I - Summary

- Introduction to protein domains
- Domain databases

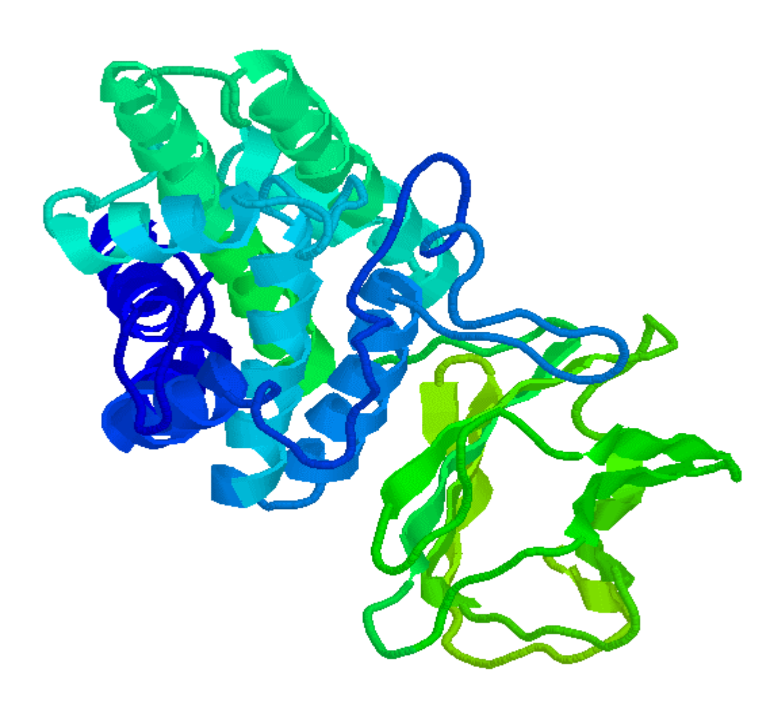
http://www.sanger.ac.uk/Software/Pfam/

Protein Domains

- From a structural perspective protein domains are discrete units.

http://www.sanger.ac.uk/Software/Pfam/

What is a Domain?

