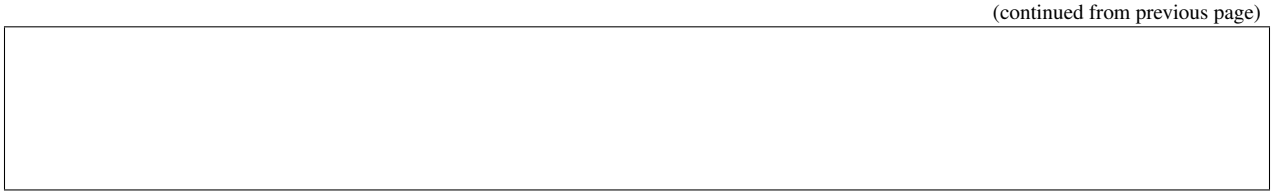
(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)



programs to access data, rather than having to rely on a browser to view a site.

URLs

b
A
su
th
se
v
p
v
th
a
R
fu
in
te
fa
ar
id
ti
fi
u
in
U

In
th
In
te
P
w
si
w
u
a
d
fe
ex
U
to
p
v
th
st
d
H
re
re
se
ta
ti
o
P
d
ar
th

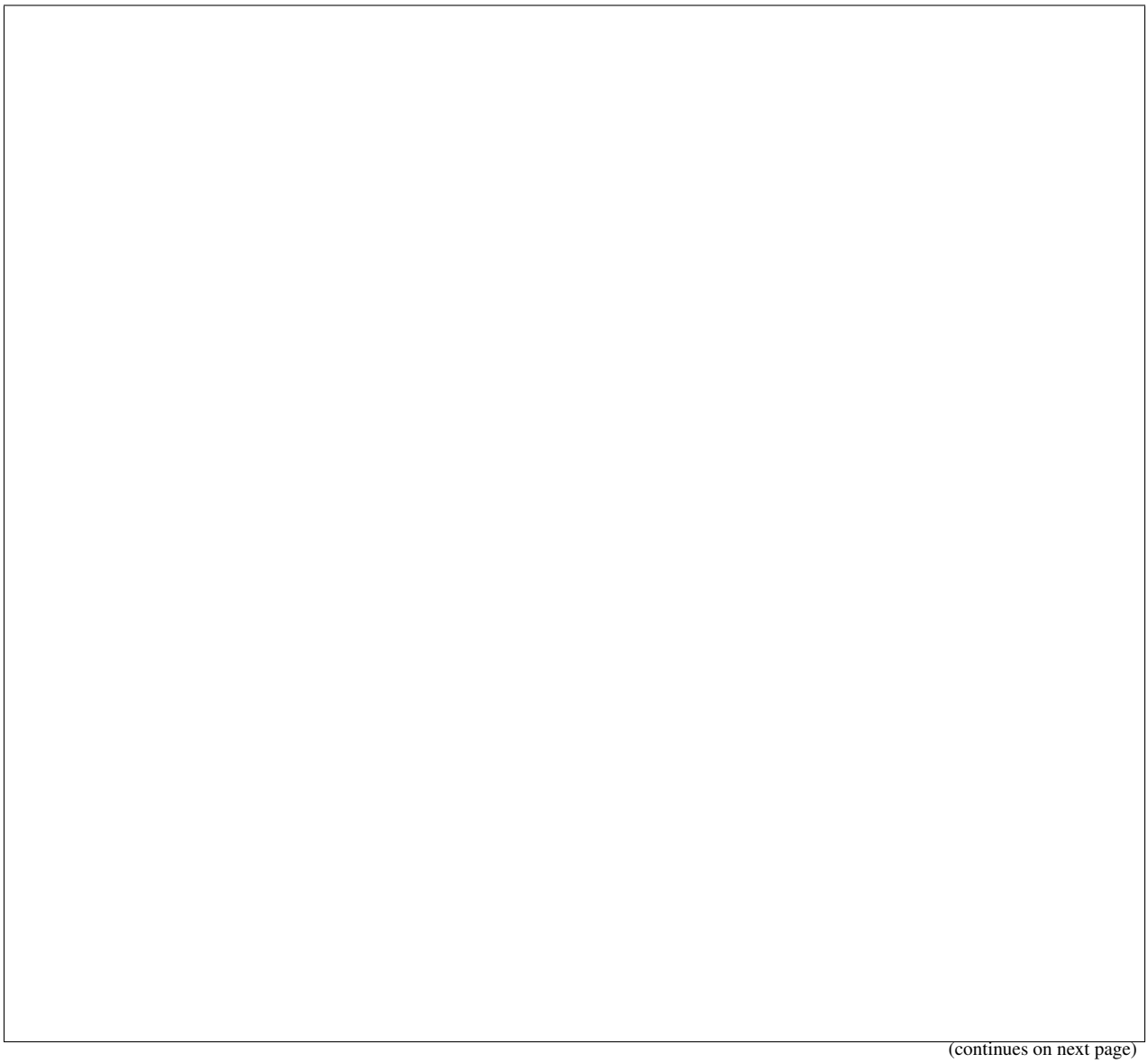
/entry/pfam/PF02171/

/api/entry/pfam/PF02171/

Data	Example website url	Example API url
List all Pfam entries	/entry/pfam/#table	/api/entry/pfam/
List all Pfam entries of type Family	/entry/integrated/pfam/?type=family#table	/api/entry/pfam/?type=family
Information about a specific Pfam entry	/entry/pfam/PF02171/	/api/entry/pfam/PF02171/
List of proteins matching a specific entry	/entry/pfam/PF02171/protein/UniProt/	/api/protein/UniProt/entry/pfam/PF02171/
