

## Module 8: Proteins, Complexes and Pathways

### Aims

- Introduce protein sequence and protein domain databases
- Perform homology searches to help elucidate protein function
- Access and interpret protein structures and complexes
- Perform basic homology modelling
- Pathway databases

### Introduction

Protein entries found in database such as UniProt, Ensembl and RefSeq can provide information about function. UniProt represents the most comprehensive source of protein sequences. Despite this, only a relative few sequences have been experimentally characterised. Homology searches allow the identification of similar sequences, and consequently allow the transfer of annotation from one sequence to another. Nevertheless, such pairwise searches have limitation. An alternative approach to understanding protein function is the studying if sets of related sequences to identify regions of similarity (which may correspond to domains). It is well known that proteins are usually comprised of one or more globular domains. As these domains are independent units, they can be combined in different ways to give rise to functional diversity. Identification of functional domains on a protein of unknown function can enable the potential function to be postulate.

Only the elucidation of the 3 dimensional (3D) structure of a protein can allow the precise molecular mechanism of catalysis and/or function to be understood. The primary protein structure deposition database (PDB) and two associated databases will be introduced. These resources highlight many functional features found in protein structures, including protein interactions. Within the cell, proteins and protein domains are in contact with each other in order to carry out their function, e.g. signal relay or catalysis. Knowledge of protein interactions allows the understanding of the role of proteins in larger networks and pathways.

In the second half of this module, the focus will shift to disease related resources that catalogue both phenotypic and genetic effects of the disease. Tools for understanding/elucidating the possible consequence of a mutated amino acid will be covered. Finally, resources from the emerging field of functional non-coding RNAs will be covered.

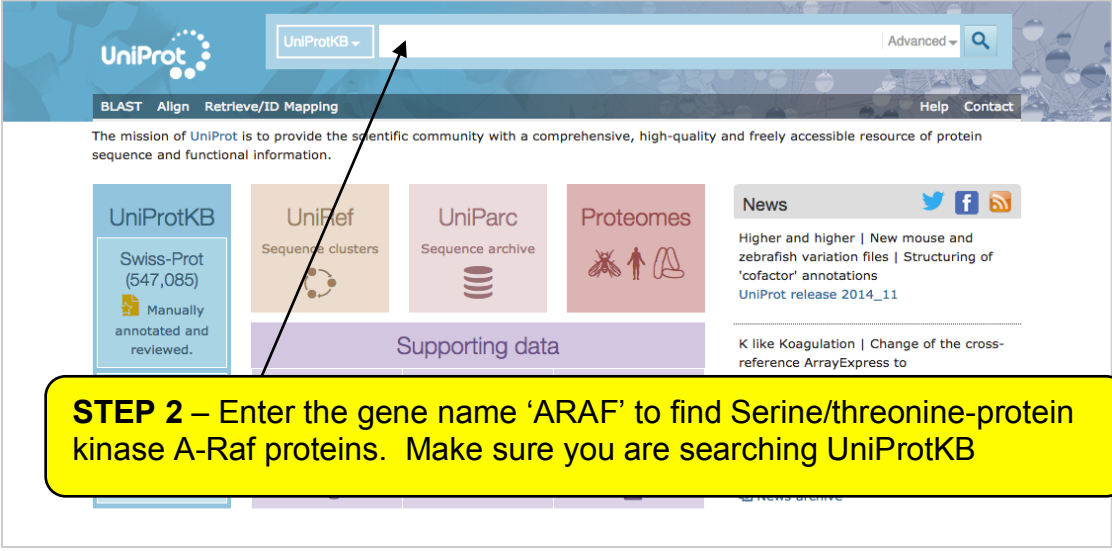
### **6.1 UniProt – protein sequence database**

In the following section UniProt (Universal Protein Resource) is the world's most comprehensive resource for protein sequence and annotation data. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR).

There are **three** parts to the UniProt databases: 1) the UniProt Knowledgebase (UniProtKB) 2) the UniProt Reference Clusters (UniRef), and 3) the UniProt Archive (UniParc). In this module we will explore UniProtKB in detail, which represents the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. The UniProtKB consists of two parts, Swiss-Prot and TrEMBL. UniProtKB/Swiss-Prot contains manually-annotated records with information extracted from literature and curator-evaluated computational analysis. The sequences in TrEMBL, which represents more than 95 % of the protein sequences in UniProtKB, are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DDBJ databases (INSDC). All these sequences, as well as the related data submitted by the authors, are automatically integrated.

## Worked Example 6.1 - using UniProt to find proteins and exploring their annotation

**STEP 1** – Open the UniProt homepage  
(<http://www.uniprot.org>)



The screenshot shows the UniProt homepage. At the top, there is a search bar with a dropdown menu set to 'UniProtKB'. Below the search bar is a navigation menu with links for 'BLAST', 'Align', 'Retrieve/ID Mapping', 'Help', and 'Contact'. The main content area features a mission statement: 'The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.' Below this are several service tiles: 'UniProtKB' (Swiss-Prot 547,085, Manually annotated and reviewed), 'UniRef' (Sequence clusters), 'UniParc' (Sequence archive), and 'Proteomes'. A 'Supporting data' section is also visible. On the right, there is a 'News' section with social media icons and a list of recent updates.

**STEP 2** – Enter the gene name 'ARAF' to find Serine/threonine-protein kinase A-Raf proteins. Make sure you are searching UniProtKB

How many sequences are from the human genome? How many sequences are really Serine/threonine kinases? (Note, UniProt updates every month, so sequence numbers are in constant flux)

**Summary of results**

**UniProt/Swiss-Prot entries**

Entry	Entry name	Protein names	Gene names	Organism	Length
P10398	ARAF_HUMAN	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf) (Proto-oncogene A-Raf-1) (Proto-oncogene Pks)	ARAF, ARAF1, PKS, PKS2	Homo sapiens (Human)	606
P04627	ARAF_MOUSE	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf)	Araf, A-raf, Araf1	Mus musculus (Mouse)	604
P14056	ARAF_RAT	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf) (Proto-oncogene A-Raf-1)	Araf, A-raf, Araf1	Rattus norvegicus (Rat)	604
O19004	ARAF_PIG	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf) (Proto-oncogene A-Raf-1)	ARAF, ARAF1	Sus scrofa (Pig)	606
Q96II5	Q96II5_HUMAN	ARAF protein (Serine/threonine-protein kinase A-Raf)	ARAF	Homo sapiens (Human)	609
Q5FBD1	Q5FBD1_DANRE	Serine/threonine protein kinase ARAF (Uncharacterized protein)	araf, ARAF	Danio rerio (Zebrafish) (Brachydanio rerio)	608
Q9SG80	ASD1_ARATH	Alpha-L-arabinofuranosidase 1 (AtASD1) (EC 3.2.1.55) (Beta-D-xylosidase) (EC 3.2.1.-)	ASD1, ARAF, ARAF1, At3g10740, T7M13.18	Arabidopsis thaliana (Mouse-ear cress)	678
Q5NSW1	Q5NSW1_TAKRU	Serine/threonine protein kinase ARAF (Uncharacterized protein)	ARAF, araf	Takifugu rubripes (Japanese)	573

Refine the search by clicking on the 'Advanced' button to the right of the search box. Repeat the search, restricting the search to the field Gene Name.

**STEP 3 – Refine the search by restricting the search to Gene Names.**

How many ARAF sequences are from Human? Do you think that the *E.coli* sequence is homologous? Why?

Now explore the entry ARAF\_HUMAN, by clicking on the sequence in the summary table:



**Results**  [About UniProtKB](#) [Basket](#)

Filter by <sup>i</sup>

- Reviewed (8) Swiss-Prot
- Unreviewed (293) TrEMBL
- Popular organisms
  - Zebrafish (12)
  - Mouse (8)
  - Human (3)
  - Rat (3)
  - A. thaliana (2)
  - Other organisms

BLAST Align Download Add to basket Columns

1 to 25 of 301 Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P04627	ARAF_MOUSE	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf)	Araf, A-raf, Araf1	Mus musculus (Mouse)	604
P10398	ARAF_HUMAN	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf) (Proto-oncogene A-Raf-1) (Proto-oncogene Pks)	ARAF, ARAF1, PKS, PKS2	Homo sapiens (Human)	606
P14056	ARAF_RAT	Serine/threonine-protein kinase A-Raf (EC 2.7.11.1) (Proto-oncogene A-Raf) (Proto-oncogene A-Raf-1)	Araf, A-raf, Araf1	Rattus norvegicus (Rat)	604
Q9SG80	ASD1_ARATH	Alpha-L-arabinofuranosidase 1	ASD1, ARAF, ARAF1, At3g10740, T7M13.18	Arabidopsis thaliana (Mouse-ear cress)	678

**STEP 4 – Click on P10398 to view the protein entry**

This represents one of the most complete entries in UniProtKB. The left

**P10398 - ARAF\_HUMAN** [Basket](#)

Protein **Serine/threonine-protein kinase A-Raf**

Gene **ARAF**

Organism *Homo sapiens (Human)*

Status Reviewed - Experimental evidence at protein level <sup>i</sup>

Display **None** BLAST Align Format Add to basket History Comment (?) Feedback Help video

**FUNCTION** **Function**<sup>i</sup>

Involves in the transduction of mitogenic signals from the cell membrane to the nucleus. May also regulate the TOR signaling cascade. [1 Publication](#)

**SUBCELL. LOCATION** Isoform 2: Serves as a positive regulator of myogenic differentiation by inducing cell cycle arrest, the expression of myogenin and other muscle-specific proteins, and myotube formation. [1 Publication](#)

**Catalytic activity**<sup>i</sup> ATP + a protein = ADP + a phosphoprotein.

**Cofactor**<sup>i</sup> Zn<sup>2+</sup> [By similarity](#)

**Note:** Binds 2 Zn(2+) ions per subunit. [By similarity](#)

**Sites**

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Metal binding <sup>i</sup>	99 – 99	1	Zinc 1 <a href="#">By similarity</a>			
Metal binding <sup>i</sup>	112 – 112	1	Zinc 2 <a href="#">By similarity</a>			
Metal binding <sup>i</sup>	115 – 115	1	Zinc 2 <a href="#">By similarity</a>			
	95 – 125	1	Zinc 1 <a href="#">By similarity</a>			
	8 – 128	1	Zinc 1 <a href="#">By similarity</a>			
	3 – 133	1	Zinc 2 <a href="#">By similarity</a>			
	6 – 136	1	Zinc 2 <a href="#">By similarity</a>			
	4 – 144	1	Zinc 1 <a href="#">By similarity</a>			
	6 – 336	1	ATP			

This column allows you to jump to different sections or toggle on/off the display

What is the function of this sequence? What metal ion does this sequence bind? How many ions are bound? Write a list of amino acids binding to each amino acid.

While the website provides an intuitive user interface, you may wish to download in different formats, either to get to the raw sequences or to parse information, or simply to provide a convenient notation for your workbook.

The screenshot shows the UniProt entry for P10398 - ARAF\_HUMAN. The entry is for the protein Serine/threonine-protein kinase A-Raf. A 'View format' dropdown menu is open, showing options: Text, FASTA (canonical), FASTA (canonical & isoform), XML, RDF/XML, and GFF. A callout box points to the 'Format' button with the text 'Click on format to reveal the list of options'.

Here is the same entry (part of) in text format.

```
ID ARAF_HUMAN Reviewed; 606 AA.
AC P10398; P07557; Q5H9B2; Q5H9B3;
DT 01-APR-1988, integrated into UniProtKB/Swiss-Prot.
DT 01-OCT-1996, sequence version 2.
DT 26-NOV-2014, entry version 177.
DE RecName: Full=Serine/threonine-protein kinase A-Raf;
DE EC=2.7.11.1;
DE AltName: Full=Proto-oncogene A-Raf;
DE AltName: Full=Proto-oncogene A-Raf-1;
DE AltName: Full=Proto-oncogene Pks;
GN Name=ARAF; Synonyms=ARAF1, PKS, PKS2;
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX PubMed=3029685; DOI=10.1093/nar/15.2.595;
RA Beck T.W., Huleihel M., Gunnell M., Bonner T.I., Rapp U.R.;
RT "The complete coding sequence of the human A-raf-1 oncogene and
RT transforming activity of a human A-raf carrying retrovirus.";
RL Nucleic Acids Res. 15:595-609(1987).
..
```

And here is the entry in FASTA format.

```
>sp|P10398|ARAF_HUMAN Serine/threonine-protein kinase A-Raf OS=Homo sapiens GN=ARAF
PE=1 SV=2 MEPPRGGPPANGAEPSPRAVGTVKVYLPNKQRTVVTVRDGMSVYDSDLKALKVRGLNQDCCV
VYRLIKGRKTVTAWDTAIAPLDGEELIVEVLEDVPLTMHNFVRKTFSSLAFCDCLKFLF
HGFRQCQTCGYKFHQCSSKVPVTVCDMSTNRQQFYHSVQDLSGGSRQHEAPSNRPLNELL
TPQGSPSPRTOHCDPEHFFPPAPANAPLQIRIRSTSTPNVHMVSTTAPMDSNLIQLTGQSFS
TDAAGSRGSDGTFRGSPSPASVSSGRKSPHSPAEQERERKSLADDDKKVKNLGYRDSG
YYWEVPPSEVQLLKRIQTGSFGTVFRGRWHGDVAVKVLKVSQPTAEQAQAFKNEMQVLRK
TRHVNILLFMGMTRPGFAIITQWCEGSSLYHHLHVADTRFDVQMLIDVARQTAQGM DYL
HAKNIIHRDLKSNINFLHEGLTVKIGDFGLATVVKTRWSGAQPLEQPSGSLVWMAAEVIRM
QDPNPYSFQSDVYAYGVVLYELMTGSLPYSHIGCRDQIIFMVGRGYSPLDSKISSNCPK
AMRRLSDCLKFORERPLFPQILATIELLQRLPKIERSASEPSLHRTQADELPACLLS
AARLVLP
```

The entry P10398 represents one of the most well experimentally characterised sequences and contains links to all of the resources covered in this module, to many covered in this course and many more. It is impossible to cover all of the resource, but please feel free to ask the tutors about the different resources.