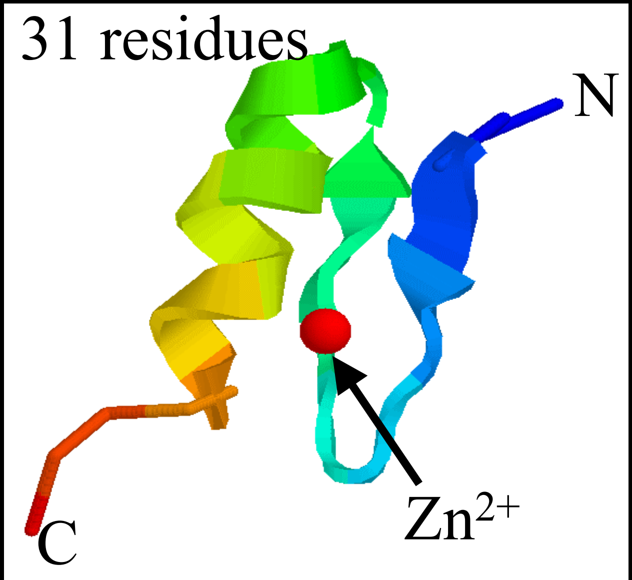


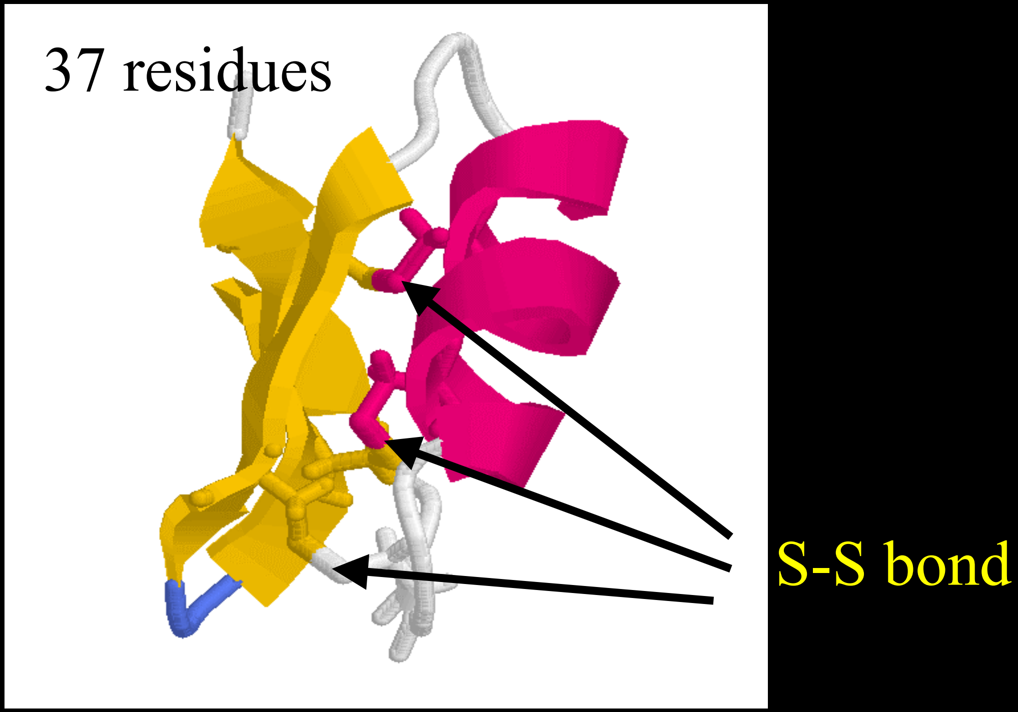
- Defined by structure
- Domain boundaries can be inferred from careful sequence analysis
- Domains are the currency of protein function

http://www.sanger.ac.uk/Software/Pfam/

Domains - size

- Domains can be 25 to 500 residues long
- Most are less than 200 residues.
- Domains can be smaller that 50 residues but these are stabilized by disulphide bonds or chelated metals.

