

---

---

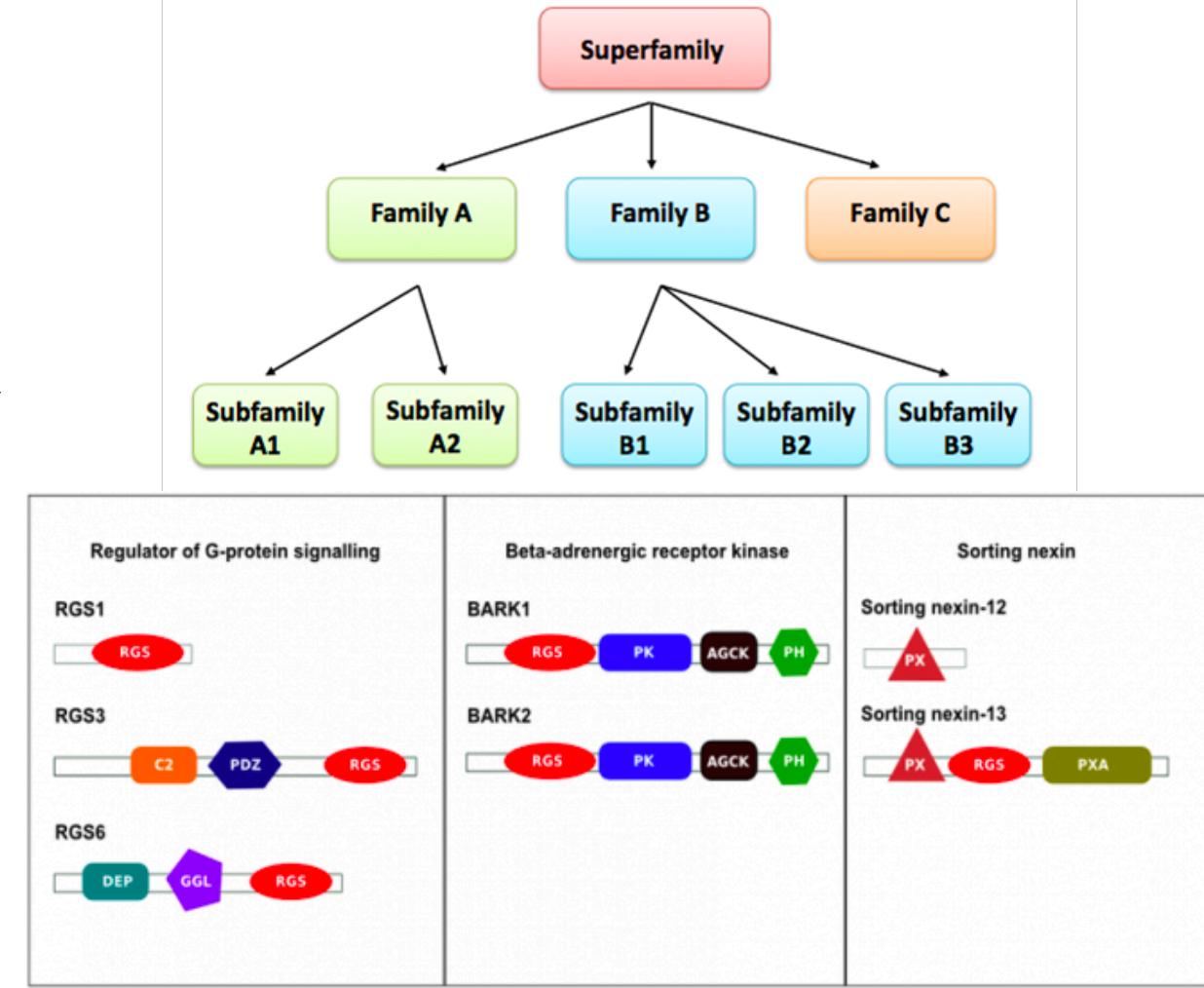
# PFAM: A PROTEIN FAMILY DATABASE

BUSHRA HAQUE

The Pfam logo consists of the word "Pfam" in a bold, blue, sans-serif font. The letters are slightly rounded and have a slight shadow effect, giving them a three-dimensional appearance.