However, not all sequences may be in UniProt (because they may be from a novel sequencing experiment) or may not have been experimentally characterised. As this is more often than not the norm, it is important to understand alternative ways of investigating protein sequences and many of the annotations present in UniProt are derived from these other databases.

## **6.2 Protein Family databases**

In the following section the protein family/domain database **Pfam** will be covered. The exemplar database has been chosen simply as Pfam is one of the most widely used databases, has high coverage of sequences, incorporation into other databases (such as InterPro and CDD) and connectivity to other major resources/tools, e.g. Ensembl, UniProt, HMMER and BLAST.

### **Pfam**

Pfam is a database of protein families and domains. This is the largest, original source of protein family data. Currently, there are over 14,000 entries in Pfam that match to nearly 80% of all sequences in UniProt. Pfam can be accessed from the following location: <http://pfam.xfam.org>.

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

**STEP 1** – Open the Pfam homepage.

wellcome trust  
**sanger**  
institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

**Pfam**  
keyword search Go

**Pfam 27.0 (March 2013, 14831 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** YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

**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. Alternatively, enter any accession or identifier into the 'Jump To' box.

**STEP 3** – Click on 'domain organisation'

Summary bar provides a quick synopsis on the entry

**Family: RBD (PF02196)**

34 architectures 710 sequences 1 interaction 96 species 14 structures

**Summary: Raf-like Ras-binding domain**

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

Wikipedia: [Raf-like Ras-binding domain](#) Pfam InterPro

This is the Wikipedia entry entitled "[Raf-like Ras-binding domain](#)". [More...](#)

**Raf-like Ras-binding domain** [Edit Wikipedia article](#)

**Raf-like Ras-binding domain** is an evolutionary conserved protein domain. This is the Ras-binding domain found in proteins related to Ras.<sup>[1]</sup>

**Examples**

Human proteins containing this domain include:

- ARAF
- BRAF
- RAF1
- RGS12, RGS14
- TIAM1

**References**

- ^ Ponting CP (October 1999). "Raf-like Ras/Rap-binding domains in RGS12- and still-life-like signalling proteins". *J. Mol. Med.* **77** (10): 695–8. doi:10.1007/s001099900054. PMID 10606204.

This protein-related article is a stub. You can help Wikipedia by expanding it.

This page is based on a [Wikipedia article](#). The text is available under the [Creative Commons Attribution](#) license.

Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk). Our cookie policy. The Wellcome Trust

Clans are collections of related Pfam entries.

The summary page contains a brief descriptions of the domain and database cross references from Wikipedia, Pfam and/or InterPro

**STEP 4 – Click on ‘Alignments’**

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

**Pfam**  
keyword search Go

**Family: RBD (PF02196)**  
34 architectures 710 sequences 1 interaction 96 species 14 structures

**Domain organisation**

Summary  
Domain organisation  
Clan  
Alignments  
HMM logo  
Trees  
Curation & model  
Species  
Interactions  
Structures

Jump to...  
enter ID/acc Go

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

**There are 231 sequences with the following architecture: RBD, C1\_1, Pkinase\_Tyr**  
Q96JIS\_HUMAN [Homo sapiens (Human)] ARAF protein (609 residues)  
RBD Pkinase\_Tyr  
[Show all sequences with this architecture.](#)

**There are 63 sequences with the following architecture: RGS, RBD x 2, GoLoco**  
Q506M0\_HUMAN [Homo sapiens (Human)] Regulator of G-protein signalling 12 (706 residues)  
RGS RBD RBD  
[Show all sequences with this architecture.](#)

**There are 48 sequences with the following architecture: PH, RBD, PDZ, RhoGEF**  
TIAM1\_HUMAN [Homo sapiens (Human)] T-lymphoma invasion and metastasis-inducing protein 1 (1591 residues)  
PH RBD PDZ RhoGEF  
[Show all sequences with this architecture.](#)

**There are 42 sequences with the following architecture: PDZ, RGS, RBD x 2, GoLoco**  
Q4SWI5\_TETNG [Tetraodon nigroviridis (Spotted green pufferfish) (Chelonodon nigroviridis)] Chromosome undetermined SCAF13617, whole genome shotgun sequence (1027 residues)  
PDZ RGS RBD RBD  
[Show all sequences with this architecture.](#)

**There are 28 sequences with the following architecture: PH, RBD, RhoGEF**  
AGAP006590-PB (2736 residues)  
PH RBD RhoGEF  
[Show all sequences with this architecture.](#)

Click to reveal all sequences with that domain organisation

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

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

**Family: RBD (PF02196)**

34 architectures 710 sequences 1 interaction 96 species 14 structures

**Alignments**

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four representative proteomes<sup>2</sup> (RP) sets, the NCBI sequence database, and our metagenomics sequence database.

**View options**

We make a range of alignments for each Pfam-A family. You can see a description of each alignment type. Please note that some types of alignment are never generated while others may be generated but are not available if the alignments are too large to handle.

	Seed (10)	Full (710)	Representative proteomes				NCBI (599)	Meta (0)
			RP15 (45)	RP35 (68)	RP55 (156)	RP75 (293)		
Jalview	✓	✓	✓	✓	✓	✓	✓	
HTML	✓	✓	✓	✓	✓	✓	✓	
PP/heatmap	X <sup>1</sup>	✓	✓	✓	✓	✓	✓	
Pfam viewer	✓	✓	X	X	X	X	X	

<sup>1</sup>Cannot generate PP/Heatmap alignments for seeds; no PP data available

Key: ✓ available, X not generated, - not available.

**Format an alignment**

Alignment:  Seed  Full

Format: Selex

Order:  Tree  Alphabetical

Sequence:  Inserts lower case  All upper case

Gaps: Gaps as "\*" or "-" (mixed)

Download/view:  Download  View

**Download options**

We make all of our alignments available in Stockholm format. You can download them here as raw, plain text files or as gzipped-compressed files.

	Seed (10)	Full (710)	Representative proteomes				NCBI (599)	Meta (0)
			RP15 (45)	RP35 (68)	RP55 (156)	RP75 (293)		
Raw Stockholm	✓	✓	✓	✓	✓	✓	-	
Gzipped	✓	✓	✓	✓	✓	✓	-	

You can also [download](#) a FASTA format file containing the full-length sequences for all sequences in the full alignment.

**External links**

MyHits provides a collection of tools to handle multiple sequence alignments. For example, one can refine a seed alignment (sequence addition or removal, re-alignment or manual edition) and then search databases for remote homologs using HMMER3.

Pfam alignments:  Seed (10)  Full (710)

Submit to MyHits

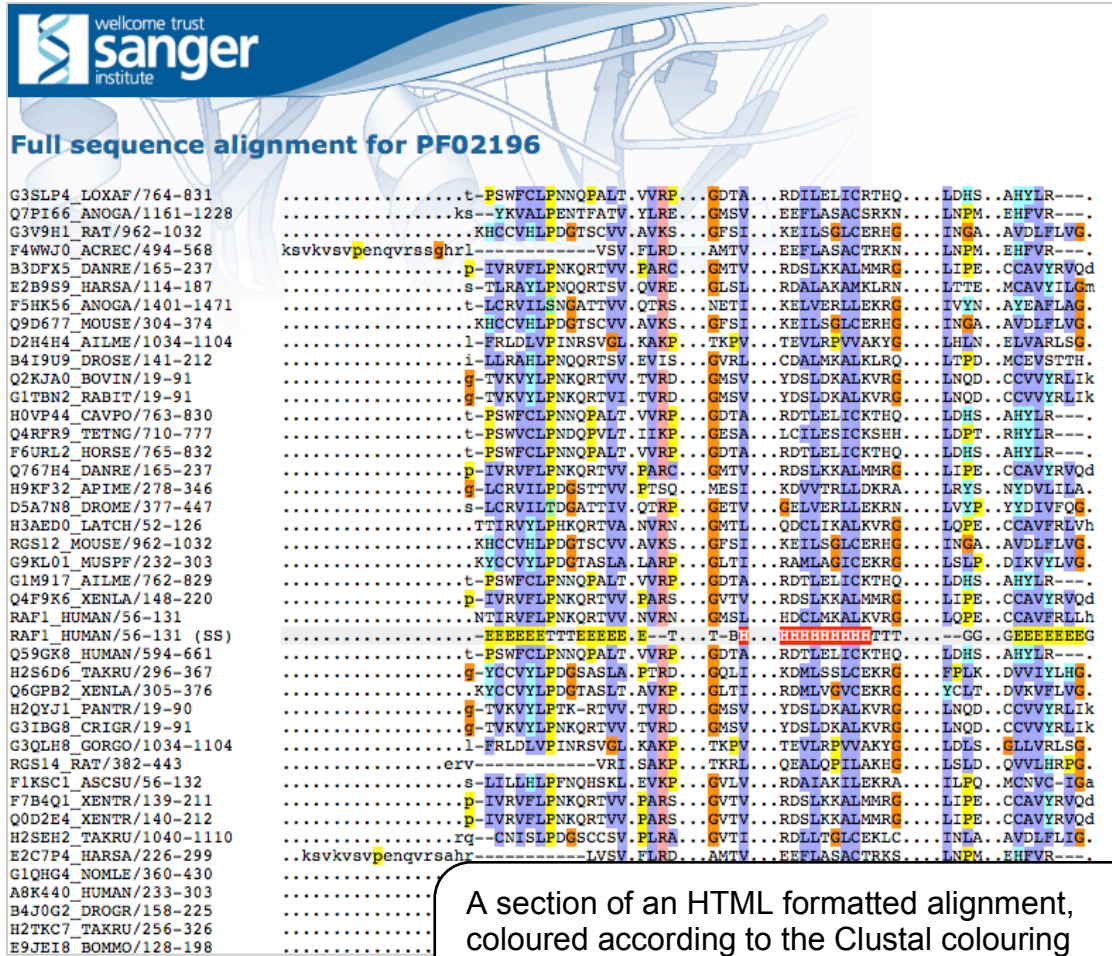
**STEP 6 – Click on 'HMM logo'**

**STEP 5 – Select 'HTML' version of the 'full' alignment**

Get the alignment in variety of formats and levels of redundancy

Each Pfam entry contains two primary alignments, termed *seed* and *full*. 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. Pfam provides additional alignments based on different sequence databases. The representative proteomes provides different levels of redundancy, from complete proteomes.





A section of an HTML formatted alignment, coloured according to the Clustal colouring scheme. The (SS) lines show secondary structure information.




## HMM logo Tab


Profile HMMs are difficult to understand if you are not used to them, converting the amino acid frequencies in the seed alignment into probabilities. To help understand them a little better, logos can be used represent the profile 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.

The screenshot shows the Pfam website interface for the RBD (PF02196) family. The top navigation bar includes links for HOME, SEARCH, BROWSE, FTP, HELP, and ABOUT. The Pfam logo is in the top right corner. The main content area displays the family name and statistics: 34 architectures, 710 sequences, 1 interaction, 96 species, and 14 structures. A sidebar on the left contains navigation links: Summary, Domain organisation, Clan, Alignments, HMM logo (highlighted), Trees, Curation & model, Species, Interactions, and Structures. A yellow callout box with an arrow pointing to the 'Species' link contains the text: **STEP 7 – Click on 'Species'**. The main content area features an HMM logo plot with the title 'HMM logo'. Below the title is a descriptive paragraph: 'HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them [here](#). [More...](#)'. The plot shows the contribution of amino acids at 26 positions. The y-axis is labeled 'Contribution' and ranges from 0 to 4. The x-axis is labeled with positions 1 to 26. The most prominent amino acid is 'D' at position 8, with a contribution of approximately 3.5. Other significant amino acids include 'V' at position 15, 'R' at position 17, 'G' at position 20, 'Y' at position 23, and 'D' at position 25. The plot also shows smaller contributions from various other amino acids at different positions.

Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk). Our [cookie policy](#).  
The Wellcome Trust



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT



**Family: RBD (PF02196)** 25 architectures 32

**Summary**

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

**Species**

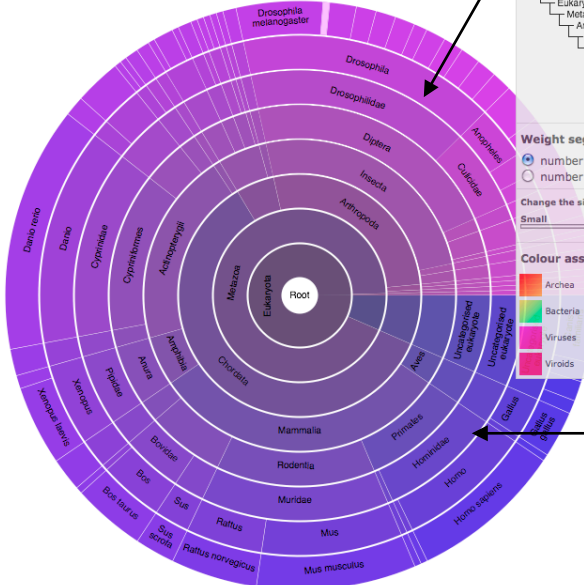
Interactions

Structures

**Species distribution**

Sunburst Tree

This visualisation provides a simple graphical representation of the distribution of across species. You can find the original interactive tree in the adjacent tab if you sub-trees and view sequence alignments. [More...](#)



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	Uncl...