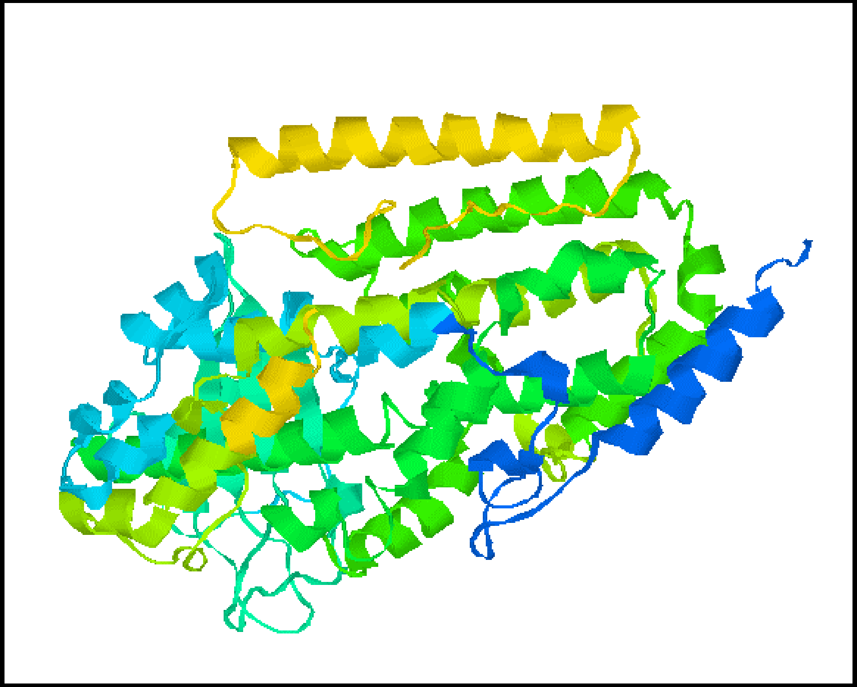




http://www.sanger.ac.uk/Software/Pfam/

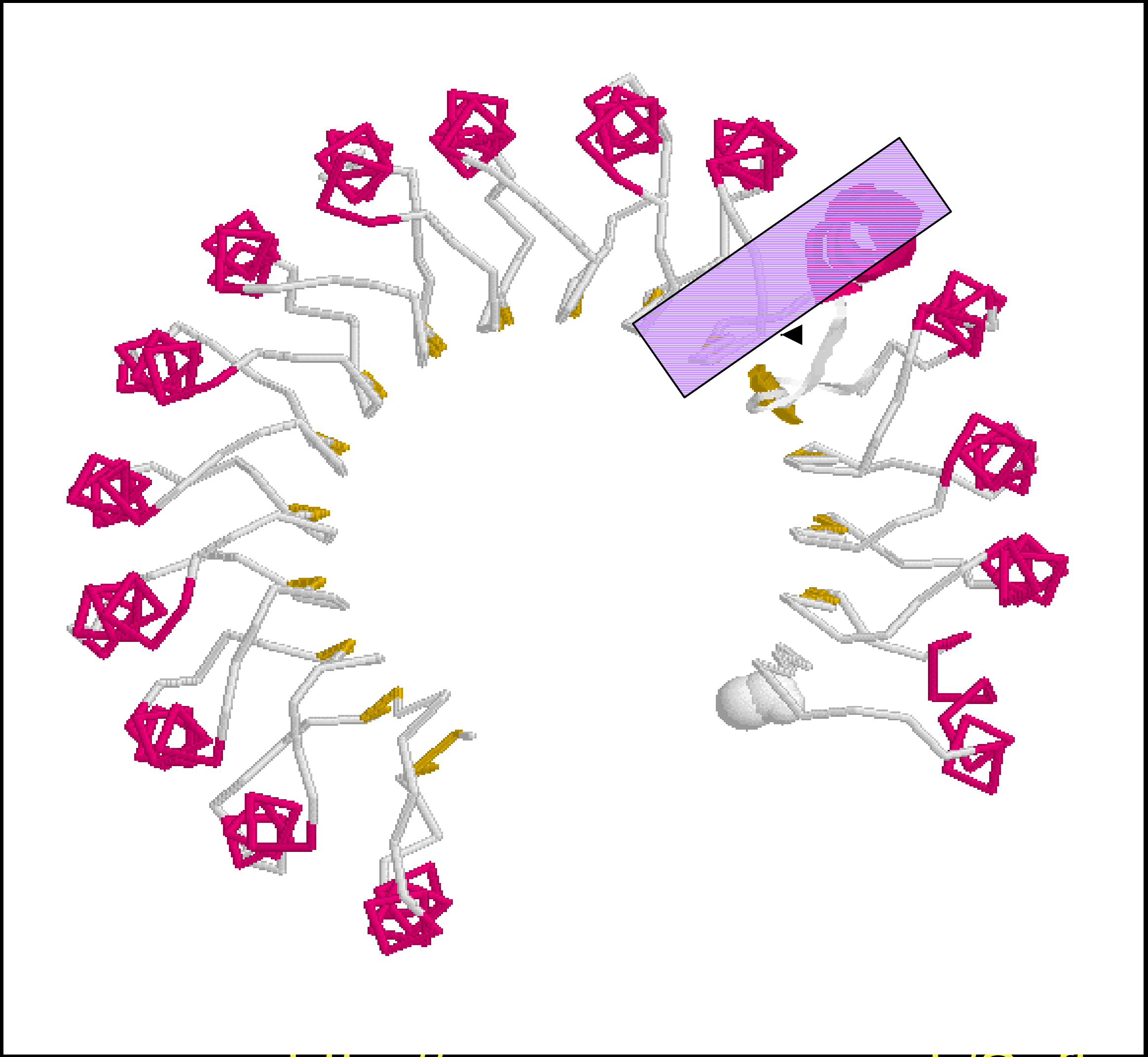
Example domains

- The lipoxygenase domain is a giant at 500 residues long.