Different domain architectures matching a specific entry	/entry/pfam/PF02171/domain_architecture/	/api/entry/pfam/PF02171?ida
List of PDB structures matching a specific entry	/entry/pfam/PF02171/structure/PDB/	/api/structure/PDB/entry/pfam/PF02171/
Download the PDB file of the predicted structure from RoseTTAFold	/entry/pfam/PF14331/rosettafold/	/api/entry/pfam/PF14331?model:structure
List all Pfam clans	/set/all/entry/pfam/#table	/api/set/pfam
List of Pfam entries in a specific clan	/set/pfam/CL0219/entry/pfam/	/api/entry/pfam/set/pfam/CL0219?page_size=100
General information about a specific clan	/set/pfam/CL0219/	/api/set/pfam/CL0219

Available outputs formats

Pfam-A annotations



(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

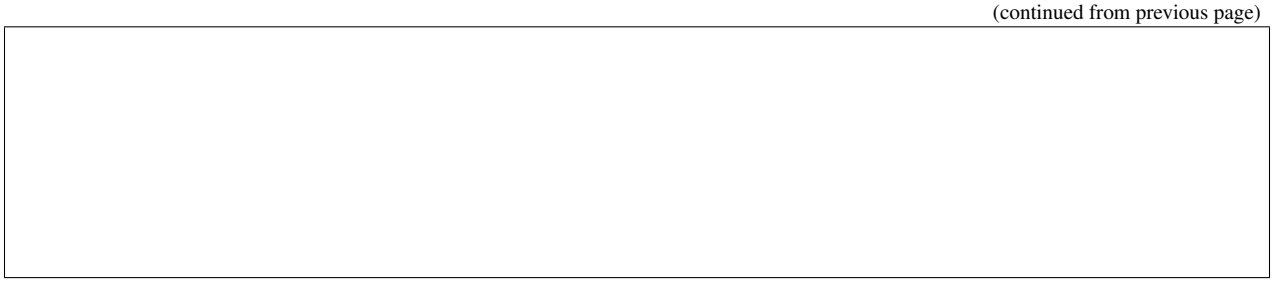
(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)



ment that only tells you that there isn't any content associated to this entry.