**STEP 8 – Finally, click on ‘Structures’**

Select segments of the species distribution to extract matches according to taxonomy

Mouse over the segments in the spiral to reveal taxonomic data

This shows the taxonomic distribution of the hits in the family. Each segment is proportional to the family fraction of sequences matched for that taxonomic level.

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

The Wellcome Trust

**Family: RBD (PF02196)**

34 architectures 710 sequences 1 interaction 96 species 14 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 [PDB](#) 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 **RBD** domain has been found. There are 14 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
<a href="#">ARAF_HUMAN</a>	19 - 91	<a href="#">1WXM</a>	A	8 - 80	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">2L05</a>	A	155 - 227	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
<a href="#">BRAF_HUMAN</a>	155 - 227	<a href="#">3NYS</a>	A	155 - 227	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	155 - 227	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			C	155 - 227	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			D	155 - 227	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				155 - 227	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				131	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
				16	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>

**STEP 9 – Select 'Jmol' to view the structure**

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

**PDB entry 1RFA**

**STEP 10 - Right click on the structure and open up the console. Left click on the structure to reveal amino acid positions. Try to modify how the structure is represented using Jmol.**

PDB			UniProt			Pfam family	Colour
Chain	Start	End	ID	Start	End		
A	56	131	RAF1_HUMAN	56	131	RBD ( PF02196)	

[Close window](#)

**Worked Example** - Search 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.**

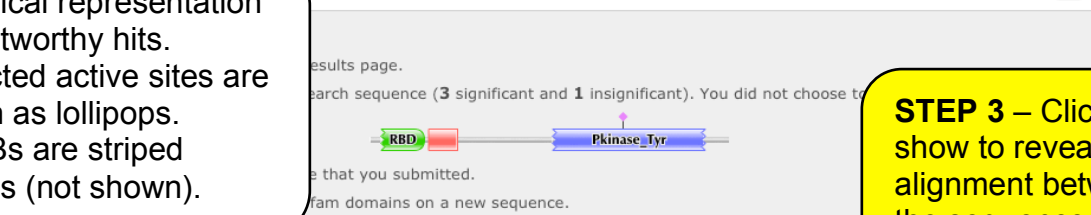
The screenshot shows the Pfam search page. At the top, there are logos for the Wellcome Trust Sanger Institute and Pfam. The navigation menu includes HOME, SEARCH, BROWSE, FTP, HELP, and ABOUT. The Pfam logo has a 'keyword search' button and a 'Go' button. Below the navigation is a 'Search Pfam' section with a 'sequence search' sub-section. A large text input field is provided for the sequence. Below the input field are 'Protein sequence options' including a 'Cut-off' section with radio buttons for 'Gathering threshold' and 'Use E-value' (which is selected), and an 'E-value' input field set to 1.0. There is also a 'Search for PfamBs' checkbox which is checked. At the bottom of the form are 'Submit', 'Reset', 'Example protein sequence', and 'Example DNA sequence' buttons. Annotations include a yellow box pointing to the 'SEARCH' menu item and another yellow box pointing to the 'Submit' button. A white box on the left explains the gathering threshold.

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 can be included.

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

Graphical representation of trustworthy hits. Predicted active sites are shown as lollipops. PfamBs are striped regions (not shown).

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



**Significant Pfam-A Matches**

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	active sites	Show/hide alignment
				Start	End	Start	End	From	To				Hide
<a href="#">RBD</a>	Raf-like Ras-binding domain	Domain	<a href="#">CL0072</a>	19	91	20	91	2	71	97.8	1.9e-28	n/a	Hide
#HMM	tirvhlPnngsrsvvevrpGmtvrDaLskalkrrgLnpsacaVrlvg...ekxpdltdisslpggeelive1												
#MATCH	t++v+LPn+qr+vv+vr+Gm+v+D+L+kalk+rgLn+++c V+++ +k+++++t+i+ L+geelive+1												
#FP	79*****999*****86												
#SEQ	LVKVIYLPKQRTVVTVRDGNISVYDSLDKALKVRGLNDDCCVYRLIKGRKTVFAMDTAIAPLDGEEELIVEV												
<a href="#">C1_1</a>	Phorbol esters/diacylglycerol binding domain (C1 domain)	Domain	<a href="#">CL0006</a>	99	147	99	145	1	51	38.8	5.1e-10	n/a	Show
<a href="#">Pkinase_Tyr</a>	Protein tyrosine kinase	Domain	<a href="#">CL0016</a>	308	565	309	563	2	257	210.2	2.4e-62	427	Show

**Insignificant Pfam-A Matches**

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				Show
<a href="#">zf-RING-like</a>	RING-like domain	Domain	<a href="#">CL0229</a>	112	142	112	138	1	29	11.8	0.17	n/a	Show

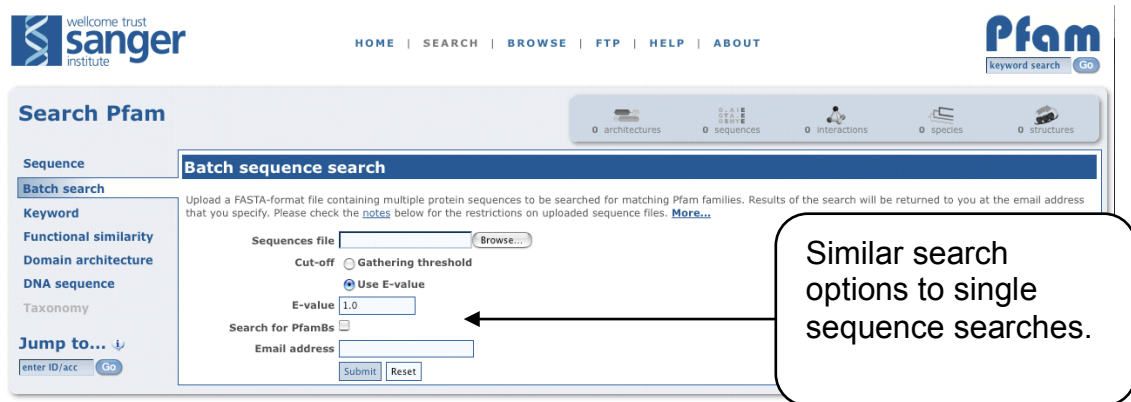
Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk). Our [cookie policy](#).  
The Wellcome Trust

What does the alignment codes mean? The top row represents the HMM and the most probably sequence to be emitted from it (you can think of it as a consensus sequence). The uppercase letters indicate the high scoring positions. The next line is the match between your query sequence and the HMM. Letters indicate an exact match, where as '+' indicate similar matches. The final line is your query sequences (or at least part of it), with the sequence region that matches this HMM aligned to it. These strings sequence can be punctuated with '-' characters 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.



The screenshot shows the Pfam search interface. At the top left is the Wellcome Trust Sanger Institute logo. The navigation bar includes links for HOME, SEARCH, BROWSE, FTP, HELP, and ABOUT. On the right is the Pfam logo with a 'keyword search' button. Below the navigation bar are icons for architectures, sequences, interactions, species, and structures. The main content area is titled 'Search Pfam' and features a sidebar with search options: Sequence, Batch search, Keyword, Functional similarity, Domain architecture, DNA sequence, and Taxonomy. The 'Batch search' section is active, showing a form to upload a FASTA file. The form includes a 'Sequences file' input with a 'Browse...' button, a 'Cut-off' section with radio buttons for 'Gathering threshold' and 'Use E-value' (selected), an 'E-value' input set to 1.0, a 'Search for PfamBs' checkbox, and an 'Email address' input. 'Submit' and 'Reset' buttons are at the bottom. A callout box on the right contains the text 'Similar search options to single sequence searches.' with an arrow pointing to the 'Use E-value' radio button.

## 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.

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 tools, but again can provide clues to key residues within 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

### **Other Domain Databases**

Two other databases that are not covered in this module, but worth mentioning are InterPro and CDD. Both of these resources take domains from other third party databases and integrate them, adding annotations and other 'value added' information. InterPro provides a hierarchical classification of the entries, so that equivalent domains from the different member databases appear as a single entry. Such integration allows the users to compare and contrast the different entries. Furthermore, InterPro added GO terms for their entries (the groupings of the database entries) that are then propagated to the sequences in UniProt. CDD provides a similar hierarchical view, but integrates fewer databases. However, the CDD curators make their own entries that complement the integrated database providing subfamily classification.

### **6.3 Protein homology searches using HMMER**

An alternative and more typical way for searching for similar protein sequences is to use homology search. For many years BLAST has been the default tool of choice. However, more recently the more sophisticated HMMER search tool has become available and is now faster than BLAST.

Worked example – find all sequences with the same domain organisation as P10398 in the UniProt reference proteomes

**STEP 1** – Go to the HMMER homepage: <http://www.ebi.ac.uk/Tools/Hmmer>, then click on the search tab



**STEP 2** – Click on ‘Accession Search’ and enter P10398 into the lookup field. Also select the reference proteome database

**STEP 3** – click submit

This will fetch the sequence and perform the search against the reference proteomes and produces the following result page that shows all of the hits.

**STEP 3** – click ‘Domain’

Row	Target	Description	Species	Known Structure	Bit Score	Hit Positions	E-value
> 1	<a href="#">ARAF_HUMAN</a>	Serine/threonine-protein kinase A-Raf	<a href="#">Homo sapiens</a>	<a href="#">RCSB</a>   <a href="#">PDBe</a>	1413.9		0.0e+00
> 2	<a href="#">K6ZK74_PANTR</a>	V-raf homolog			1413.9		0.0e+00

There are over 64,000 matches to the sequence. This is because the sequences contains a protein kinase domain. To refine the search go to the ‘Domain’ tab.

Home Search Results Software Help About Search Again

**PHMMER Results**

**STEP 4 – Scroll down until you find the matching domain architecture**

**Sequence Matches and Features**

disorder  
  coiled-coil  
  tm & signal peptide

[Show hit details](#)

[Jump to the exact match for your query architecture](#)

Domain Architectures « First « Previous Page 1 of 75 Next » Last »

26873 SEQUENCES with domain architecture: **Pkinase**, example:H7C455\_HUMAN [View Scores](#)  
[Show All](#) Sequence Features 102

4917 SEQUENCES with domain architecture: **Pkinase\_Tyr**, example:B4DV85\_HUMAN [View Scores](#)  
[Show All](#) Sequence Features 472

250 SEQUENCES with domain architecture: **cNMP\_binding, cNMP\_binding, Pkinase**, example:H3H125\_PHYRM [View Scores](#)  
[Show All](#) Sequence Features 761

**STEP 5 – Click 'View Scores'**

228 SEQUENCES **Exact match with query architecture: RBD, C1\_1, Pkinase\_Tyr**, example:ARAF\_HUMAN [View Scores](#)  
[Show All](#) Sequence Features 606

228 SEQUENCES with domain architecture: **Ephrin\_lbd, GCC2\_GCC3, fn3, fn3, EphA2\_TM, Pkinase\_Tyr, SAM\_1**, example:M3WX10\_FELCA [View Scores](#)  
[Show All](#) Sequence Features

This will show the scores for all sequences that have the same domain architecture as the query sequence.

Score Taxonomy Domain Download

**Sequence Matches and Features**

disorder  
  coiled-coil  
  tm & signal peptide

[Show hit details](#)

Your results have been filtered [Cancel](#)

All Results → [RBD C1\\_1 Pkinase\\_Tyr](#)

« First « Previous Page 1 of 3 Next » Last »

Query Matches (228) in uniprotrefprot (v.2014-10-16) [Customize](#)

Row	Target	Description	Species	Known Structure	Bit Score	Hit Positions	E-value
> 1	<a href="#">ARAF_HUMAN</a>	Serine/threonine-protein kinase A-Raf	<a href="#">Homo sapiens</a>	<a href="#">RCSB</a>   <a href="#">PDBe</a>	1413.9		0.0e+00
> 2	<a href="#">K6ZK74_PANTR</a>	V-raf murine sarcoma 3611 viral oncogene hom	<a href="#">Pan troglodytes</a>		1413.9		0.0e+00

**STEP 7 – Click 'Taxonomy'**

This will allow the investigation of the hits according to their taxonomic distribution.

Home Search Results Software Help About

Your results have been filtered [Cancel](#)

All Results → RBD C1\_1 Pkinase\_Tyr

All Hits Representative

Taxonomic distribution of all search hits

All(228) Eukaryota(228) Metazoa(228)

- Cnidaria(1) 1
- Platyhelminthes(1) 1
- Nematoda(7) 2
- Arthropoda(10) 9
- Echinodermata(1) 1
- Chordata(208) 39

**STEP 8 – Click on the arrows to navigate to the human hits**

PHMMER Results [Search Again](#)

Score Taxonomy Domain Download

Your results have been filtered [Cancel](#)

All Results → RBD C1\_1 Pkinase\_Tyr

All Hits Representative

Taxonomic distribution of all search hits

All (228) / Eukaryota (228) / Metazoa (228) / Chordata (208) / Mammalia (113) / Primates (44) / Hominidae (19) /

← back Homo(6) Homo sapiens(6)

Species Distribution			
Species	Count	View	
<a href="#">Homo sapiens</a>	6	<a href="#">Show</a>	

[Show Scores For All](#)

**STEP 10 – Click on ‘Show Scores For All’**

The resulting page shows all scores from human – how does this list compare to the list of human hits from UniProt?

