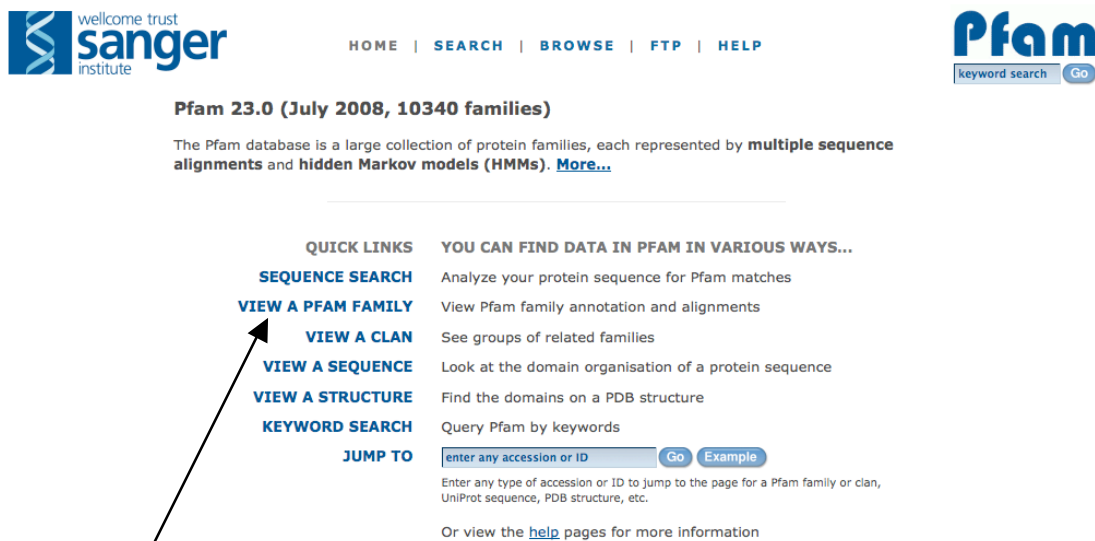


Pfam Domain Databases

Pfam is a database of protein families and domains. Currently, there are over 10,000 entries in Pfam that match to 75% of all sequences in UniProt / GenPept. Pfam can be accessed from the following locations: <http://pfam.sanger.ac.uk> and <http://janelia.sanger.ac.uk>.

In the following **worked example** you will be guided through a Pfam entry.

STEP 1 – Open the Pfam homepage at either of the two sites.



The screenshot shows the Pfam 23.0 homepage. At the top left is the Wellcome Trust Sanger Institute logo. In the center is a navigation menu with links for HOME, SEARCH, BROWSE, FTP, and HELP. At the top right is the Pfam logo with a keyword search box and a Go button. Below the navigation is the text "Pfam 23.0 (July 2008, 10340 families)" and a brief description of the database. A section titled "QUICK LINKS" lists several options: SEQUENCE SEARCH, VIEW A PFAM FAMILY, VIEW A CLAN, VIEW A SEQUENCE, VIEW A STRUCTURE, KEYWORD SEARCH, and JUMP TO. An arrow points from the "VIEW A PFAM FAMILY" link to a yellow box containing the instruction for Step 2.

wellcome trust
sanger
institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam
keyword search Go

Pfam 23.0 (July 2008, 10340 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS

- SEQUENCE SEARCH** Analyze your protein sequence for Pfam matches
- VIEW A PFAM FAMILY** View Pfam family annotation and alignments
- VIEW A CLAN** See groups of related families
- VIEW A SEQUENCE** Look at the domain organisation of a protein sequence
- VIEW A STRUCTURE** Find the domains on a PDB structure
- KEYWORD SEARCH** Query Pfam by keywords
- JUMP TO**

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

STEP 2 – Click on view a Pfam Family and entry 'RBD' in the textfield.

STEP 3 – Click on 'domain organisation'

Summary bar provides a quick synopsis on the entry

Clans are collections of related Pfam Entries.