# WHAT IS PFAM?

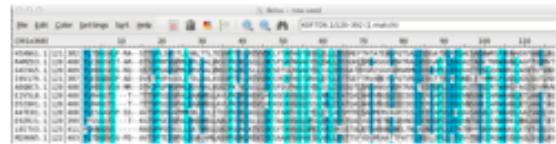
- Pfam is a database comprised of protein families and domains
- A protein family is defined to be a group of proteins that share a common evolutionary origin
- The evolutionary origin can be deduced from conserved sequences
  - Protein domains



# PFAM DATABASE ENTRY

## 1. Seed Alignment

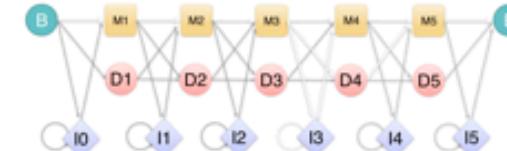
Representative Members



HMMER

## 2. Profile HMM

Seed weighted probabilities

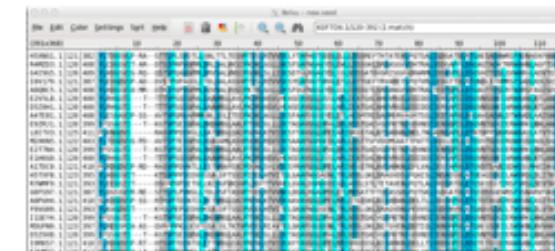


Iterate

HMMER  
against  
Reference  
Proteomes

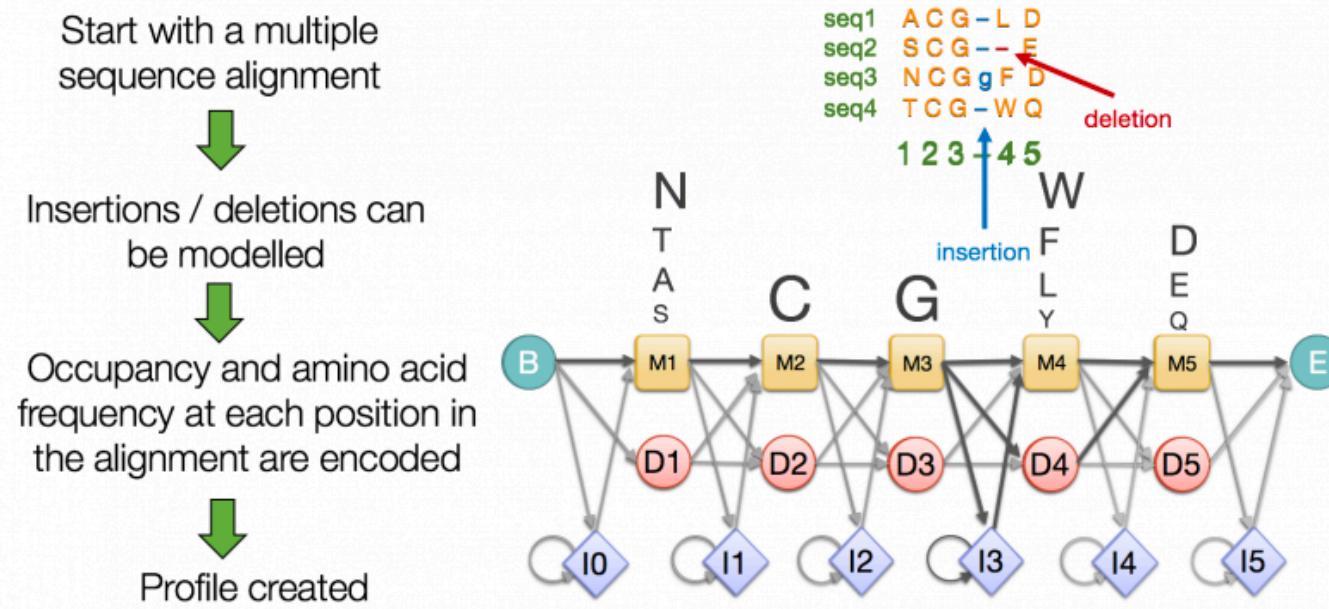
## 3. Full Alignment

All members

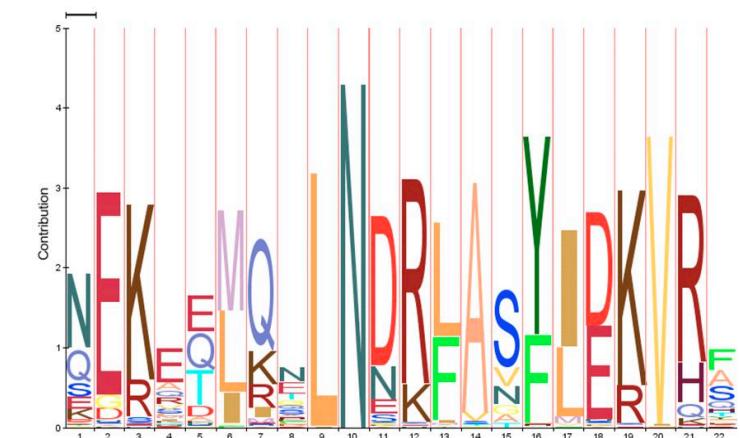


# PROFILE HMM & HMMER

- Profile HMMs use position-specific probabilistic modelling
- HMMER is a webserver and a tool used to both create profile HMMs and search against other large databases



HMM Profile Logo



# HOME PAGE

HOME | SEARCH | **BROWSE** | FTP | HELP | ABOUT

## Pfam 32.0 (September 2018, 17929 entries)

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](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

### YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches

View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Query Pfam by keywords

[Go](#)

[Example](#)

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