**Sequence Matches and Features**

Plam: RBD, C1, Pkinase\_Tyr (606)

disorder: (606)

✓ disorder ✓ coiled-coil ✓ tm & signal peptide

Show hit details

Your results have been filtered **Cancel**

All Results → RBD\_C1\_1\_Pkinase\_Tyr → Homo sapiens

Query Matches (6) in uniprotrefprot (v.2014-10-16) **Customize**

Row	Target	Description	Species	Known Structure	Bit Score	Hit Positions	E-value
> 1	<a href="#">ARAF_HUMAN</a>	Serine/threonine-protein kinase A-Raf	<a href="#">Homo sapiens</a>	RCSB   PDBe	1413.9		0.0e+00
> 2	<a href="#">Q96IIS_HUMAN</a>	ARAF protein	<a href="#">Homo sapiens</a>		1405.3		0.0e+00
> 3	<a href="#">RAF1_HUMAN</a>	RAF proto-oncogene serine/threonine-protein kinase	<a href="#">Homo sapiens</a>	RCSB   PDBe	809.2		4.3e-241
> 4	<a href="#">RAF1_HUMAN</a>	Isoform 2 of RAF proto-oncogene serine/threonine-protein kinase	<a href="#">Homo sapiens</a>		806.4		3.0e-240
> 5	<a href="#">BRAF_HUMAN</a>	Serine/threonine-protein kinase B-raf	<a href="#">Homo sapiens</a>	RCSB   PDBe	762.9		4.5e-227
> 6	<a href="#">H7C155_HUMAN</a>	RAF proto-oncogene serine/threonine-protein kinase (Fragment)	<a href="#">Homo sapiens</a>		733.3		4.1e-218

(show all) alignments Your search took: 11.69 secs

## 6.4 Protein Structures and Complexes

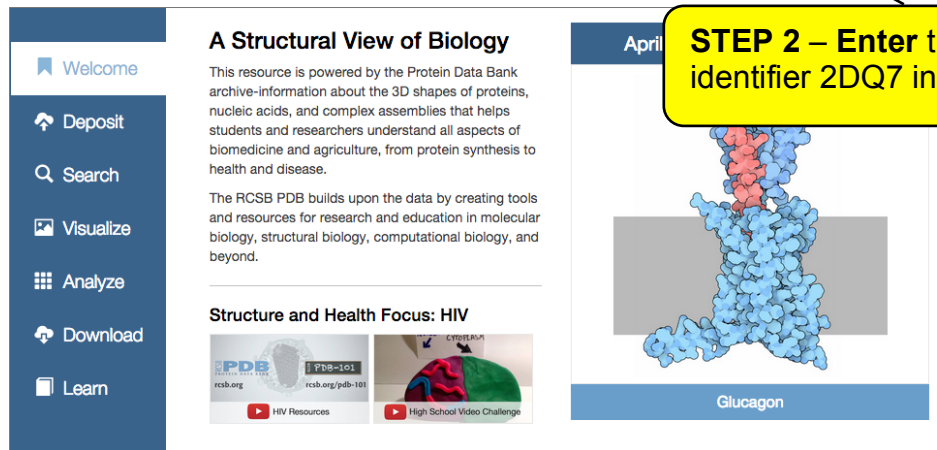
### Finding and Understanding Protein Structures (PDB, PDBsum)

In the previous section, we investigated different protein domains and features on a protein sequence, such as active sites. However, knowing the 3-dimensional structure of a protein is often vital for understanding protein function.

The Protein Data Bank (PDB) is the primary database for storing protein structure data. Here, it is possible to search for structures by their identifier or by keyword.

**Worked Example** – Find a structure with the PDB and use the site tool to investigate the bound ligand.

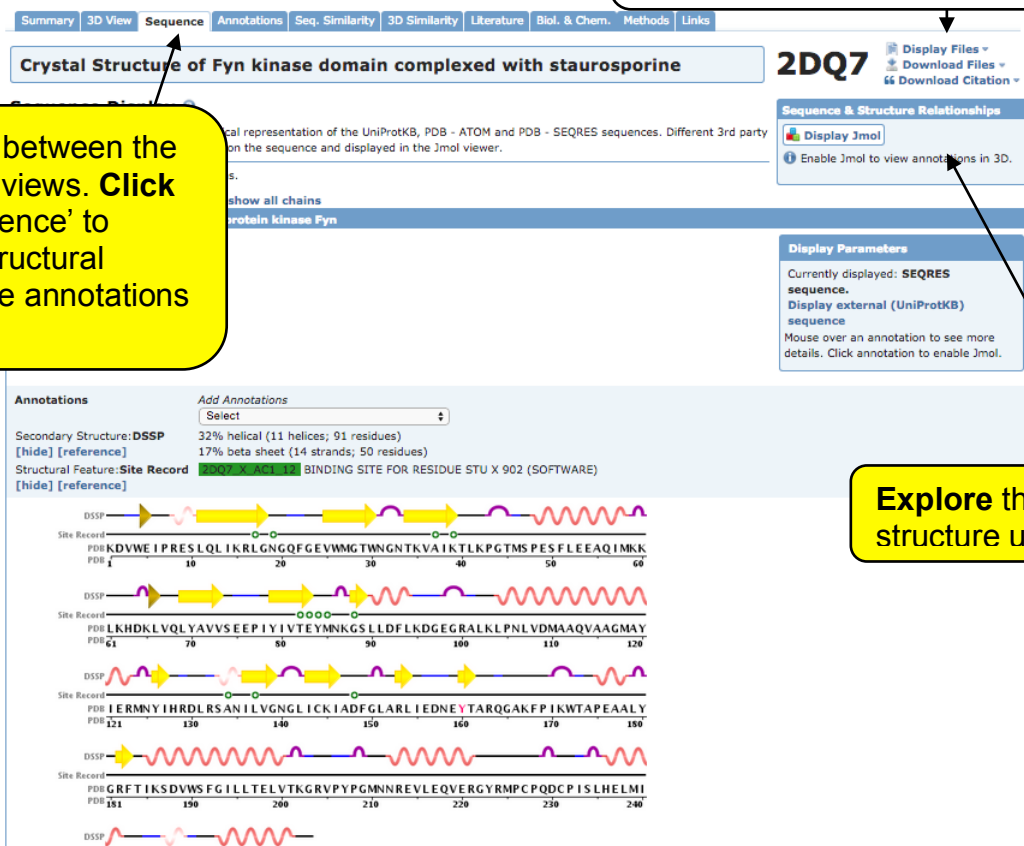
**STEP 1** – Go to the PDB homepage:  
<http://www.rcsb.org/>



**STEP 2 – Enter the PDB identifier 2DQ7 into the textfield**

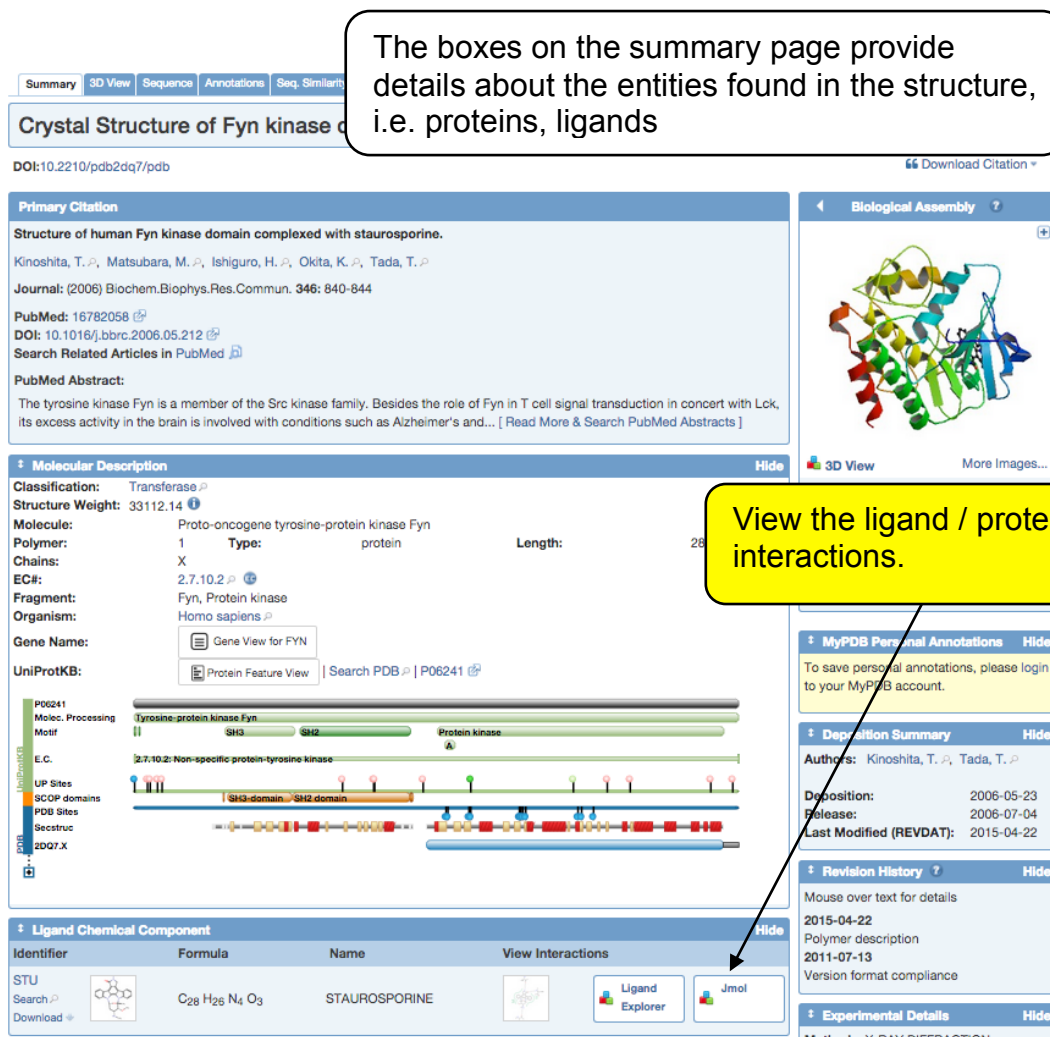
**Download entry for use in local structure viewers.**

**Change between the different views. Click on 'sequence' to obtain structural sequence annotations**



**Explore the protein structure using Jmol**

The boxes on the summary page provide details about the entities found in the structure, i.e. proteins, ligands



**Primary Citation**  
Structure of human Fyn kinase domain complexed with staurosporine.  
Kinoshita, T., Matsubara, M., Ishiguro, H., Okita, K., Tada, T.  
Journal: (2006) Biochem.Biophys.Res.Commun. 346: 840-844  
PubMed: 16782058  
DOI: 10.1016/j.bbrc.2006.05.212  
Search Related Articles in PubMed

**Molecular Description**  
Classification: Transferase  
Structure Weight: 33112.14  
Molecule: Proto-oncogene tyrosine-protein kinase Fyn  
Polymer: 1 Type: protein Length: 268  
Chains: X  
EC#: 2.7.10.2  
Fragment: Fyn, Protein kinase  
Organism: Homo sapiens  
Gene Name: FYN  
UniProtKB: P06241

**Ligand Chemical Component**

Identifier	Formula	Name	View Interactions
STU Search Download	C <sub>28</sub> H <sub>26</sub> N <sub>4</sub> O <sub>3</sub>	STAUROSPORINE	Ligand Explorer Jmol

**View the ligand / protein interactions.**

The PDB website provides a number of additional pages and tools that allow interrogation of protein structures. These can be very powerful tools if you have a protein structure, but are beyond the scope of this tutorial. However, some of the most useful are outlined in the appendix.

**PDBSum** – As 3D protein structures can be very difficult to interpret, the PDBSum provide a series of display that provide users with detailed information about the structure via a user-friendly interface.

**Worked Example** – Exploring a structure in PDBsum

**STEP 1** – Go to the PDBsum home page  
<http://www.ebi.ac.uk/pdbsum>

The screenshot shows the PDBsum website interface. At the top, the title "PDBsum Pictorial database of 3D structures in the Protein Data Bank" is displayed. The left sidebar contains a navigation menu with categories like "Highlights", "List of PDB codes", "Het Groups", "Ligands", "Drugs", "Enzymes", "UniProt", "Pfam", "ProSite", "Species", "Generate", "Gallery", "Figure stats", "Documentation", "Downloads", and "Contact us". The main content area includes a breadcrumb trail "Databases > Structure Databases > PDBsum", a description of PDBsum, and three search sections: "PDB code (4 chars)", "Text search", and "Search by sequence". A yellow callout box with a black border points to the "Text search" field and contains the text: "STEP 2 – Enter 2dq7 into textfield and click find. You can also search by keyword and sequence". The right sidebar features a "Contents" section with statistics and an "In-house version" section with a "DrugPort" section at the bottom.

**PDBsum** Pictorial database of 3D structures in the Protein Data Bank

Databases > Structure Databases > PDBsum

PDBsum is a pictorial database that provides an at-a-glance overview of the contents of each 3D structure deposited in the Protein Data Bank (PDB). It shows the molecule(s) that make up the structure (ie protein chains, DNA, ligands and metal ions) and schematic diagrams of the interactions between them. [Read more ...](#)

**PDB code (4 chars)**  **Find** Example: "1ktv"

**Text search**  **Search**  
Scans all TITLE, HEADER, COMPND, SOURCE and AUTHOR records in the PDB (eg to find a given protein by name)

**Search by sequence**  
  
**Search**  
Perform FASTA search vs all sequences in the PDB to get a list of the closest matches.

**Notes**  
**NEW** The PDBsum **Highlights** and **Species** analyses have now been retired and replaced by links to the more sophisticated

**Contents**  
PDBsum contains 96,852 entries, including 1,950 superseded  
Last update: 7 September, 2013

**In-house version**

**DrugPort**  
Structures of drugs and their target proteins in the PDB.

**STEP 2 – Enter 2dq7 into textfield and click find. You can also search by keyword and sequence**

EMBL-EBI   [Terms of Use](#) [Privacy](#) [Cookies](#)

Databases

**STEP 3 – Click on ligands to get summary of interaction**

Go to PDB code:

**PDB id: 2dq7** [Links](#)

**Name: Transferase**

**Title:** Crystal structure of fyn kinase domain complexed with staurosporine

**Structure:** Proto-oncogene tyrosine-protein kinase fyn. Chain: x. Fragment: fyn, protein kinase. Synonym: p59-fyn, protooncogene syn, slk. Engineered: yes

**Source:** Homo sapiens. Human. Organism\_taxid: 9606. Expressed in: spodoptera frugiperda. Expression\_system\_taxid: 7108. Expression\_system\_cell\_line: sf21.