The summary page contains a brief description of the domain and database cross references

The screenshot displays the Pfam website interface for the 'Raf-like Ras-binding domain' (PF003116). The page is organized into several sections:

- Summary Bar:** Located at the top, it provides a quick synopsis of the entry, including statistics such as 20 architectures, 167 sequences, 1 interaction, 33 species, and 6 structures.
- Navigation Menu:** On the left side, there are links for Summary, Domain organisation, Alignments, Trees, Curation & models, Species, Interactions, Structures, and a 'Jump to...' section with an 'enter ID/acc' field.
- Domain Information:** The main content area includes the domain name 'Raf-like Ras-binding domain', a brief abstract, the Interpro entry 'IPR003116', and a description of the domain's function and association with Ras-related proteins.
- Clan Membership:** A section titled 'Clan' lists the 19 members of the 'Ubiquitin' clan (CL0072), including APG12, DWNN, PI3K_rbd, ThiS, UPPF0185, CIDE-N, FERM_N, RA, ubiquitin, Urm1, DUF1315, MAP1_LG3, RBD, UBX, YukuD, DUF933, PB1, TGS, and UPPF0125.
- Gene Ontology:** This section lists biological processes like 'signal transduction' and molecular functions like 'receptor signaling protein activity'.
- Database Links:** There are sections for 'Internal database links' (e.g., similarity to PfamA) and 'External database links' (e.g., FUNSHIFT, PANDIT, SCOP, SMART, SYSTEMS).
- Example Structure:** A 3D ribbon structure of the domain is shown, with a caption indicating it is the 'PDB entry 1wfy' solution structure of the Ras-binding domain of mouse RGS14.

STEP 4 – Click on 'Alignments'

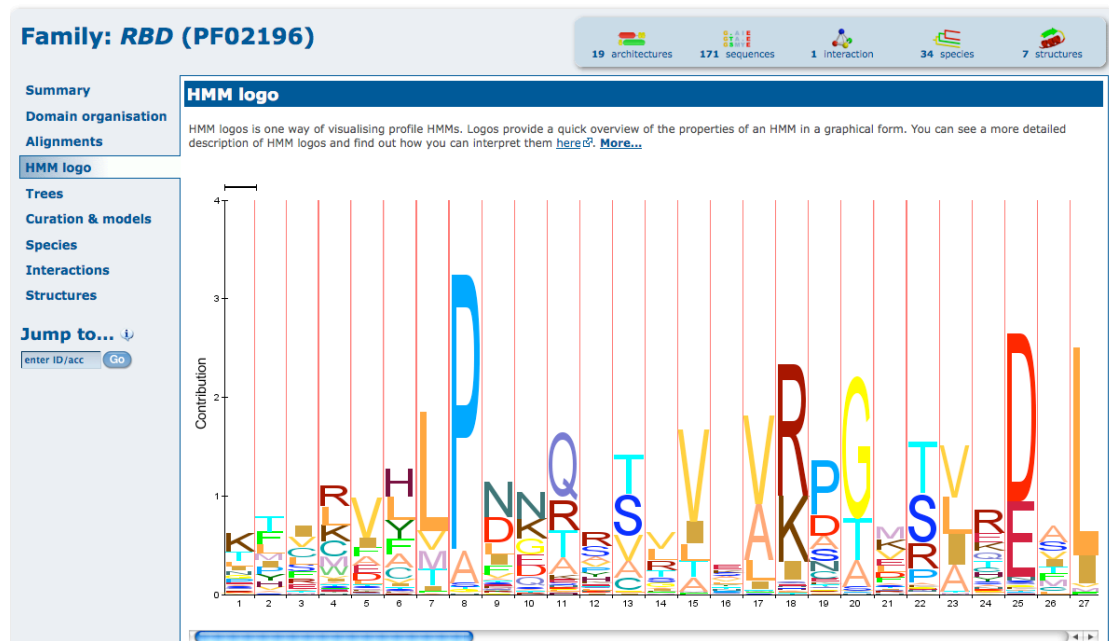
Click to reveal all sequences with that domain organisation

Solid, coloured regions are Pfam domains. Striped regions represent Pfam-Bs, which are low quality 'potential' domains.

The RBD is found associated with many different domains, many of which are involved in signalling

New HMM logo Tab

Profile HMMs are difficult to understand. To help understand them a little better, there has been the introduction of the HMM logo tab. This is a graphical representation of the HMM, where the height of the letter denotes the likelihood of that amino acid. Thus, the key residues that define the family can easily be identified.



Now back on track.....

STEP 6 – Click on 'Species'

STEP 5 – Select 'Pfam Viewer' and the 'full' alignment

Get the alignment in variety of formats

Each Pfam entry contains two alignments. The seed alignment contains a set of representative sequences that are used to build a profile HMM. The full alignment contains *all* examples of the domains.

wellcome trust sanger institute

Pfam full alignment for PF02196

Currently showing rows 1 to 30 of 167 rows in this alignment. Show rows of alignment

Click to view Protein sequence entry

Change this number to fill the screen, if you want to

Page through the alignment.