(continues on next page)

(continued from previous page)

Pfam-A family list

fier and description:

- /api/entry/all/pfam?format=json
- /api/entry/all/pfam?format=tsv

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

Protein data

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

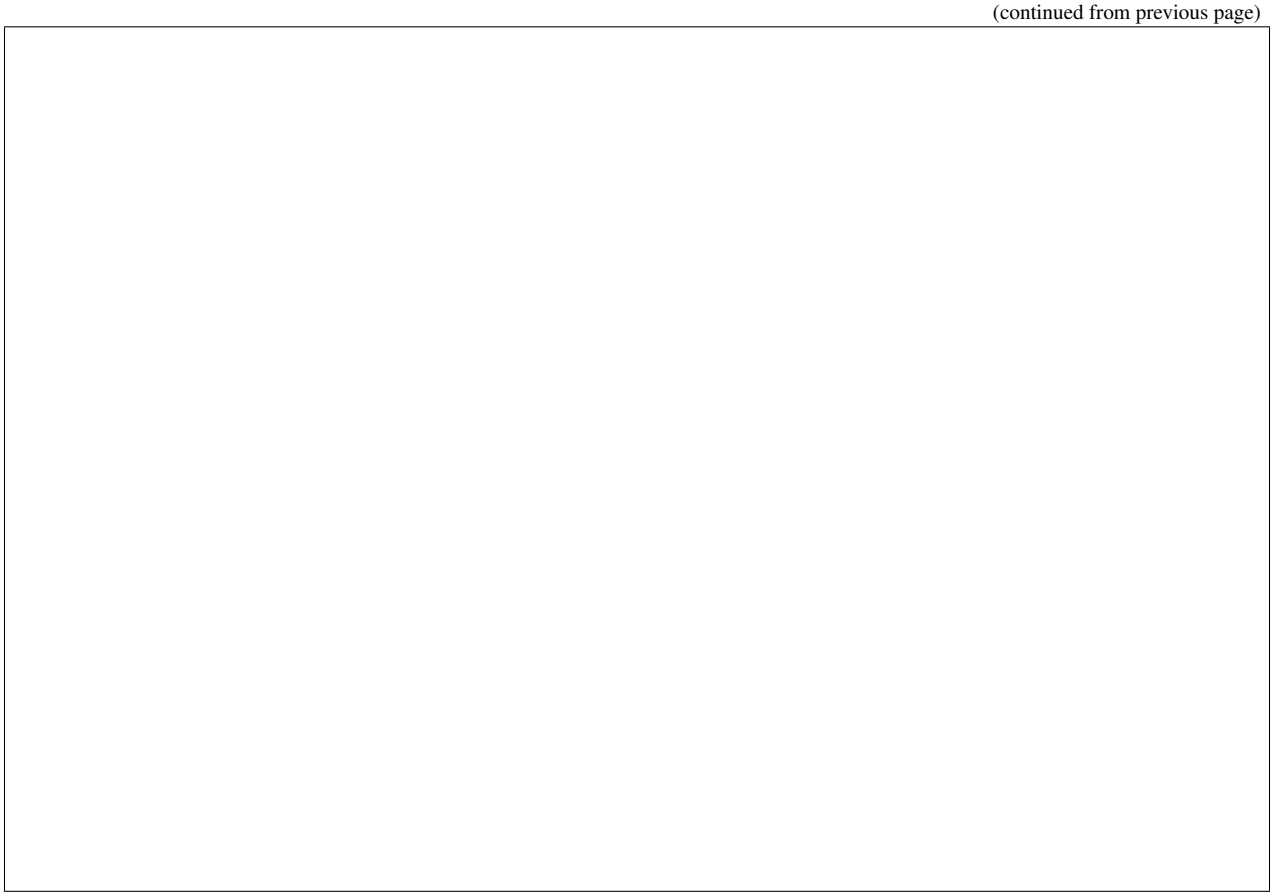
(continued from previous page)