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

# SEARCHING WITH A PROTEIN SEQUENCE



HOME | **SEARCH** | BROWSE | FTP | HELP | ABOUT

**Pfam**  
keyword search **Go**

**Search Pfam**

**Sequence**   **Batch search** **Keyword** **Domain architecture** **Taxonomy**

**Jump to...** ⓘ **enter ID/acc** **Go**

**Sequence search**

The internal search feature on this website will be switched off soon, so we recommend you run your searches using [PfamScan](#). Alternatively, you can run your searches on the [HMMER website](#) or using [InterProScan](#). These services are actively maintained, and they provide searches against other databases in addition to Pfam.

Find Pfam families within your sequence of interest. Paste your **protein** or **DNA** sequence into the box below to have it searched for matching Pfam families. [More...](#)

**Sequence**

**Protein sequence options**

**Cut-off**  **Gathering threshold**  **Use E-value**

**E-value**  **Submit** **Reset** [Example protein sequence](#) [Example DNA sequence](#)

# SEQUENCE RESULTS



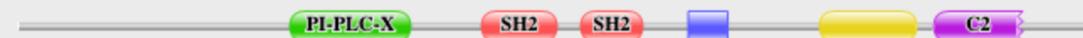
[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



## Sequence search results

[Show](#) the detailed description of this results page.

We found **6** Pfam-A matches to your search sequence (**all** significant)