The Pfam viewer loads parts of the alignment at a time, which is then coloured according the conservation of the whole alignment

There are 6 pages in this alignment. Show page 1

Close window

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam keyword search

Family: RBD (PF02196)

20 architectures 167 sequences 1 interaction 33 species 6 structures

Summary

Domain organisation

Alignments

Trees

Duration & models

Species

Interactions

Structures

Jump to...

STEP 7 – Finally, click on ‘Structures’

Expand/Collapse the tree

Select nodes using check-boxes

Use the panel on the right to view select nodes of the tree as sequence graphics, as an alignment or as text

Tree controls Hide

Fully expand tree

Fully collapse tree

Expand to depth

Annotation

Hide highlighting of species in seed

Hide summaries

Key: species, sequences, domains

Download tree

Save a text representation

Selected sequences (uncheck all)

View

- graphically
- as an alignment

Download

- sequence accessions
- sequences in FASTA format

Family: RBD (PF02196)

20 architectures | 167 sequences | 1 interaction | 33 species | 6 structures

Structures

For those sequences which have a structure in the [Protein DataBank](#), we use the mapping between [UniProt](#), PDB and Pfam coordinate systems from the [MSD](#) group, to allow us to map Pfam domains onto UniProt three-dimensional structures. The table below shows the structures on which the **RBD** domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
ARAF_HUMAN	19 - 91	1wxm	A	8 - 80	Jmol AstexViewer SPICE
		1c1v	B	56 - 131	Jmol AstexViewer SPICE
RAF1_HUMAN	56 - 131	1qua	B	56 - 131	Jmol AstexViewer SPICE
		1rfa	n/a	56 - 131	Jmol AstexViewer SPICE
RAF1_RAT	56 - 131	1rrb	n/a	56 - 131	Jmol AstexViewer SPICE
	366 - 374	1wfv	A	5 - 16	Jmol AstexViewer SPICE
RGS14_MOUSE	376 - 446	1wfv	A	18 - 88	Jmol AstexViewer SPICE

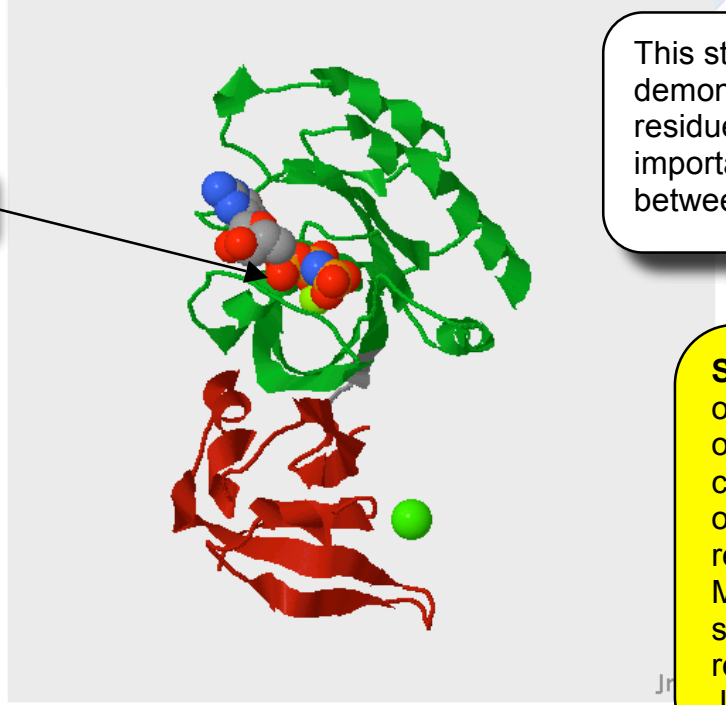
RBD domains with a known structure. Often a structure can be solved multiple times.

STEP 8 – Select 'Jmol' to view the structure

Pfam: Jmol

PDB entry 1gua

Bound ligands



This structure demonstrates the residues that are important for binding between RAS and RBD.

STEP 9 - Right click on the structure and open up the console. Left click on the structure to reveal position. Modify how the structure is represented using Jmol

PDB			UniProt			Pfam family	Color
Chain	Start	End	ID	Start	End		
A	5	167	RAP1A_HUMAN	5	167	Ras (PF00071)	Green
B	56	131	RAF1_HUMAN	56	131	RBD (PF02196)	Red

[Close window](#)

Worked Example - Searching your sequence against Pfam to identify domains. In the following example, we will analyse the sequence P14056 (<http://www.uniprot.org/uniprot/P14056.fasta>)

STEP 1 – Select Search from the menu at the top of any Pfam Page.