**Resolution:** 2.80Å **R-factor:** 0.255 **R-free:** 0.281

**Authors:** T.Kinoshita,T.Tada

**Key ref:** T.Kinoshita et al. (2006). Structure of human Fyn kinase domain complexed with staurosporine. *Biochem Biophys Res Commun*, 346, 840-844. PubMed id: [16782058](#) DOI: [10.1016/j.bbrc.2006.05.212](#)

**Date:** 23-May-06 **Release date:** 04-Jul-06

**PROCHECK**

**Protein chain**

**P06241 (FYN\_HUMAN) - Tyrosine-protein kinase Fyn**

Seq: SH2 SH3

Struc:

Seq: 537 a.a.

Struc: 263 a.a.

**Enzyme reactions**

**Enzyme class:** [E.C.2.7.10.2 - Non-specific protein-tyrosine kinase.](#)

**Reaction:**  $ATP + a \text{ [protein]-L-tyrosine} = ADP + a \text{ [protein]-L-tyrosine phosphate}$

**Gene Ontology (GO) functional annotation**

	<b>Biological process</b>	protein amino acid phosphorylation	2 terms
	<b>Biochemical function</b>	protein binding	5 terms

**Click for orthogonal views of the structure**

**Summary of information contained in PDB/MSD**

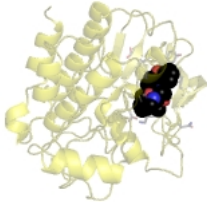


**PDBsum** Go to PDB code:

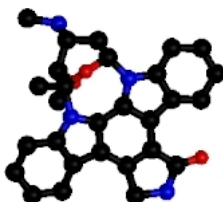
Top page  
  Protein  
  **Ligands**  
  Clefts  
  Links


Ligand/metal interactions PDB id **2dq7**


**Ligand STU - Staurosporine**  
Formula:  $C_{28}H_{26}N_4O_3$

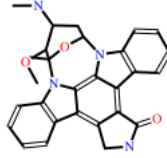


Ligand highlighted  
**STU**





Postscript version  


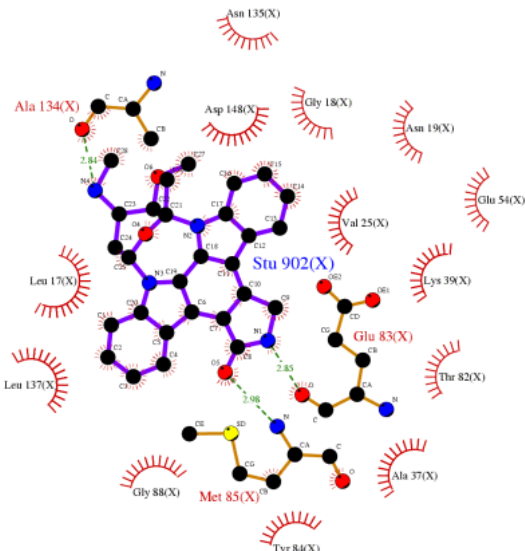



**Ligands**


STU  
STU 902(X)

**LIGPLOT of interactions involving ligand**

Simple 2D representation of ligand interaction





Postscript version  


Now explore a different structure, 1gua that we looked at on the Pfam site.

**STEP 4** – Enter **1gua** in the 'Go to pdb code' box in the top right of the page and click on go.

**PDBsum** Go to PDB code:

**Top page** | Protein | Ligands | Metals | **Prot-prot** | Clefts | Links

**Complex (gtp-binding/atp-binding)** PDB id: **1gua**

**STEP 5 – Click on the Prot-prot tab**

**PDB id: 1gua** [Links](#)

**Name:** Complex (gtp-binding/atp-binding)  
**Title:** Human rap1a, residues 1-167, double mutant (e30d,k31e) complexed with the ras-binding-domain of human c-raf1, residues 51-131  
**Structure:** Rap1a. Chain: a. Fragment: residues 1-167. Engineered: yes. Mutated: no. Other\_details: complexed to 5'-guanosyl-imido-triphosphate. C-terminus: residues 51-131.  
**Source:** Homo sapiens. Human. Organism\_taxid: 9606. Gene: human c-raf1 gene residues 51. Expressed in: escherichia coli. Expression\_system\_taxid: 562. Expressed in: escherichia coli bl21(de3). Expression\_system\_taxid: 469008. Other\_details: purified as a gst-fusion protein with  
**Biol. unit:** Dimer (from PQS)  
**Resolution:** 2.00Å **R-factor:** 0.220  
**Authors:** N.Nassar,A.Wittinghofer  
**Key ref:** N.Nassar et al. (1996). Ras/Rap effector specificity determined by charge reversal. *Nat Struct Biol*, 3, 723-729. PubMed id: 8756332 DOI: 10.1038/nsb0896-723  
**Date:** 18-Jun-96 **Release date:** 11-Jan-97

**Contents**

- Protein chains
  - A 167 a.a.\*
  - B 76 a.a.\*
- Ligands
  - GNP
- Metals
  - CA
  - MG
- Waters x89

\* Residue conservation analysis

**Protein chain A** (P62834) (RAP1A\_HUMAN) - Ras-related protein Rap-1A

Seq: 184 a.a.  
 Struc: 167 a.a.\*

**Protein chain B** (P04049) (RAF1\_HUMAN) - RAF proto-oncogene serine/threonine-protein kinase

Seq: 646 a.a.  
 Struc: 76 a.a.

Key: ■ Family PfamA domain ▬ Secondary structure ↔ CATH domain  
 \* PDB and UniProt seqs differ at 2 residue positions (black crosses)

**Enzyme reactions**

**Enzyme class:** Chain B: [E.C.2.7.11.1](#) - Non-specific serine/threonine protein kinase. [\[IntEnz\]](#) [\[ExPASy\]](#) [\[KEGG\]](#) [\[BRENDA\]](#)

**Reaction:** ATP + a protein = ADP + a phosphoprotein

This produces a similar page to the 'top page' for 2dq7. However, you may have notice that there are two proteins in this structure. Also, there is an additional tab 'Prot-prot'.

The Prot-prot tab is divided in to two parts. The first is a gross summary of the protein interactions. Below this is a detailed schematic of the amino-acid contacts and the bonds formed between them.



Go to PDB code:

[Top page](#)
[Protein](#)
[Ligands](#)
[Metals](#)
[Prot-prot](#)
[Clefts](#)
[Links](#)

Protein-Protein interface: A|B

PDB id **1gua**

**Protein-protein interface: A|B**



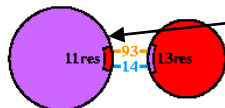
Chains A and B highlighted (click to view)



A|B (11:13 res)

\* Coloured by residue conservation

**Chain A Chain B**



Schematic of the protein-protein interactions

Key: — Salt bridges — Disulphide bonds — Hydrogen bonds — Non-bonded contacts

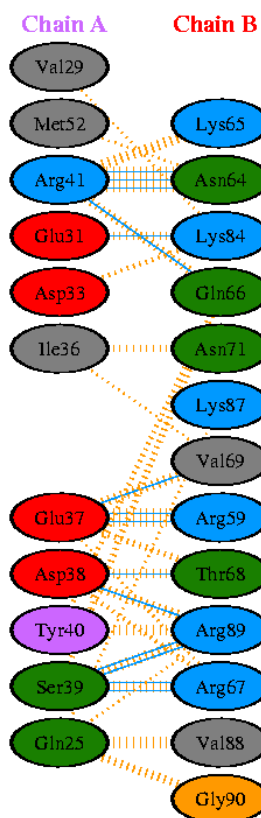
**Schematic diagram of interactions between protein chains.** Interacting chains are joined by coloured lines, each representing a different type of interaction, as per the key above. The area of each circle is proportional to the surface area of the corresponding protein chain. The extent of the interface region on each chain is represented by the black wedge whose size signifies the interface surface area. Statistics for this interface are given below.

**Interface statistics**

Chain	No. of interface residues	Interface area (Å <sup>2</sup> )	No. of salt bridges	No. of disulphide bonds	No. of hydrogen bonds	No. of non-bonded contacts
A	11	669	-	-	14	93
B	13	622	-	-	-	-

**Residue interactions across interface**

Coloured by residue type



Detailed interaction data is listed at the bottom of the page

- List of interactions
- Diagram in PDF format
- Diagram in PostScript format

Key: — Salt bridges — Disulphide bonds — Hydrogen bonds — Non-bonded contacts

The number of H-bond lines between any two residues indicates the number of potential hydrogen bonds between them. For non-bonded contacts, which can be plentiful, the width of the striped line is proportional to the number of atomic contacts.

Residue colours: Positive (H,K,R); negative (D,E); S,T,N,Q = neutral; A,V,L,I,M = aliphatic; F,Y,W = aromatic; P,G = Pro&Gly; C = cysteine.

## SWISS-MODEL

Although the number of known 3D structures is now over 100,000, the number of sequences in UniProt is over 100 times greater. To bridge the gulf in numbers, it is possible to use homology modelling to estimate the structural arrangement of a protein with an undetermined 3D structure. In the following example, the SWISS-MODEL server will be used to perform homology modelling. More details about SWISS-MODEL can be found at <http://swissmodel.expasy.org>.

**Worked Example** – Use the SWISS-MODEL server to identify structural homologues of a sequence and construct a homology model.

**STEP 1** – Go to the SWISS-MODEL server: <http://swissmodel.expasy.org/> and click on ‘

The screenshot shows the SWISS-MODEL web interface. At the top, there is a navigation bar with 'Modelling', 'Tools', 'Repository', 'Documentation', 'Log in', and 'Create Account'. A yellow banner below the navigation bar contains a notice about a service disruption from Thursday 30th April to Sunday 3rd May. The main content area is titled 'Start a New Modelling Project'. It features a 'Target Sequence' input field with a callout box pointing to it that says 'Full search (slow) modelling'. Below the input field is an 'Upload Target Sequence File...' button. There are also fields for 'Project Title' (Untitled Project) and 'Email' (Optional). At the bottom of the form are 'Search For Templates' and 'Build Model' buttons. To the right of the form is a 'Supported Inputs' section with a list of input types: Sequence, Uniprot AC, Target-Template Alignment, Upload Template, and Deepview Project. At the bottom of the page, there is a disclaimer: 'By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles. I have read the terms of use, and hereby state that I am an academic non-commercial user (Please select)'.

**STEP 2 – Click Repository and Paste your sequence accession P14056 into the textfield and click submit to find structural homologs, or look to see if it is in the repository.**

If the sequence has a structural template then a page like the following will be displayed.

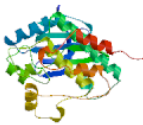
**STEP 3 - Click on a structural model to switch between results**


Model information:		Quaternary structure information:
Modelled residue range:	301 to 574	NA
Based on template:	[ 1uwj ]	
Sequence Identity [%]:	74%	
Model date:	2008-08-29	
Revision date:	2008-08-29	
Ligand information:		
		NA

**Model 3D Structure [+/-]**

**Based on template: 1uwj** [SMTL] [RCSB] [PDBe] [SCOP] [CATH]

Sequence identity: 74%  
 Residue range: 301 to 574  
 Model date: 2008-08-29  
 Revision date: 2008-08-29  
[\[ display \]](#) [\[ download \]](#) [\[ download project \]](#)




 This model has not been updated since 2008-08-29. In the meantime, new template structures have become available which would allow building a more reliable model. Would you like to target protein to SWISS-MODEL Workspace and build a new model now? [\[ Submit \]](#)

**Alignment [+/-]** [\[ start new quality assessment in Workspace \]](#)

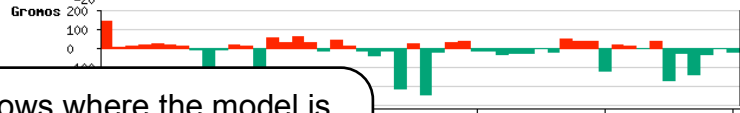
**Quality [+/-]**

**Model Quality Assessment**

**Anolea**



**Gromos**



FRGRWHGDVAVKVLKVAQP331TAEQAQAFK341

**STEP 4 – Display the template and model in DeepView**

Alignment of sequence to the template used, in this case 1uwj, chain A.

This graph shows where the model is expected to be bad (red) and good (green). As we have found an exact match, the model is very good.

**STEP 5 -** Look at both models. Notice how the quality differs, as the sequence identity of the alignment changes.

Target is shown in yellow and template as a ribbon representation of the structure. This view takes a bit of work!