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

## Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
<a href="#">PI-PLC-X</a>	Phosphatidylinositol-specific phospholipid-binding protein domain	Family	<a href="#">CL0384</a>	322	465	322	465	1	145	145	218.3	2.9e-65	n/a	<a href="#">Hide</a>
#HMM	<a href="#">lsaplhfyfissshtnyltgnqlsssesveayrnalirGcRcvEidwdgedeppvlhgyltltkikfrdvleakdyafktselPvisLevhcsaeqqekmaeilkeifgdllleeplekeeskelpspeelkgkilikvk</a>													
#MATCH	<a href="#">+++Plshy+issshtnyltg+q+ssess eay++ L++GcRc+E+d+wdg+d+ pv++hg+tlt+kikf+dvl++ik++af++se+Pvi+s+e+hcs qqq+ma+++k+++gd+Ll++p + +++ lpsp++lk+Kilik+kk</a>													
#PP	<a href="#">899*****</a>													
#SEQ	<a href="#">MNNPLSHYWISSHNTYLTGDQFSSESSLEAYARCLRMGCRCIELDCWDGPDGMPIVYHGHTLTKIKFSDVLHTIKEHAFVASEYPVILSIEDHCSIAQQRNMAQYFKVLDTLTKPVDTI AADGLPSPNQLKRKILIKHK</a>													
<a href="#">SH2</a>	SH2 domain	Domain	<a href="#">CL0541</a>	550	639	550	639	1	77	77	85.3	2.2e-24	n/a	<a href="#">Show</a>
<a href="#">SH2</a>	SH2 domain	Domain	<a href="#">CL0541</a>	668	741	668	741	1	77	77	72.5	2.3e-20	n/a	<a href="#">Show</a>
<a href="#">SH3_1</a>	SH3 domain	Domain	<a href="#">CL0010</a>	797	843	797	843	1	48	48	54.2	7.7e-15	n/a	<a href="#">Show</a>
<a href="#">PI-PLC-Y</a>	Phosphatidylinositol-specific phospholipid-binding protein domain	Family	<a href="#">CL0384</a>	953	1068	953	1067	1	114	115	133.7	3.6e-39	n/a	<a href="#">Show</a>
<a href="#">C2</a>	C2 domain	Domain	<a href="#">CL0154</a>	1088	1193	1090	1187	3	95	103	53.3	2.7e-14	n/a	<a href="#">Show</a>



Pfam is part of the ELIXIR infrastructure

Pfam is an Elixir service [Read more](#)

Comments or questions on the site? Send a mail to [pfam-help@ebi.ac.uk](mailto:pfam-help@ebi.ac.uk).

European Molecular Biology Laboratory

(El-Gebali et al., 2019)

# SEARCHING WITH PFAM ENTRY NAME



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam  
keyword search Go

## Family: Hormone\_recep (PF00104)

>Loading page components (1 remaining)...

**Summary**

**Domain organisation**

**Clan**

**Alignments**

**HMM logo**

**Trees**

**Curation & model**

**Species**

**Interactions**

**Structures**

**Jump to...** ⓘ

**Summary: Ligand-binding domain of nuclear hormone receptor**

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: Nuclear receptor](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "[Nuclear receptor](#)". [More...](#)

**Nuclear receptor** [Edit Wikipedia article](#)

In the field of **molecular biology**, **nuclear receptors** are a class of **proteins** found within cells that are responsible for sensing **steroid** and **thyroid hormones** and certain other molecules. In response, these receptors work with other proteins to regulate the **expression** of specific **genes**, thereby controlling the **development**, **homeostasis**, and **metabolism** of the organism.

Nuclear receptors have the ability to directly bind to **DNA** and regulate the expression of adjacent genes, hence these receptors are classified as **transcription factors**.<sup>[2][3]</sup> The regulation of gene expression by nuclear receptors generally only happens when a **ligand** — a molecule that affects the receptor's behavior — is present. More specifically, ligand binding to a nuclear receptor results in a **conformational** change in the receptor, which, in turn, activates the receptor, resulting in **up- or down-regulation** of gene expression.