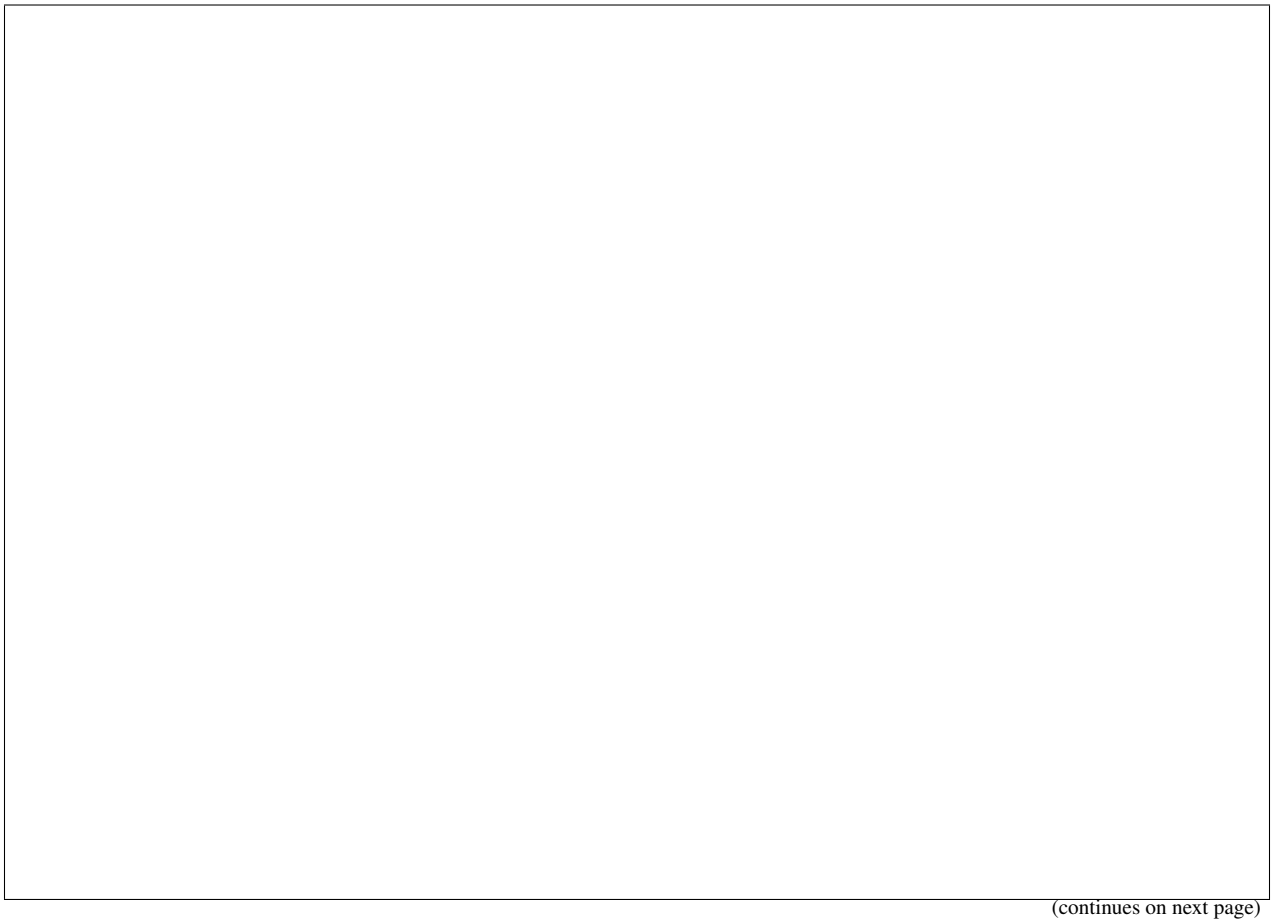
(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)