http://www.sanger.ac.uk/Software/Pfam/

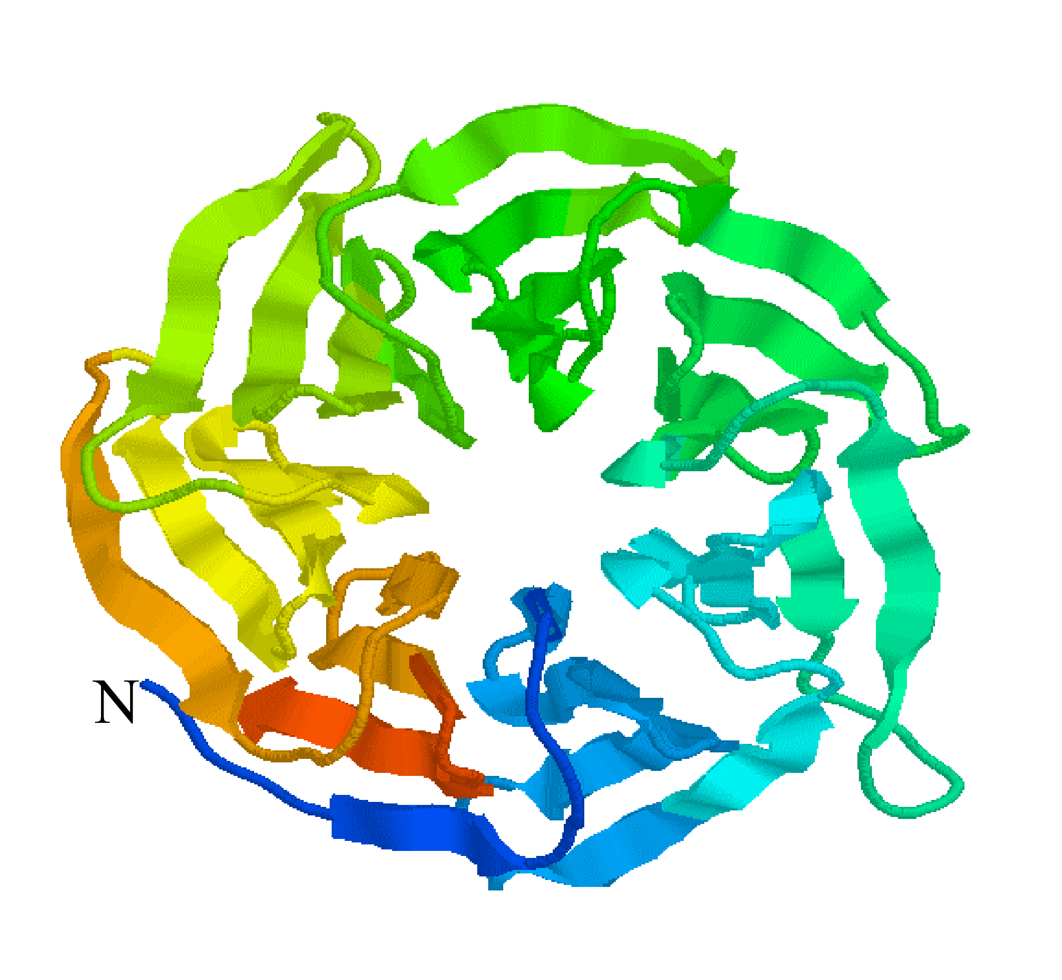
Leucine Rich repeats



- A single repeat is not stable
- Multiple repeats are stable
- Each repeat is represented separately
- Unlimited number

http://www.sanger.ac.uk/Software/Pfam/

WD40 repeats



- 7 repeats
- beta sheet per repeat
- Limited number (6-8)

http://www.sanger.ac.uk/Software/Pfam/

Structural domains

- Domains are most easily defined in known structures
- Several automatic programs available
- They don’t always/often agree!