A unique property of nuclear receptors that differentiates them from other classes of **receptors** is their ability to directly interact with and control the expression of **genomic DNA**. As a consequence, nuclear receptors play key roles in both embryonic development and adult homeostasis. As discussed below, nuclear receptors may be classified according to either **mechanism**<sup>[4][5]</sup> or **homology**.<sup>[6][7]</sup>

**Contents** [hide]

- 1 Species distribution
- 2 Ligands
- 3 Structure
- 4 Mechanism of action
  - 4.1 Type I
  - 4.2 Type II
  - 4.3 Type III
  - 4.4 Type IV
- 5 Coregulatory proteins
  - 5.1 Coactivators



Crystallographic structure of a heterodimer of the nuclear receptors PPAR- $\gamma$  (green) and RXR- $\alpha$  (cyan) bound to double stranded DNA (magenta) and two molecules of the NCOA2 coactivator (red). The PPAR- $\gamma$  antagonist GW9662 and RXR- $\alpha$  agonist retinoic acid are depicted as space-filling models (carbon = white, oxygen = red, nitrogen = blue, chlorine = green).<sup>[1]</sup>



172 architectures

## Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

### There are 9876 sequences with the following architecture: zf-C4, Hormone\_recep

[F4WYZ5\\_ACREC](#) [Acromyrmex echinatior (Panamanian leafcutter ant) (Acromyrmex octospinosus echinatior)] Transcription factor HNF-4-like protein {ECO:0000313|EMBL:EGI60649.1} (440 residues)



[Show](#) all sequences with this architecture.

### There are 2457 sequences with the following architecture: Hormone\_recep

[O45954\\_CAEEL](#) [Caenorhabditis elegans] Nuclear Hormone Receptor family {ECO:0000313|EMBL:CAA16390.2} (243 residues)



[Show](#) all sequences with this architecture.

### There are 400 sequences with the following architecture: Nuc\_recep-AF1, zf-C4, Hormone\_recep

[W5PP44\\_SHEEP](#) [Ovis aries (Sheep)] Retinoid X receptor gamma {ECO:0000313|Ensembl:ENSOARP00000012220} (463 residues)



[Show](#) all sequences with this architecture.

### There are 165 sequences with the following architecture: ERbeta\_N, zf-C4, Hormone\_recep

[W5QJ29\\_SHEEP](#) [Ovis aries (Sheep)] Estrogen receptor beta {ECO:0000313|Ensembl:ENSOARP00000022731} (527 residues)



[Show](#) all sequences with this architecture.

### There are 145 sequences with the following architecture: GCR, zf-C4, Hormone\_recep

[W5NVF6\\_SHEEP](#) [Ovis aries (Sheep)] Glucocorticoid receptor {ECO:0000313|Ensembl:ENSOARP00000002152} (781 residues)