Alignment in DeepView

group	show	side	labl	ribs	col	B/S
X	TYR207					
X	PRO208					
X	GLY209					
X	MET210					
X	ASN211					
h	ASN212					
h	ARG213					
h	GLU214					
h	VAL215					
h	LEU216					
h	GLU217					
h	GLN218					
h	VAL219					
h	GLU220					
h	ARG221					
X	GLY222					
X	TYR223					
X	ARG224					
X	MET225					
X	PRO226					
X	CYS227					
X	PRO228					
X	GLN229					
X	ASP230					
X	CYS231					
X	PRO232					
X	ILE233					
X	SER234					
X	LEU235					
X	HIS236					
h	GLU237					
h	LEU238					
h	MET239					
h	ILE240					
X	HIS241					
h	CYS242					
X	TRP243					
X	LYS244					
X	LYS245					
X	ASP246					
X	PRO247					
X	GLU248					
X	GLU249					
X	ARG250					
X	PRO251					
X	THR252					
h	PHE253					
h	GLU254					
h	TYR255					
h	LEU256					
h	GLN257					
h	SER258					
h	PHE259					
h	LEU260					
X	GLU261					

```

> TARGET      36.1 0.0HG D - - YAVKVLKVAOPTAEQAQAFKNEMOVLRKTRHVNILLFMGF MTR
> refsim     100.000NGNTKVAIKTLK - - - PGTMSPE SFLEEAOIMKKLKHKDKLVOLYAVVSE
> 2dq7X

```

Unfortunately, it is beyond the scope of the module to go in to any great detail as to how to use DeepView/Swiss-PDB viewer. However, the tutors may be able to give you a quick introduction.

Currently only a small fraction of proteins have been determined structurally. Thus, for the majority of proteins there are no structural homologs so homology modelling is impossible. If you are really interested in structure, the only remaining form of analysis is secondary structure prediction. As this is only of limited use, the uses of a secondary structure prediction tool as been consigned to the appendix.

## 6.5 Protein Interactions

The IntAct database contains protein interactions that are curated from the literature. IntAct is part of a wider consortium, which regularly exchanges curated interaction datasets. As such, IntAct contains one of the larger collections of protein interaction data. The following **worked example** illustrates how to access the data contained with this database.

**STEP 1 – Go to the IntAct homepage:**  
<http://www.ebi.ac.uk/intact>

EMBL-EBI Services Research Training About us

# IntAct

Home Search Browse Data Submission Downloads Datasets Statistics FAQ Developer Resources Contact Us About IntAct Feedback

## IntAct Molecular Interaction Data

IntAct provides a freely available, open source database system and analysis tool. Interactions are derived from literature curation or direct user submissions and are available in the IntAct database use the search box above.

Search in IntAct  Search Clear [Show Advanced Fields >](#)

[MIQL syntax reference](#)

### Search Tips

- Free text search will look by default for interactor identifier, (e.g. gene name *BRCA2*, UniProtKB Ac *Q06609* or UniProtKB Id *dmc1*), species, interaction id, detection method, interaction type, publication identifier or author (e.g. Pubmed Id *10831611*), interactor xrefs, interaction xrefs.
- For a more specific search, use MIQL syntax or advanced search
- Search based on exact word matches e.g. *BRCA2* will not match *BRCA2B*
- Search for isoforms of 'P12345' by using 'P12345\*'

### Contributors

Manually curated content is added to IntAct by curators at the EMBL-EBI and the following organisations:

- MINT
- UniProt
- SIB



**STEP 3 – Click on interaction details.**

145 binary interactions found for search term *P10398*

Interactions (145) | Browse | Lists | **Interaction Details** | Molecule View | Graph

Filter out the spoke expanded co-complexes (70) | Your query also matches 8,664 interaction evidences from 8 other databases. | What is this view?

Your query also matches 1 interaction evidences from 1 other IMEX databases.

Dts	Molecule 'A'	Links 'A'	Molecule 'B'	Links 'B'	Interaction Detection Method	Interaction AC	Source Database
1	ARAF	P10398 EBI-365961	YWHAZ	P63104 EBI-347088	peptide array	EBI-7616897 MINT-8009422	MINT
2					coimmunoprecipitation	EBI-7702412 MINT-8009569	MINT
3					two hybrid pooling approach	EBI-3438014 imex : IM-17049-85	IntAct
4					anti tag coimmunoprecipitation	EBI-10101513 imex : IM-23674-5	IntAct
5					anti tag coimmunoprecipitation	EBI-10101587 imex : IM-23674-6	IntAct
6	ARAF	P10398 EBI-365961	MAP2K2	P36507 EBI-1056930	two hybrid	EBI-1164984 imex : IM-19677-1	IntAct
7					two hybrid	EBI-1165021 imex : IM-19677-5	IntAct
8					pull down	EBI-1165031	IntAct

interaction\_id:EBI-7616097

Examples: BRCA2, Q06609, dmc1, 10831611

Home | Search | Browse | Data Submission | Downloads | Datasets | Statistics | FAQ | Developer Resources | Contact Us | About IntAct

IntAct > Interaction Details

Interactions (1) | Browse | Lists | **Interaction Details** | Molecule View | Graph

### Publication

PubMed Id: 19933840 | Title: Regulation of IRSp53-dependent filopodial dynamics by antagonism between 14-3-3 binding and SH3-mediated localization.

Journal: Mol. Cell. Biol. (0270-7306) | Author List: Robens JM., Yeow-Fong L., Ng E., Hall C., Manser E.

Year of Publication: 2010

Cross References:

Database	Identifier	Secondary identifier	Qualifier
pubmed	19933840	-	primary-reference
mint	MINT-8009399	-	primary-reference
mint	MINT-8009405	-	identity
doi	10.1128/MCB.01574-08	-	primary-reference

### Experiment (3 interactions)

Accession: EBI-7616023 | Host organism: Unknown

Name: robens-2010-1 | Interaction Detection Method: peptide array

Participant Identification Method: predetermined

Cross References:

Database	Identifier	Secondary identifier	Qualifier
mint	MINT-8009405	-	identity
doi	10.1128/MCB.01574-08	-	primary-reference
pubmed	19933840	-	primary-reference

Annotations:

Topic	Text
partial coverage	partial coverage
rapid curation	rapid curation

Details of the protein-protein interactions and how they were determined and by whom

### Interaction

Accession: EBI-7616097      Description: -

Name: araf-ywhaz-1      Type: [association](#)

[Links](#)  
[Find similar interactions](#)

Cross References:

Database	Identifier	Secondary Identifier	Qualifier
mint	MINT-8009422	-	identity

Annotations:

Topic	Text
comment	homomint
comment	domino
comment	mint

### Participants (2)

Legend: ■ Annotation and Cross Reference   ■ Experimental Parameter   ■ Stoichiometry   ■ Experimental Feature   ■ Participant Confidence

#	Name	Links	Primary Identifier	Aliases	Description	Species	Expression system	Experimental role	Biological role	Interactor type	More...
1	EBI-347088	<a href="#">UniProt</a> <a href="#">BioGRID</a> <a href="#">Mint</a>	P63104	YWHAZ Protein kinase C inhibitor protein 1	14-3-3 protein zeta/delta	Homo_sapiens	-	unspecified role	unspecified role	protein	
2	EBI-365961	<a href="#">UniProt</a> <a href="#">BioGRID</a> <a href="#">Mint</a>	P10398	ARAF PKS PKS2	Serine/threonine-protein kinase A-Raf	Homo_sapiens	-	unspecified role	unspecified role	protein	

**STEP 4 – Return** to the original search and **Click on Graph**.

### Network visualisation

[Open in Cytoscape](#)

Click [here](#) or on the icon above to start Cytoscape.

This is going to open the current search in a WebStart [version](#) of Cytoscape 2.6.3

**CytoscapeWeb Controls**

Layouts: **force directed** | radial  
| circle

Merge edges: **on** | off

People often ask about domain-domain interaction. 3DID – is a database of high quality domain interactions. (<http://3did.irbbarcelona.org/>)

## 6.6 Pathway databases

At the cellular level, life is a network of molecular reactions that can be organized into higher order interconnected pathways. Molecules are synthesized, degraded, transported from one location to another and assembled into complexes and higher order structures with other molecules. This module will cover two pathways databases Reactome and MetaCyc. Reactome has been chosen as it is a curated database, primarily aimed at human pathways. The second database, MetaCyc, is a broad pathway database.

**Reactome** – Investigate the signal transduction pathways for the gene RAF.

**STEP 1 – Go to the Reactome homepage at:**  
<http://www.reactome.org>

The image shows the Reactome homepage with several annotations. A yellow box at the top contains the text 'STEP 1 – Go to the Reactome homepage at: http://www.reactome.org'. A white box on the right side contains the text 'Three main entry points to the database' and points to three icons: 'Browse Pathways', 'Analyze Data', and 'Reactome FI Network'. A yellow box at the bottom contains the text 'STEP 2 – Click on “browse pathways”' and points to the 'Browse Pathways' icon. The homepage itself features a navigation bar with links for 'About', 'Content', 'Documentation', 'Tools', 'Community', 'Download', and 'Contact'. Below the navigation bar are six icons for 'Browse Pathways', 'Analyze Data', 'Reactome FI Network', 'User Guide', 'Data Download', and 'Contact Us'. A 'Tweets' section is visible on the right, showing two tweets from @reactome. At the bottom, there are logos for OICR, NYU Langone, CSH Cold Spring Harbor, and EMBL-EBI.

**STEP 2 – Click on “browse pathways”**

The resulting page contains a list of all pathways found in Reactome. The drop down list of species contains a list of all eukaryotic species contained within Reactome.

The left panel contains a list of all pathway categories, while the image represents the same thing graphically

**STEP 2 – Click on “signal transduction” in the left panel and then the “RAF/MAP cascade”**

**STEP 3 – Click on “RAF activation” in the left panel and then “p-RAF binds 14-3-3”**

Focus the window on the highlighted part of the pathway network (bottom left)

Overview Molecules Structures Expression Analysis Processes Downloads

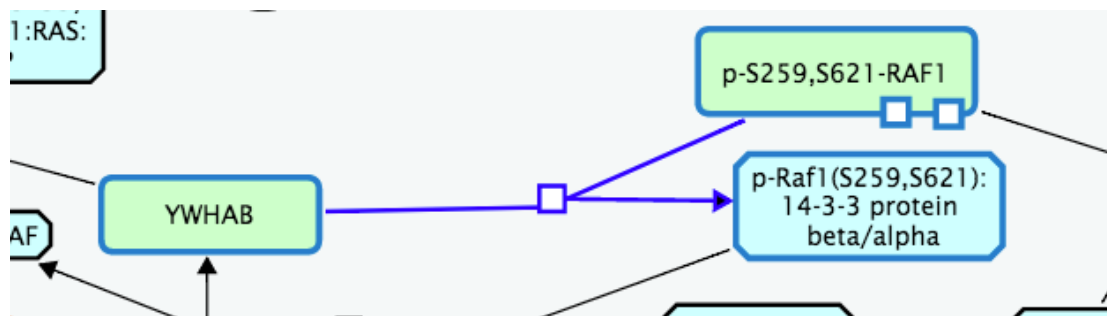
Species: Homo sapiens

Stable Identifier  
REACT\_2158.2

Summation

Inactive Raf-1 is associated in the cytoplasm with 14-3-3. 14-3-3 binds to Raf-1 via the Ser259 phosphorylation site (S1). This interaction stabilises the inactive conformation of Raf-1 in which the Ras-binding Cysteine-rich domain (CRD) is obscured. The Raf-1 molecule contains an additional p21ras-binding domain (RBD), a second serine phosphorylation site at S621 (S2) and two tyrosine phosphorylation sites (at 340, Y1 and 341, Y2).

This highlights a small sub-section of the entire signal transduction cascade.



This depicts the two protein components (YWHAB and pRAF1) coming together to form a complex. The overview describes the events that occur as the complex is formed.

Overview Molecules (2/21) Structures (0) Expression Analysis Processes Downloads

↳ p-RAF binds 14-3-3 beta/alpha Species: Homo sapiens

Stable Identifier  
[REACT\\_2158.2](#)

Summation

Inactive Raf-1 is associated in the cytoplasm with 14-3-3. 14-3-3 binds to Raf-1 via the Ser259 phosphorylation site (S1). This interaction stabilises the inactive conformation of Raf-1 in which the Ras-binding Cysteine-rich domain (CRD) is obscured. The Raf-1 molecule contains an additional p21ras-binding domain (RBD), a second serine phosphorylation site at S621 (S2) and two tyrosine phosphorylation sites (at 340, Y1 and 341, Y2).

You can then focus the annotation on the complex. This indicates the two proteins involved and shows that are know structures of this complex. In the network diagram is shows what the complex interacts with. What happens?

**STEP 4 – Click on the “Raf1/14-3-3” complex**

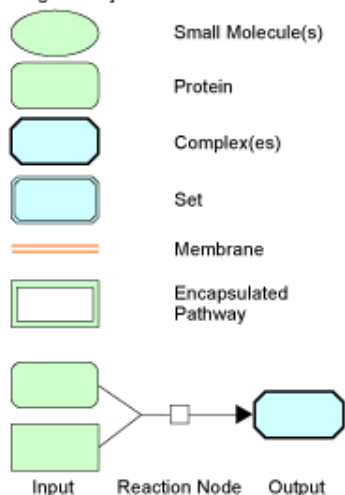
Showing 2 of 2 genes found:

Gene	adipose tissue	adrenal gland	animal ovary	appendix	bladder	bone marrow	cerebral cortex	colon	duodenum	endometrium	esophagus	gall bladder	heart	kidney	liver
YWHAB	Low	Low	Low	Low	Low	Low	High	Low	Low	Low	Low	Low	Low	Low	Low
RAF1	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low

**STEP 4 – Click on the Expression tab to reveal where these two proteins are expressed**

### Diagram Key

#### Diagram Objects

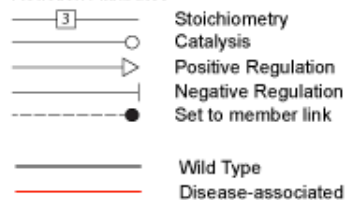


Key to the Reactome Pathway diagram

#### Reaction Types



#### Reaction Attributes



[Click here for more detailed diagram key](#)

Now return to the Reactome homepage. Reactome uses manually-curated human pathways to electronically 'infer' their equivalents in 19 other species. To compare annotations, say between human and mouse, we can use the Reactome comparison tool.