(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

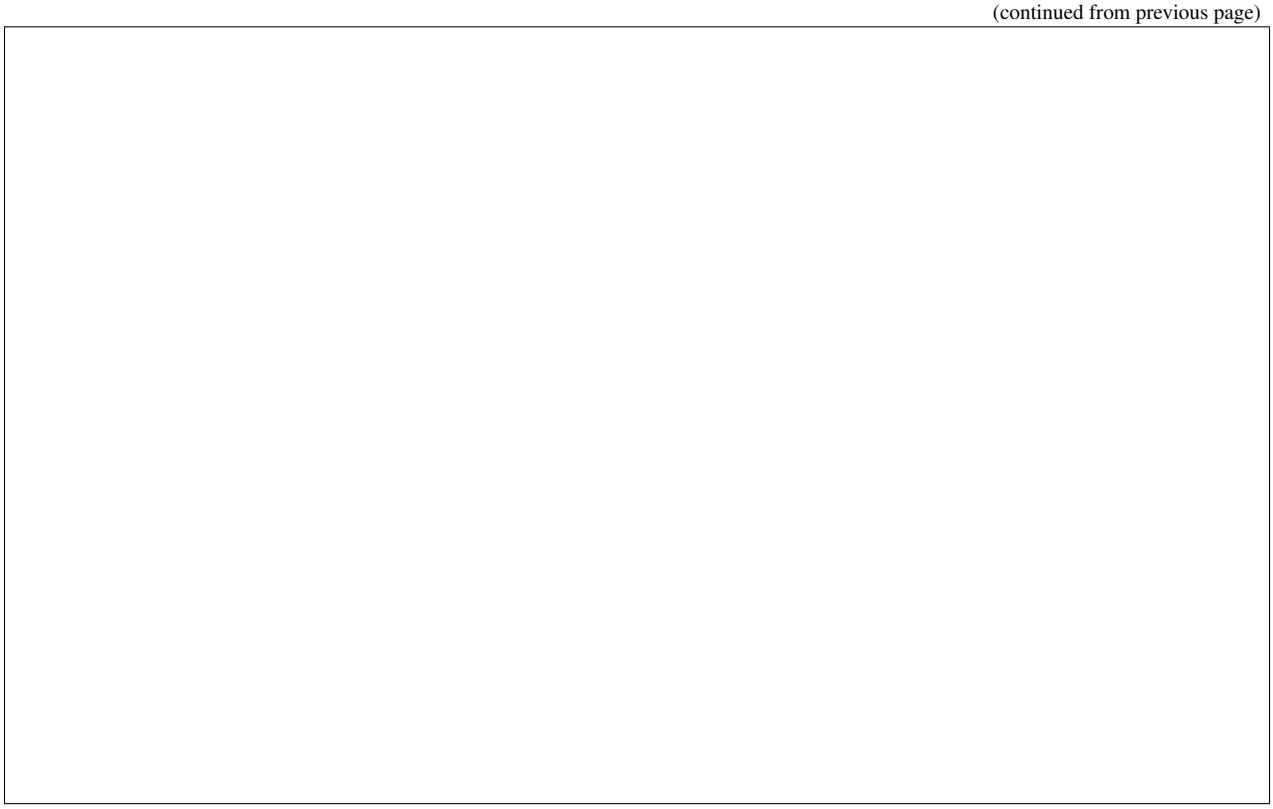
(continued from previous page)

(continues on next page)

(continued from previous page)

(continues on next page)

(continued from previous page)



- *AntiFam*
- *RoseTTAfold_aln*
- *Tools*
- *current_release*
- *mappings*
- *papers*
- *releases*

be easily incorporated into other Perl scripts. The *ChangeLog* file lists the versions and changes to the current version of *pfam_scan.pl* (and modules).

a slow connection. The files, most of which are compressed using gzip, are:

a
th
w
o
U
th

Pfam-A

T
fu
a
m
o
th
c
ra
fa
i-
li
se
a
U
th

Pfam-A

T
fu
a
m
o
th
c
ra
fa
i-
li
se
a
M
n
p
te

Pfam-A

T
fu
a
m
o
th
c
ra
fa
i-
li

vides mappings to Pfam which might be of interest.

b
sl

active_

T
b
o
d
re
q
fo
th
p
d
ti
o
ac
ti
si
b
P
sc

databa

A
ta
b
o
th
d
d
re
to

databa

D
co
ta
tw
fi
p
ta
b
fr
th
M
d
T
.s
fi
co
ta
th
ta
b

quote (‘).

M
c
s
f
e
r
l
f

metase

M
s
q
d
u
i
t
r
l

ncbi.gz

N
C
P
s
q
d
u
i
t
r
l

pdbma

M
b
t
P
s
t
a
P
d
m

pfamse

A
f
v
s
o
P
u
d
l
i

relnote
swisspf
uniprot
uniprot

n
ta
ti

userma

F
c
ta
in
in
fo
m
ti
a
th
fl
fi
fo
m

Pfam-A

A
ta
sc
a
ra
fi
c
ta
in
U
th
re
e
e
p
te
sc
q
a
P
A
fa
il
in
fo
m
ti

Pfam-A

A
ta
sc
a
ra

older releases the contents do change.