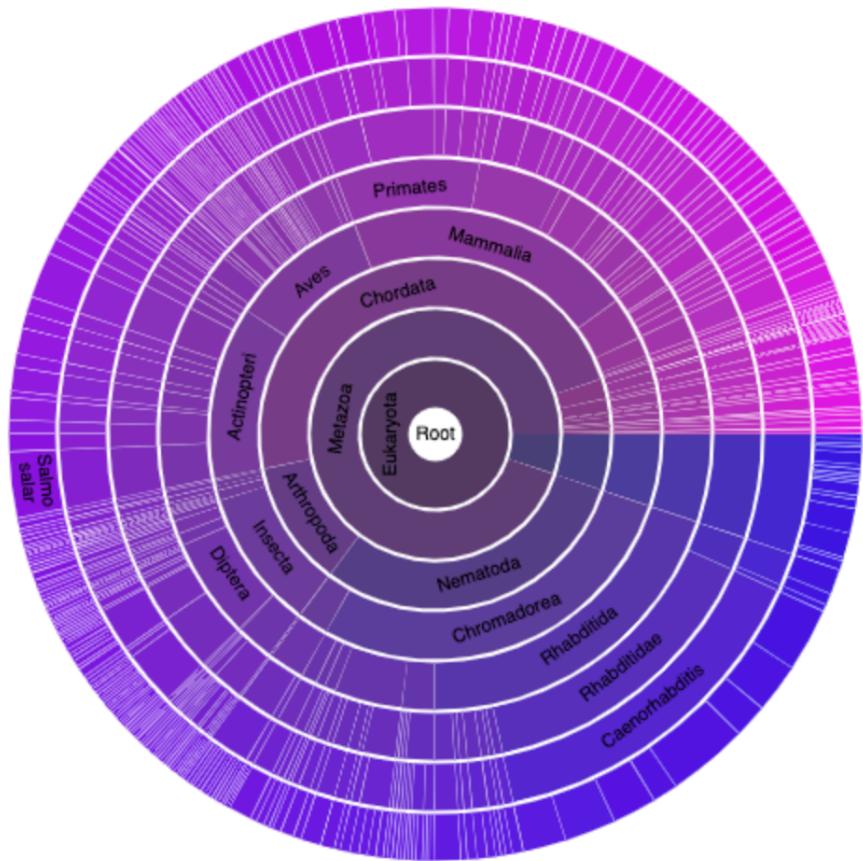




## Species distribution

Sunburst Tree

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the [adjacent tab](#). [More...](#)



### Sunburst controls

Hide

#### Xenopus

Root  
└ Eukaryota  
  └ Metazoa  
    └ Chordata  
      └ Amphibia  
        └ Anura  
          └ Pipidae  
            └ Xenopus

### Weight segments by...

- number of sequences
- number of species

### Change the size of the sunburst

Small Large

### Colour assignments

Archea	Eukaryota
Bacteria	Other sequences
Viruses	Unclassified
Viroids	Unclassified sequence

### Selections

Align selected sequences to HMM  
Generate a FASTA-format file (El-Gebali et al., 2019)



## Interactions

There are **8** interactions for this family. [More...](#)

[fn3](#)  
[Hormone recep](#)

[ARA70](#)

[zf-C4](#)

[FTZ](#)

[Androgen recep](#)

[Arm](#)

[SRC-1](#)



**Pfam is part of the ELIXIR infrastructure**

Pfam is an Elixir service [Read more](#)

Comments or questions on the site? Send a mail to [pfam-help@ebi.ac.uk](mailto:pfam-help@ebi.ac.uk).  
European Molecular Biology Laboratory

# SEARCHING & VISUALIZING STRUCTURES

## Family: *Piwi* (PF02171)

117 architectures   3833 sequences   4 interactions   764 species   96 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 [PDBe group](#), to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Piwi** domain has been found. There are 96 instances of this domain found in the PDB. Note that there may be multiple copies of the domain in a single PDB structure, since many structures contain multiple copies of the same protein sequence.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View		
A7TMA9 VANPO	888 - 1209	4F1N	A	888 - 1209	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
			B	888 - 1209	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
AGO1_ARATH	676 - 738	4G0P	A	678 - 740	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4G0Q	A	678 - 740	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4G0X	A	85 - 147	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4G0Y	A	678 - 740	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4G0Z	A	678 - 740	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		3VNA	A	678 - 741	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
AGO1_HUMAN	676 - 739	3VNB	A	678 - 741	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4KRE	A	515 - 816	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4KRF	A	515 - 816	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		4KXT	A	515 - 816	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		
		AGO2_ARATH	665 - 719	4G0M	B	665 - 719	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>
			665 - 720	4G0M	A	665 - 720	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>
	517 - 571	3LUD	B	517 - 571	<a href="#">Jmol</a> <a href="#">OpenAstexViewer</a>		

(El-Gebali et al., 2019)



astex

### PDB entry 4KRE

The applet to the left shows the structure of PDB entry 4KRE, along with the Pfam domains that map onto that structure.

The *protein backbone* is drawn as secondary structure elements, and is coloured according to the Pfam domains found on the structure. Regions of the protein that represent Pfam domains are also surrounded by a semi-transparent *molecular surface*, representing the van der Waal's surface for the atoms that fall within the domain. Regions of protein that are not assigned to a Pfam-A region are shown in grey.

Molecular surfaces can be turned on and off using the buttons in the table below. Details of the mapping between the PDB structure, UniProt sequence and Pfam domain are also given in the table.

All surfaces are initially drawn as semi-transparent membranes, but you may find it quicker to manipulate the view if the surfaces are drawn as solids. [Toggle surface transparency](#).