http://www.sanger.ac.uk/Software/Pfam/

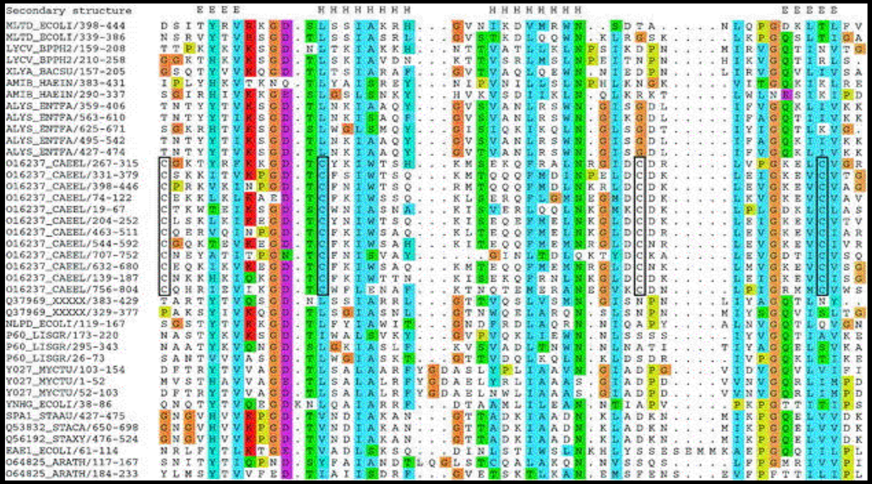
Defining domains from sequence

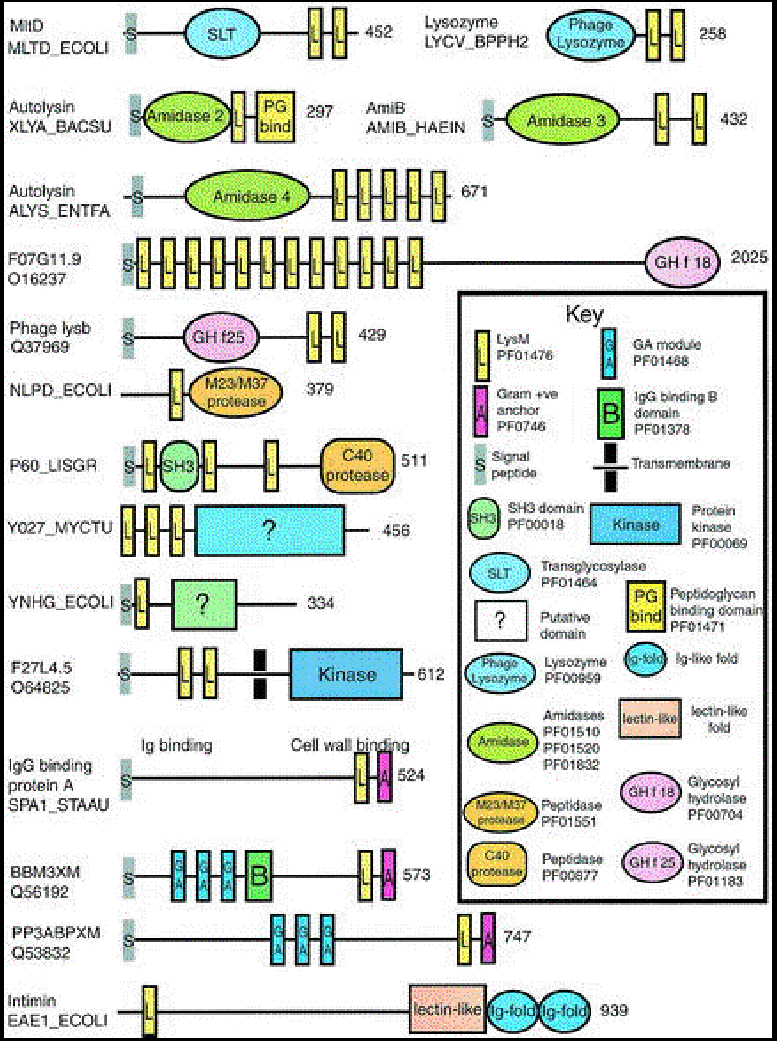
- Has been done successfully hundreds of times
- Cannot always be done
- Usually requires the domain to be mobile