EMBL is EMBL-EBI's parent organisation. It provides core funding (staff, space, equipment) for Pfam.

The Wellcome Trust has supported Pfam since the database inception, via core funding when based at the Wellcome Trust Sanger Institute. As well as providing and maintaining the campus on which the EMBL-EBI is located, the Wellcome Trust also now provides significant funding for Pfam (grant 221320/Z/20/Z). The current grant runs from October 2020 to September 2025.

Supported by

BBSRC is supporting Pfam activities (BB/S020381/1) from November 2019 to October 2023 and has previously supported Pfam activities via grants BB/L024136/1 and BB/N00521X/1.

The Howard Hughes Medical Institute supports the Eddy group.

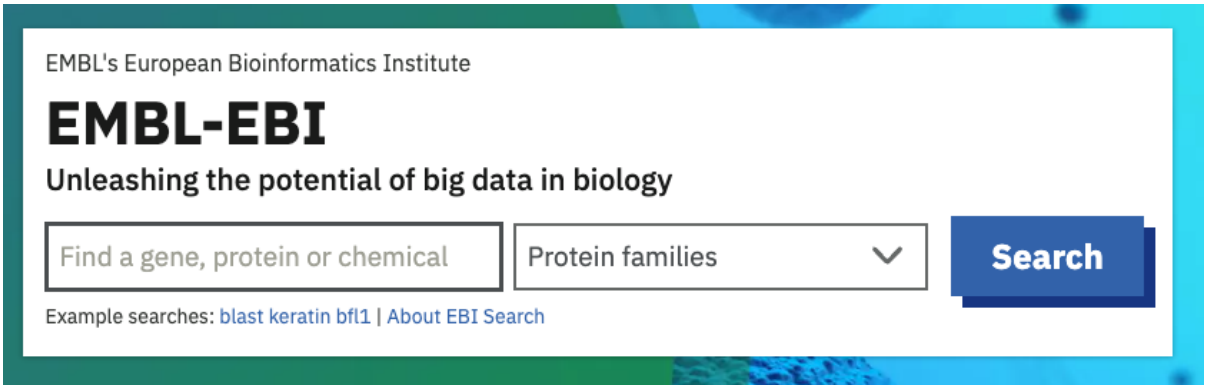
Many organisations have supported Pfam activities in the past.

For more information, please contact the Pfam helpdesk.

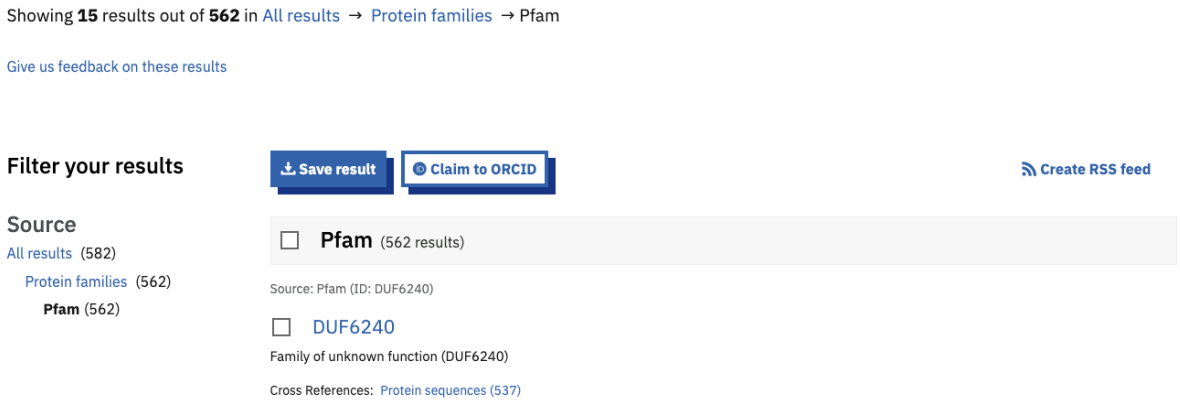
1.18 Authorship

We greatly appreciate the contribution made to Pfam from our user community. To acknowledge these contributions, and allow them to be an integral part of researchers' profiles, we have incorporated ORCID identifiers, displaying these in the 'curation and model' tab of each Pfam entry.