PDB			UniProt			Pfam family	Colour	Show/hide surface
Chain	Start	End	ID	Start	End			
A	174	224	AGO1_HUMAN	174	224	ArgoL1 (.PF08699)		Show
A	229	363	AGO1_HUMAN	229	363	PAZ (.PF02170)		Show
A	373	418	AGO1_HUMAN	373	418	ArgoL2 (.PF16488)		Show
A	427	509	AGO1_HUMAN	427	509	ArgoMid (.PF16487)		Show
A	26	164	AGO1_HUMAN	26	164	ArgoN (.PF16486)		Show
A	515	816	AGO1_HUMAN	515	816	Piwi (.PF02171)		Show
Not assigned to a Pfam-A region								

Show or Hide all surfaces

# OBTAINING ALIGNMENT DATA

- Viewing, formatting and downloading sequence alignments from ‘alignment’ tab
- Download raw HMM profile from ‘curation & model’ tab
- Retrieve sequence of a particular species from ‘species’ tab

**Summary**

**Domain organisation**

**Clan**

**Alignments**

**HMM logo**

**Trees**

**Curation & model**

**Species**

**Interactions**

**Structures**

# IMPORTANCE OF PFAM

- Our ability to generate sequence data exceeds the rate at which we can characterize sequences through experiments
- Computation methods are needed to identify regions of similarity between sequences
- Contains a wide variety of tools and resources
- Serves as a platform for researchers in the scientific community to collaborate

# REFERENCES

- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(Database issue), D427–D432. <https://doi.org/10.1093/nar/gky995>
- European Molecular Biology Laboratory. (2011a, November 18). Family- and domain-based protein classification. Retrieved September 14, 2019, from EMBL-EBI Train online website: <https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/family-and-domain-based-protei>
- European Molecular Biology Laboratory. (2011b, November 18). What are protein families? Retrieved September 14, 2019, from EMBL-EBI Train online website: <https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-families>
- European Molecular Biology Laboratory. (2014, December 4). What is the Pfam resource? Retrieved September 14, 2019, from EMBL-EBI Train Online website: <https://www.ebi.ac.uk/training/online/course/what-pfam>
- European Molecular Biology Laboratory. (2018a, March 26). Modelling in Pfam. Retrieved September 14, 2019, from EMBL-EBI Train online website: <https://www.ebi.ac.uk/training/online/course/pfam-database-creating-protein-families/modelling-pfam>
- European Molecular Biology Laboratory. (2018b, March 26). What are profile hidden Markov models (HMMs). Retrieved September 14, 2019, from EMBL-EBI Train online website: <https://www.ebi.ac.uk/training/online/course/pfam-database-creating-protein-families/what-are-profile-hidden-markov-models-hmms>
- Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., ... Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research*, 36(Database issue), D281–D288. <https://doi.org/10.1093/nar/gkm960>
- Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1), W200–W204. <https://doi.org/10.1093/nar/gky448>
- Sonnhammer, E. L. L., Eddy, S. R., Birney, E., Bateman, A., & Durbin, R. (1998). Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, 26(1), 320–322. <https://doi.org/10.1093/nar/26.1.320>
- Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., ... Finn, R. D. (2013). Dfam: A database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*, 41(Database issue), D70–D82. <https://doi.org/10.1093/nar/gks1265>



<https://create.kahoot.it/share/pfam/b0b59e20-370a-4a64-92ed-a23056eb9675>

**KAHOOT**

# PFAM DATABASE ENTRY

FILE	FUNCTION
Seed alignment	<ul style="list-style-type: none"><li>Manually verified multiple alignment of a representative set of sequences of the family</li></ul>
Profile hidden Markov model (HMM)	<ul style="list-style-type: none"><li>Built from the seed alignment for database searching and alignment purposes</li></ul>
Full alignment	<ul style="list-style-type: none"><li>Generated automatically from the seed profile HMM by searching large protein sequence database and aligning them to the HMM-profile</li><li>Contains all detectable protein sequences belong to the family</li></ul>

- Then classified into one of six types: *family, domain, motif, repeat, coiled coil or disordered*