http://www.sanger.ac.uk/Software/Pfam/

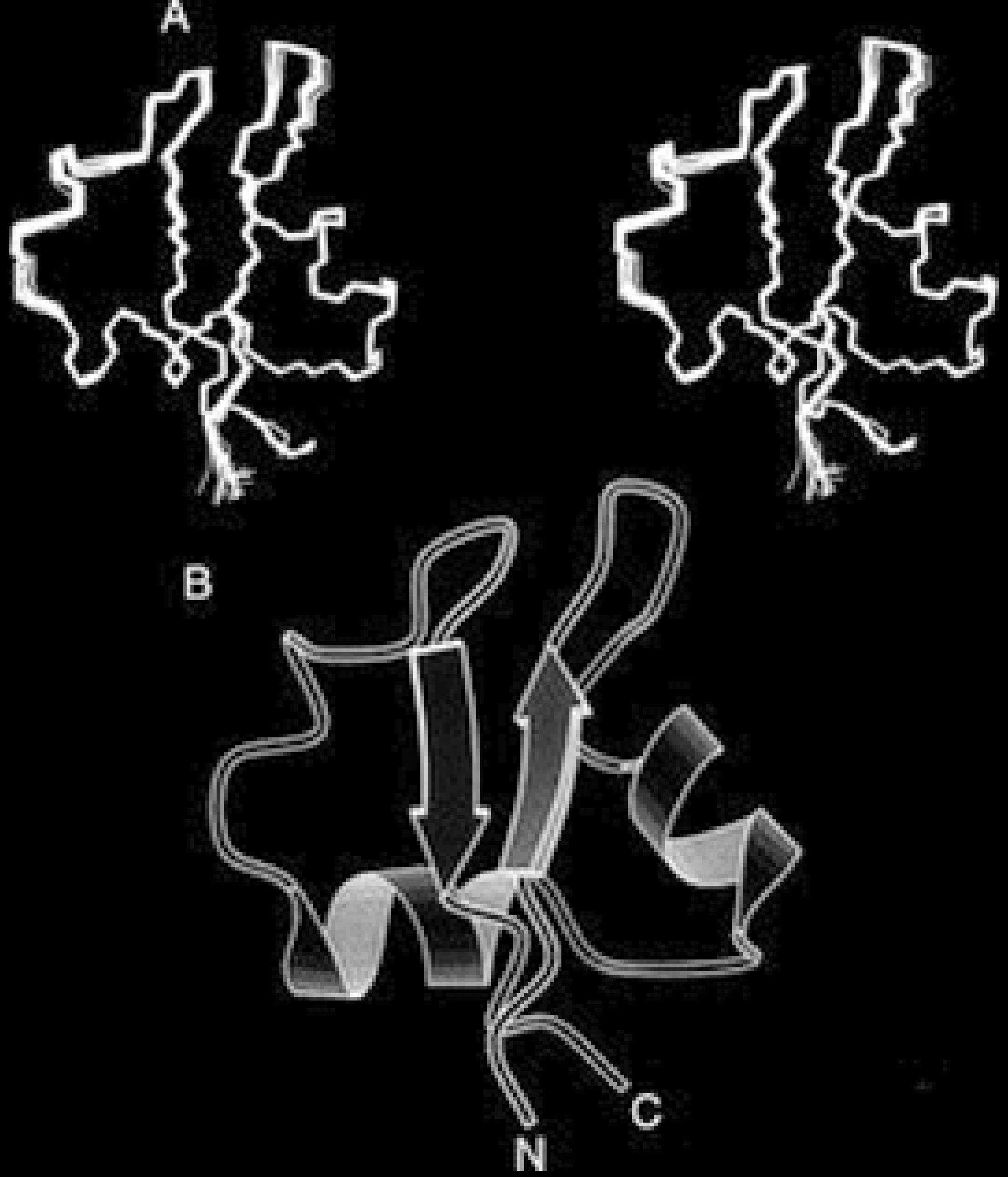
Domains and structure determination

- Hard to get structure of complete protein
- Expressing smaller segments is easier