The gathering threshold is defined by the Pfam curators. A score at or above this threshold is trustworthy, but using E-value based cut-offs means that borderline hits are more likely to be included.

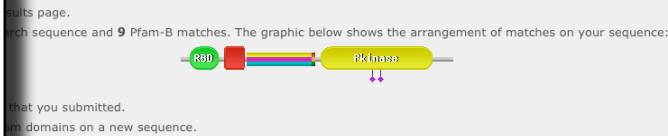
STEP 2 – Paste your sequence in the textfield*. Check the 'Search for PfamBs' checkbox and click **submit**

*The module appendix contains details of how to get the sequence used in this example



Graphical representation of trustworthy hits. Predicted active sites are shown as lollipops. PfamBs are striped regions.

STEP 3 – Click on show to reveal the alignment between the sequence and Pfam entry



Pfam-A	Description	Entry type	Sequence		HMM		Bits score	E-value	Alignment mode	Predicted active sites	Show/hide alignment
			Start	End	From	To					
RBD	Raf-like Ras-binding domain	Domain	19	91	1	77	131.4	2.9e-36	ls	n/a	Show
C1_1	Phorbol esters/diacylglycerol binding domain (C1 domain)	Domain	99	147	1	55	61.5	2.6e-16	fs	n/a	Show
Pkinase	Protein kinase domain	Domain	308	565	1	287	264.2	3e-76	ls	D427, D445	Show

List of PfamB matches

Pfam-B Matches
Show or hide all alignments.

Pfam-B	Sequence Start	Sequence End	Score	E-value	Show/hide alignment
Pfam-B_12109	148	292	739	9.1e-73	Show
Pfam-B_181853	163	233	109	7.1e-05	Show
Pfam-B_33438	163	280	129	5.1e-07	Show

Result from step 3 – revealed alignment.

```
#HMM      *->ktirvhLPnnqrsVvEvRpGmtvrDaLakalkkRGLnpsacvVrrsgdpgegekplDldtdissLpgPeElvvEnl<-*
#MATCH   t++v+LPn+qr+vV+vR+Gm+v+D+L+kalk+RGLn+++cvV+r++ +g+k++++dt+i++L+g eEl+vE+l
#SEQ     GTVKVYLPNKQRTVTVTVRDMGMSVYDSLKALKVRGLNQCVCVYRLI---KGRKTVTAWDTAIAPLDG--EELIVEVL 91
```

What does this mean? The top row represent the HMM and the most probably sequence to be emitted from it (you can think of it as a consensus sequence). The upper case letters are the important match states, the lower case letters represent insert states. The next line is the match between your query sequence and the HMM. Letter indicate a good match, where as '+' indicate similar matches. Then you have your query sequences (or at least part of it) that matches this HMM, aligned to it. These strings sequence can be punctuated with '-' charactes denoting that your sequence is missing residues compared to what is expected in the HMM (delete states) or '.' that indicate that your sequence has extra residues in it compared to what is expected (insert states).

Multiple Searches

If you have a lot of sequences to search against Pfam, rather than searching them one after the other, if you generate a fasta file containing these sequences in them, you can upload this fasta file and have the results

emailed to you. The fasta file is limited to 500 sequences at a time, but there is nothing stopping you submitting multiple files.

wellcome trust
sanger
institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam
keyword search Go

Search Pfam

0 architectures 0 sequences 0 interactions 0 species 0 structures

Batch sequence search

Upload a FASTA-format file containing multiple protein sequences to be searched for matching Pfam families. Results of the search will be returned to you at the email address that you specify. Please check the [notes](#) below for the restrictions on uploaded sequence files. [More...](#)

Sequences file

Search strategy

Cut-off Gathering threshold Use E-value

E-value

Email address

Comments or questions on the site? Send a mail to pfam-help@sanger.ac.uk
The Wellcome Trust

Similar search options to single sequence searches.

Exploring Individual Proteins Using Pfm

In the next section the use of Pfm for exploring individual proteins will be demonstrated. To use this part of the site, you must know either a UniProt accession (e.g. P00789) or identifier (PAPA1_CARPA). Although you can use NCBI genPept *gi* numbers or some metagenomics sequence accession, not all of the tools work for these alternative accessions.