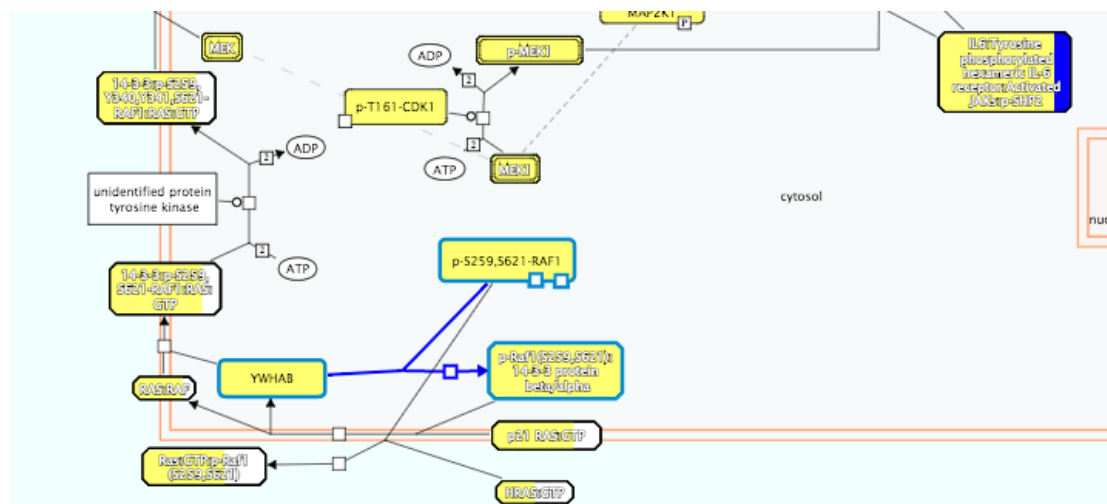
**STEP 5 – Click on ‘Analyze Data’**

**STEP 6 – Select ‘Mus musculus’ and click on compare**

Yellow indicates that the protein has an inferred equivalent in the comparison species. Blue indicates that no equivalent was identified. This protein may not exist in the comparison species.

**STEP 7 – Descend down to the same pathway as before**

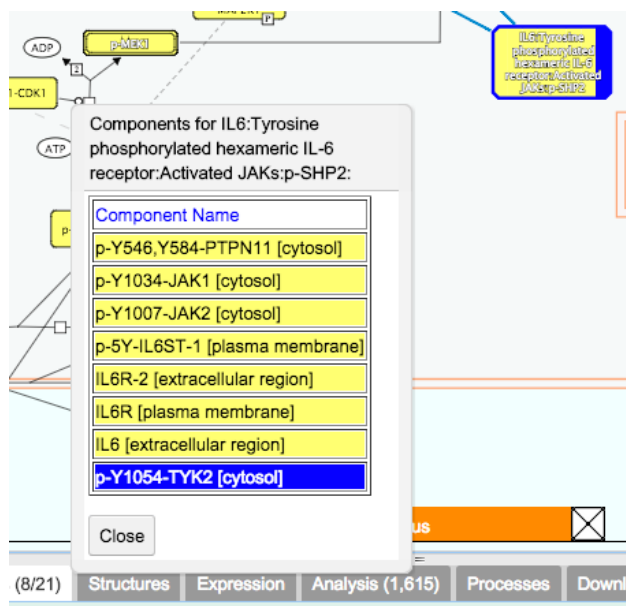
Is this pathway present in mouse (at least inferred). Look at the surrounding reaction network



There are some entities (proteins/complexes and molecules) that are coloured differently. White regions indicates that inference was not possible. This is always the case for small molecules, DNA and other objects that have no UniProt entry (or did not at the time the pathway was constructed). Objects with bands of colour represent complexes or sets containing more than one molecule. The bands of colour reflect the inference success for the molecules within the complex/set.

To view species comparison results for a complex or set right click it and select the option Display Participating Molecules. This reveals a table representing all the proteins involved in the complex/set. Each square in the grid represents one component of the complex/set, coloured as described above.





Finally, it is also possible to test whether a list of proteins are random, or are enriched for a particular pathway. Paste the following list of accession into

P27695  
Q13216  
Q16531  
P19388  
P36954  
P62875  
P23025  
P19447  
Q01831  
P18074  
Q92889  
P28715  
P51948  
P50613  
P51946  
P49005  
P27694  
P15927  
P35244  
P35251

### Analysis Tools

This tool merges pathway identifier mapping, overrepresentation and expression analysis into a single tabbed data analysis portal, with integrated visualization and summary features.

Select a file from your computer and click on the "Analyse" button to perform the analysis.

Select data file for analysis     No file chosen     Project to human

▼ Click here to paste your data or try example data sets...

Paste the data to analyse

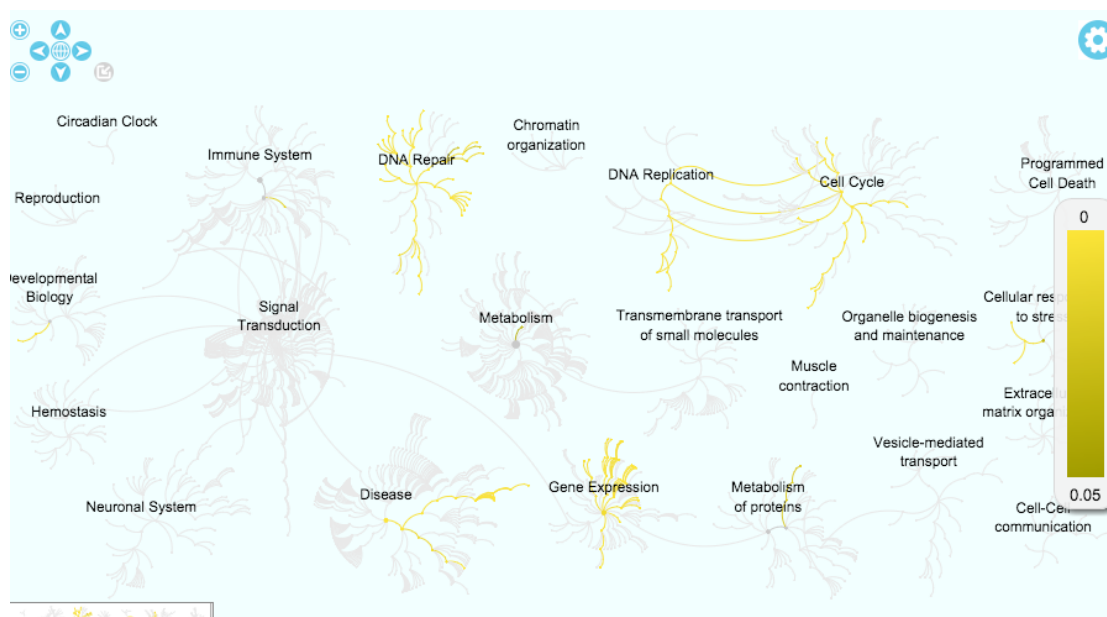
P27695  
 Q13216  
 Q16531  
 P19388  
 P36954  
 P62875  
 P23025  
 P19447  
 Q01831  
 P18074  
 Q92889  
 P28715  
 P51948  
 P50613  
 P51946  
 P40000

Project to human

Some examples:

- 
- 
- 
- 
- 
- 
-

Then click 'Analyse'.

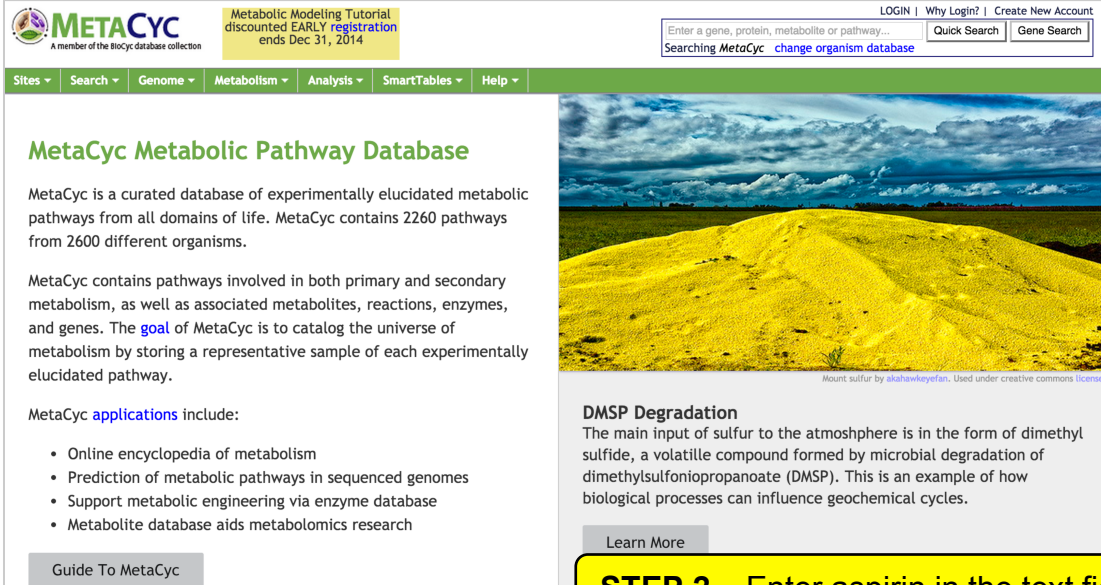


This shows the pathways where the proteins match and in this particular case that the highlighted pathways are over-represented in the set. You can then use the table below, graphic or left menu to investigate these pathways further.

**MetaCyc** – The MetaCyc database is a comprehensive and freely accessible database describing metabolic pathways and enzymes from all domains of

life. MetaCyc pathways are experimentally determined, mostly small-molecule metabolic pathways and are curated from the primary scientific literature.

**STEP 1 – Go to the MetaCyc homepage at:**  
<http://www.metacyc.org>



**MetaCyc Metabolic Pathway Database**

MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life. MetaCyc contains 2260 pathways from 2600 different organisms.

MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes. The **goal** of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each experimentally elucidated pathway.

MetaCyc **applications** include:

- Online encyclopedia of metabolism
- Prediction of metabolic pathways in sequenced genomes
- Support metabolic engineering via enzyme database
- Metabolite database aids metabolomics research

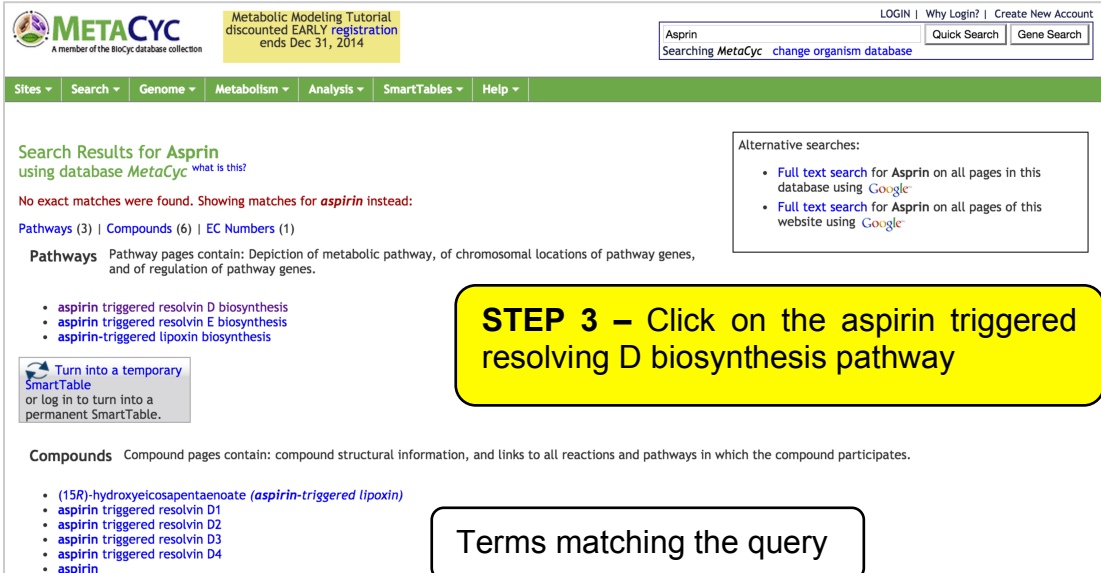
[Learn More](#)

**DMSP Degradation**  
 The main input of sulfur to the atmosphere is in the form of dimethyl sulfide, a volatile compound formed by microbial degradation of dimethylsulfoniopropanoate (DMS). This is an example of how biological processes can influence geochemical cycles.

[Learn More](#)

**STEP 2 – Enter aspirin in the text field**

Rather than looking at a particular gene as we did with Reactome, lets considered a widely used drug – aspirin.



**Search Results for Aspirin**  
 using database *MetaCyc* [what is this?](#)

No exact matches were found. Showing matches for *aspirin* instead:

**Pathways** (3) | **Compounds** (6) | **EC Numbers** (1)

Pathway pages contain: Depiction of metabolic pathway, of chromosomal locations of pathway genes, and of regulation of pathway genes.

- aspirin triggered resolin D biosynthesis
- aspirin triggered resolin E biosynthesis
- aspirin-triggered lipoxin biosynthesis

[Turn into a temporary SmartTable](#)  
 or log in to turn into a permanent SmartTable.

**Compounds** Compound pages contain: compound structural information, and links to all reactions and pathways in which the compound participates.