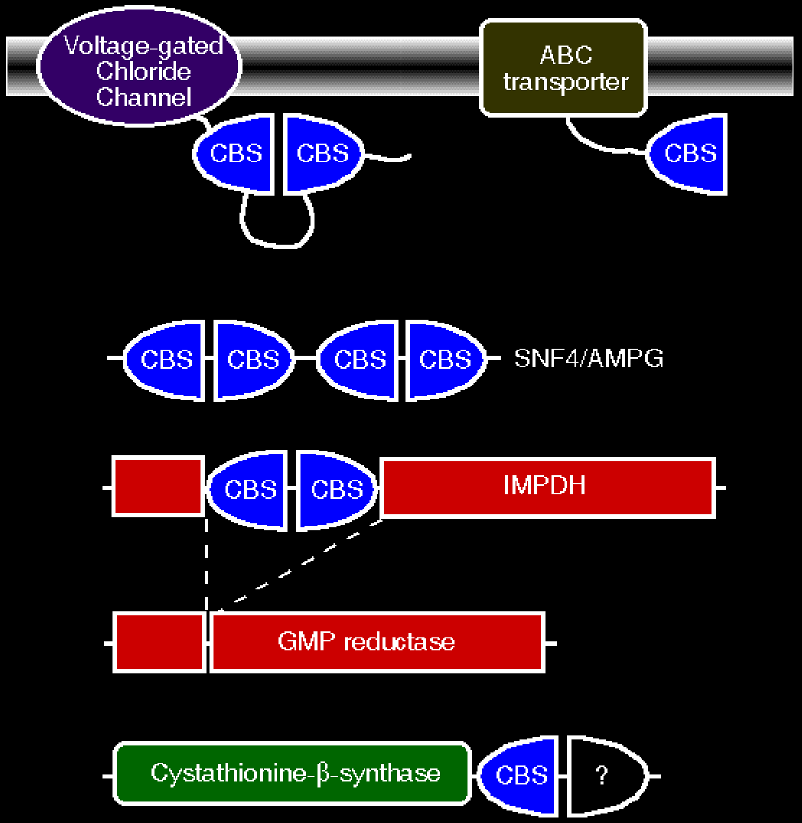
http://www.sanger.ac.uk/Software/Pfam/

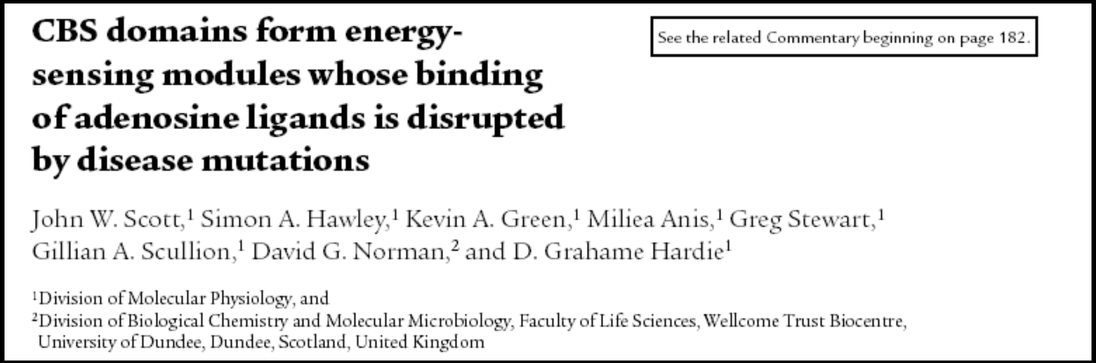


http://www.sanger.ac.uk/Software/Pfam/

Domain Hunting: CBS domains

- Discovering new domains can reveal new biology





CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations

See the related Commentary beginning on page 182.