STEP 1 - Go back to the Pfm home page and enter the accession P00789 into either the 'jump to' box or the 'view a sequence' page, then click 'go'

This should produce a page that looks something like this:

The screenshot shows the Pfm website interface for the protein CANX_CHICK (P00789). The page is titled "Protein: CANX_CHICK (P00789)" and includes a navigation menu with options like HOME, SEARCH, BROWSE, FTP, and HELP. The main content area is divided into several sections:

- Summary:** This section provides a brief overview of the UniProt entry. It includes the following information:
 - Description:** Calpain-1 catalytic subunit (EC 3.4.22.52) (Calpain-1 large subunit)(Calcium-activated)
 - Source organism:** *Gallus gallus* (Chicken) (NCBI taxonomy ID 9031)
 - Length:** 705 amino acids
- Pfam domains:** This section shows the arrangement of Pfam domains found on the sequence. It includes a diagram with two domains: Peptidase_C2 (green) and Calpain_III (red). Below the diagram is a table of domain boundaries:

Source	Domain	Start	End
PfamA	Peptidase_C2	48	347
		358	513

Annotations on the screenshot include:

- A yellow box labeled "STEP 2 – Click on the features tab" with an arrow pointing to the "Features" tab in the left sidebar.
- A callout box pointing to the "Summary" section, stating: "Summary of sequence information, including description, organism and length."
- A callout box pointing to the "Pfam domains" section, stating: "Representation of Pfm domains and active site residues."

Step 2 will take you to a similar graphical view of the protein, however, there will be some additional graphics shown below.

STEP 3 – Click on the show link to reveal many more sources of information

Vertical ruler tracks position

Mouse over the graphics to reveal more information

All of the data under the Pfam domain image is retrieved via DAS (distributed annotation system). There is no data duplication, so when the sources update, the information in displayed in this page is kept up to date. As new sources of protein annotations become available the list of sources they will be included in the sources listing. This feature allows users to tailor the view to the sorts of information we are interested in. For example, if we are interested in protein interactions, we can try and see if any protein interactions are known for this sequence by switching on protein interactions sources.

netphos (source)
 PDBsum_DNAbinding (source)
 PDBsum_protprot (source)
 PDBsum (source)
 OMA (source)
 PDBsum_ligands (source)
 Pfam (source)
 PDBsum (source)

STEP 4 – Click on the **PDBsum_ligands check box to see if there are any residues known to interaction with a small molecule ligand. Then scroll to the bottom of the page and click on **update**. Feel free to add other sources in.**

Protein: CANX_CHICK (P00789)

Summary | Sequence annotations

This section shows a graphical representation of this sequence, with Pfam domains shown in the standard Pfam format. Under the Pfam domain image we show various tracks, illustrating features on this sequence that we found in other databases. You can choose which databases to include using the drop-down panel under the image.

Note: It can take a few seconds for this image to be generated and loaded.

UniProt Protein Sequence (1) [Show](#)

P00789

Pfam: Peptidase_Cp, Calpain_III

PDBsum_ligands: Additional ligand interacting residues highlighted, in this case corresponding to peptidase inhibitors.

Residue number: 232

[Show](#) sources update panel.

Pfam Clans

Pfam clans are groups of related families that have arisen from a single common evolutionary ancestor. A variety of tools are used for finding related families: structural similarity, sequence similarity, functionally similarity and profile-profile comparison tools.

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam keyword search Go

Clan: Ubiquitin (CL0072)

433 architectures | 14438 sequences | 35 interactions | 1505 species | 226 structures

Summary

Ubiquitin superfamily Add annotation

This family includes proteins that share the ubiquitin fold. It currently unites four SCOP superfamilies.

This clan contains **21** families and the total number of domains in the clan is **14438**.

Members

This clan contains the following 21 member families:

APG12	CIDE-N	DUF1017	DUF1315	DWNN
FERM_N	MAP1_LC3	PB1	PI3K_rbd	RA
RBD	SLBB	TGS	ThS	ubiquitin
UBX	Ufm1	UPF0125	Urm1	YchF-GTPase_C
YukD				

External database links

CATH:	3.10.20.90
SCOP:	54236

PDB entry 2bas: UBIQUITIN-LIKE PROTEIN YUKO OF BACILLUS SUBTILIS

Comments or questions on the site? Send a mail to pfam-help@sanger.ac.uk
The Wellcome Trust

So why are they useful? Clans can provide functional insights for domains with otherwise unknown function. For example, the DUFs (domains of unknown function) in the ubiquitin clan are likely to function as small binding

domains. It also allows the identification of more distantly related structural homologs. The alignments are at the extreme edge of what can be achieved with current sequence analysis tool, but again can provide clues to key residues with the families. One can also look to see if domains are commonly combined with members of the same clan or if they are specific. There are two points of caution:

- i) Do not over interpret the transfer of knowledge
- ii) They are not currently scaling well on the website, hence the lack of screen shots