To claim Pfam entries against your ORCID, first go to the EMBL-EBI website and search by putting your ORCID into the search box and selecting 'Protein Families' from the drop down.



From the results page, select **Pfam** on the left-hand side and you should then see a link at the top of the results inviting you to **Claim to ORCID**. Select all the entries you want to add to your ORCID and click on the button. A pop-up window will appear, inviting you to authenticate in the ORCID website. Once you are logged-in, click on the **Claim** button.



1.19 Team Members

1.19.1 The Pfam Consortium

Pfam is the product from an international consortium of researchers that has been borne out of its original development by Erik Sonnhammer, Sean Eddy and Richard Durbin. The current list of consortium members, their institutes and primary roles are listed below.

European Bioinformatics Institute (EMBL-EBI), UK

- – Alex Bateman - Pfam team leader and head of Protein Sequence resources at EMBL-EBI
- Antonina Entcheva Andreeva - Biocurator
- Sara Chuguransky - Biocurator
- Tiago Grego - Software developer
- Beatriz Lazaro Pinto - Biocurator
- Luis Sanchez Pulido - Biocurator

Harvard University, USA

- Sean Eddy - Founding developer and author of HMMER software

Stockholm Bioinformatics Center, Sweden

- Erik Sonnhammer - Coordinator of Pfam-Sweden and founding developer

1.19.2 External contributors

Pfam includes families that have been built by external contributors:

NCBI, USA

- Lakshminarayan Iyer
- L. Aravind
- Zhang Dapeng
- Vivek Anantharaman

Sanford-Burnham Medical Research Institute, USA

- Adam Godizk

1.19.3 Previous contributors

- Gabriel Aldam
- Shimelis Assefa
- Matthew Bashton
- Ewan Birney
- Lorenzo Cerrutti
- Yuanyuan Chang
- Jody Clements
- Penny Coggill
- Lachlan Coin
- Robson De Souza
- Richard Durbin
- Ruth Eberhardt
- Sara El-Gebali
- Kyle Ellrott
- Matthew Fenech
- Kristoffer Forslund
- O. Luke Gavin
- Prasad Gunasekaran
- Sam Griffiths-Jones
- Kevin Howe
- Lukasz Jaroszewski
- Nicola Kerrison
- Marta Llagostera
- Aurélien Luciani
- Mhairi Marshall
- Nina Mian
- William Mifsud
- Jaina Mistry
- Simon Moxon
- Simon Potter
- Joanne Pollington

- Marco Punta
- Matloob Qureshi
- Lorna Richardson
- Stephen-John Sammut
- Benjamin Schuster-Böckler
- David Studholme
- John Tate
- Benjamin Vella-Briffa
- Lowri Williams
- Arthur Wuster
- Corin Yeats

Pfam is a collaborative venture and we hope to be able to interact with as many people as possible, in order to provide a quality database. Please get in touch with any one of us for more information about Pfam. You can contact us trough the Pfam helpdesk.

1.20 Contact us

1.20.1 Helpdesk

We run a helpdesk , which handles annotation comments, data enquiries and general problems with the Pfam database. We use a request tracking system to monitor emails to the helpdesk, so you should receive an automated response to your email, letting you know that the system has logged your mail and notified us of its arrival.

1.20.2 Xfam blog

The Pfam group contributes to the Xfam blog. The blog is used to announce releases, new features and important changes to Pfam, as well as for posts discussing general issues surrounding the Pfam resource. You can see blog posts that are specific to Pfam here

1.20.3 Twitter

You can follow the @PfamDB team at EMBL-EBI.

CHAPTER

TWO

LICENSE

Pfam is freely available under the Creative Commons Zero ("CC0") licence.

CHAPTER

THREE

CITING PFAM

If you use Pfam in your work, please consider citing the *Pfam References*.

CHAPTER

FOUR

GET IN TOUCH

If you have any questions or feedback, contact us through the Pfam helpdesk.