- (15R)-hydroxyicosapentaenoate (*aspirin-triggered lipoxin*)
- aspirin triggered resolin D1
- aspirin triggered resolin D2
- aspirin triggered resolin D3
- aspirin triggered resolin D4
- aspirin

**Alternative searches:**

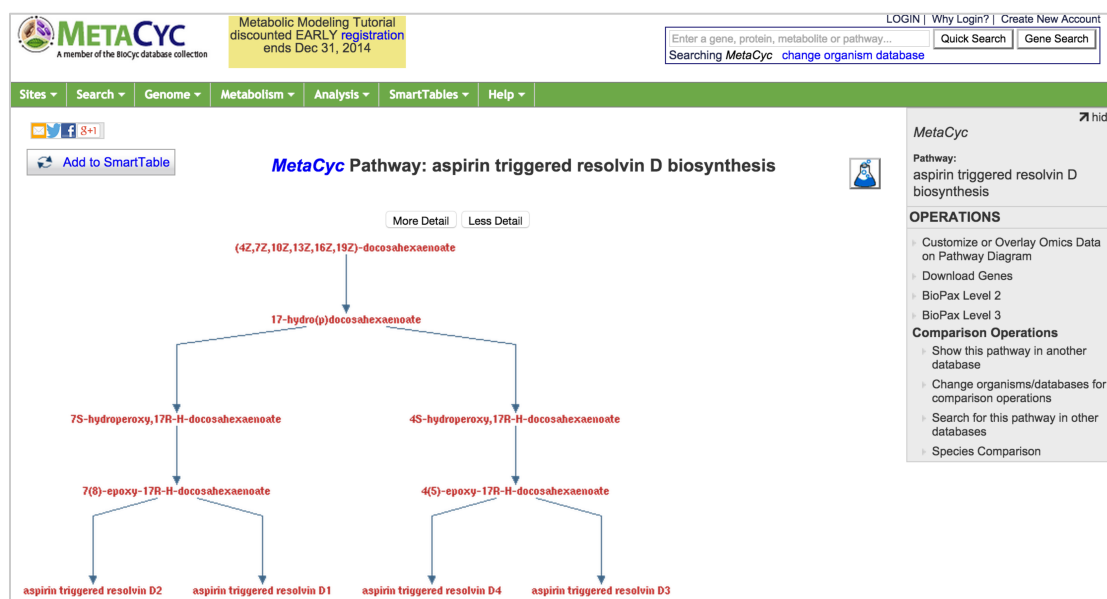
- Full text search for Aspirin on all pages in this database using [Google](#)
- Full text search for Aspirin on all pages of this website using [Google](#)

**STEP 3 – Click on the aspirin triggered resolving D biosynthesis pathway**

Terms matching the query

The page above lists all pathways, compounds and EC numbers that have matched the query term. Clicking the link in step three, produces the following page, which has been broken down into a series of parts for this module.

The top of the page shows the pathway for the generation of resolving D, which has been triggered by aspirin. The first enzyme if the pathway contains the aspirin acetylated COX2 enzyme.



Below this section, is a description of the pathway and the literature references used in generating the pathway.



**KEGG** – The KEGG database may represent one of the best-known pathway databases, contains a description of cellular pathways. However, the future of the database is unclear, so we are no longer presenting this database. However, for your information, we have retained the KEGG pathway information as a supplementary, to enable you to still understand the features of this resource. KEGG is more commonly used to analyse metabolic pathways, but it also contains disease related pathways. In the following **worked example** you will be shown how to find information on disease related pathways.

**STEP 1** – Go to the KEGG homepage at:

<http://www.genome.ad.jp/kegg/pathway.html>



## KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

KEGG2 PATHWAY BRITE MODULE DISEASE DRUG KO GENES GENOME LIGAND DBGET

Select prefix   Enter keywords   [Help](#)

---

### Pathway Maps

**KEGG PATHWAY** is a collection of manually drawn pathway maps (see [new maps](#), [change history](#), and [last updates](#)) representing our knowledge on the molecular interaction and reaction networks for:

- 0. Global Map**
- 1. Metabolism**
  - [Carbohydrate](#) [Energy](#) [Lipid](#) [Nucleotide](#) [Amino acid](#) [Other amino acid](#) [Glycan](#)
  - [Cofactor/vitamin](#) [Terpenoid/PK](#) [Other secondary metabolite](#) [Xenobiotics](#) [Overview](#)
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Organismal Systems**
- 6. Human Diseases** ←
- and also on the structure relationships (KEGG drug structure)
- 7. Drug Development**

### Pathway Mapping

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretation of higher-level systemic functions.

- [Search Pathway](#) - basic pathway mapping tool
- [Search&Color Pathway](#) - advanced pathway mapping tool
- [Color Pathway](#) - selected pathway map coloring tool

**STEP 2** – Select 'human diseases'

## 6. Human Diseases

### 6.1 Cancers: Overview

- Pathways in cancer
- Transcriptional misregulation in cancer
- MicroRNAs in cancer *New!*
- Proteoglycans in cancer
- Chemical carcinogenesis
- Viral carcinogenesis

KEGG DISEASE  
KEGG Cancer

Human diseases

**STEP 3 – Select ‘thyroid cancer’**

### 6.2 Cancers: Specific types

- Colorectal cancer
- Pancreatic cancer
- Glioma
- Thyroid cancer
- Acute myeloid leukemia
- Chronic myeloid leukemia
- Basal cell carcinoma
- Melanoma
- Renal cell carcinoma
- Bladder cancer
- Prostate cancer
- Endometrial cancer
- Small cell lung cancer
- Non-small cell lung cancer

## KEGG Thyroid cancer - Homo sapiens (human)

Help

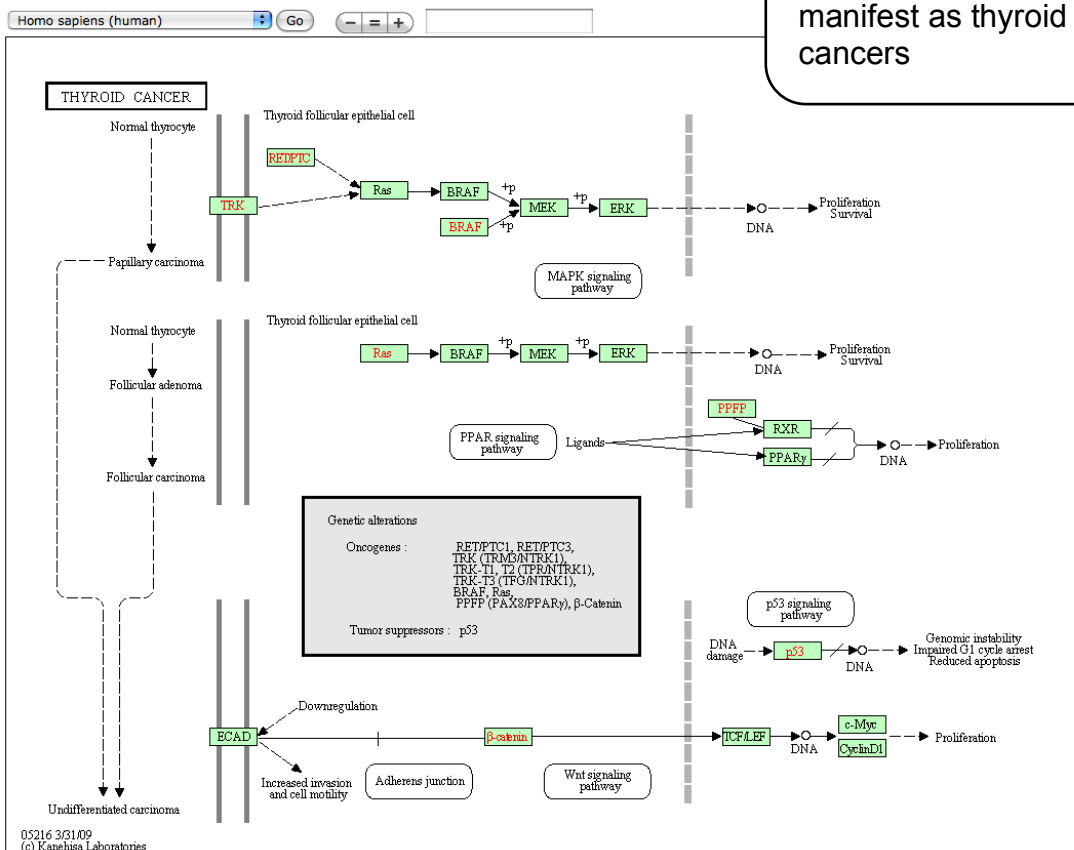
[ Pathway menu | Pathway entry | Hide description ]

Papillary thyroid carcinoma (PTC), the most frequent neoplasia originating from the thyroid epithelium, accounts for about 80% of all thyroid cancers. Chimeric oncogenes, created by chromosomal rearrangements involving prevalently RET and, to a less extent, NTRK1 loci, are implicated in the development of papillary carcinoma. These are inappropriately expressed and stimulate constitutive signaling, bypassing the need for receptor activation by growth factors. Alternatively, mutant RAS directly stimulates BRAF, whereas mutant BRAF directly stimulates MEK.

Of all thyroid cancers, 15-20% are follicular thyroid carcinoma (FTC). The most distinctive molecular features of follicular carcinoma are the prominence of aneuploidy and the high prevalence of RAS mutations and PAX8-PPAR-gamma rearrangements. The PPAR-gamma rearrangement functions through a dominant-negative effect on the transcriptional activity of wild-type PPAR-gamma. The fusion oncoprotein contributes to malignant transformation by targeting several cellular pathways, some of which are normally engaged by PPAR-gamma.

Most poorly differentiated and undifferentiated thyroid carcinomas are considered to derive from pre-existing well-differentiated thyroid carcinoma through additional genetic events, including beta-catenin nuclear accumulation and inactivation, but de novo occurrence might also occur.

Different pathways where mutations in constituent proteins manifest as thyroid cancers







KEGG - Table of Contents

KEGG2 PATHWAY BRITE MODULE DISEASE DRUG KO GENOME GENES LIGAND DBGET

Search  for

Category	Entry Point	Release Info	Search & Compute
Systems information	<a href="#">KEGG PATHWAY</a> <a href="#">KEGG BRITE</a> <a href="#">KEGG MODULE</a> <a href="#">KEGG Mapper</a> <a href="#">KEGG Atlas</a>	New maps Update history New hierarchies Update history	<a href="#">Search Pathway</a> <a href="#">Search Brite</a> <a href="#">Search Module</a> KEGG pathway maps BRITE functional hierarchies KEGG modules
	<a href="#">KEGG DISEASE</a> <a href="#">KEGG DRUG</a> <a href="#">KEGG ENVIRON</a> <a href="#">KEGG MEDICUS</a>	New drug maps Update history	Human diseases Infectious diseases ATC drug classification
Genomic information	<a href="#">KEGG ORTHOLOGY</a>		<a href="#">KEGG Orthology (KO)</a>
	<a href="#">KEGG GENES</a> <a href="#">KEGG GENOME</a> <a href="#">KEGG Organisms</a>	New organisms Update history	<a href="#">SSDB search</a> <a href="#">BLAST / FASTA search</a> KAAS automatic annotation Map organisms to taxonomy Generate taxonomy tree KEGG organisms
Chemical information	<a href="#">KEGG LIGAND</a> <a href="#">KEGG COMPOUND</a> <a href="#">KEGG GLYCAN</a> <a href="#">KEGG REACTION</a> <a href="#">Reaction Modules</a>		<a href="#">SIMCOMP / SUBCOMP search</a> <a href="#">KCaM search</a> E-zyme reaction prediction PathPred pathway prediction PathComp path computation PathSearch reaction search

See Kanehisa et al. (2012) for the new features of KEGG.

**KEGG for specific organisms**

**KEGG Organisms** - the list of currently available organisms

Select    (examples) hsa mmu sce eco bsu syn

**KEGG Pangenomes** - the list of pangenomes defined from KEGG organisms

It is also possible to search KEGG using keywords/genes name. Select KEGG2 from the home page and enter the term in the textfield.

Alternative search strategies, such as BLAST, are shown under 'search and compute'



**TASKS**

- 1) Search the sequence P52647 against the three different protein domain databases (Pfam and InterPro) outlined in the manual and appendix. How do they differ?  
Tip: Compare <http://pfam.xfam.org/protein/P52647> and <http://www.ebi.ac.uk/interpro/protein/P52647>
  
- 2) Compare the ligand interactions found in PDBsum and with the ligand interaction view from PDB for the structure 2dq7. Are they the same? Which are the most important interactions?
  
- 3) Using PDBsum, find the cleft where the ligand is bound in the structure 2dq7. What is the size of the cleft?
  
- 4) Perform homology modelling for the sequence P14056 using the template structure 2src, chain A. Look at the structure and the quality graphs. How do they compare to the automatically chosen template?
  
- 5) Look at the aspirin example in MetaCyc. How is taking aspirin useful after a stroke?

**Answers**

1. InterPro integrates many different protein family databases. Each database has a different take on protein families. Some describe protein domains, others provide protein functions, and others provide sites/motifs. InterPro gives you the access to the complete repertoire of annotations. Pfam is found within this set of annotations. Many of the databases agree on the domain definitions, and have been group together accordingly.
2. They are not the same. The most important bonds are those higher order hydrogen bonds that are consistently called between the two different sites. The weaker Van der Waals differ, but this is based on parameters/cut-off used for the calculation of these bonds.
3. This is the largest cleft, at  $4195.97\text{\AA}^3$ .
4. The model 2src does not have as high percentage identity to the query (just over 30% identity), so there are fewer confidently modelled regions. This is level of sequence identity is about the limit of what is useful for homology modelling, as there is such little confidence in the model, especially towards the C-terminus of the model.
5. Aspirin cause the acetylations of COX2 to produce the 17R resolvins facilitates the inhibition of both leukocyte infiltration and cytokine expression, which cause an inflammatory response after a stroke and result in tissue damage.