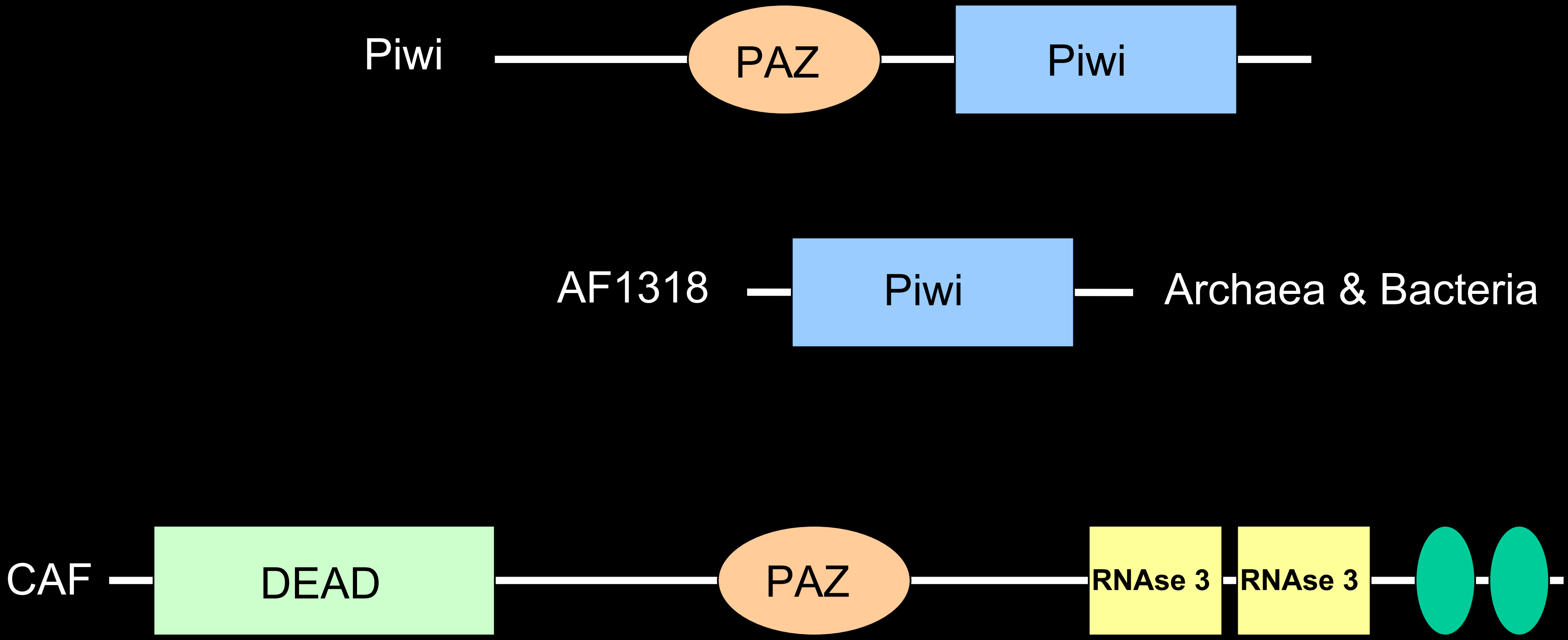
John W. Scott,[1] Simon A. Hawley,[1] Kevin A. Green,[1] Miliea Anis,[1] Greg Stewart,[1] Gillian A. Scullion,[1] David G. Norman,[2] and D. Grahame Hardie[1]

[1]Division of Molecular Physiology, and
[2]Division of Biological Chemistry and Molecular Microbiology, Faculty of Life Sciences, Wellcome Trust Biocentre, University of Dundee, Dundee, Scotland, United Kingdom

J. Clin. Inv. 113:274-284.

http://www.sanger.ac.uk/Software/Pfam/

Domain Hunting: RNAi



Cerrutti, Mian & Bateman. Trends Biochem Sci. 25:481-482 (2000)

http://www.sanger.ac.uk/Software/Pfam/

I - Summary

- Introduction to protein domains
- Domain databases

http://www.sanger.ac.uk/Software/Pfam/

Domain databases

- Many of the common domains have already been defined in domain databases.

- Advantages:
  - Pre-annotated domains
  - Easy interpretation of domain structure
  - Sensitivity can be higher

- The most used databases are:
  - Pfam
  - Prosite Profiles
  - SMART
  - Prints
  - Blocks
  - ProDom

http://www.sanger.ac.uk/Software/Pfam/

Pfam

- Good coverage
- No specific bias
- Good graphical views
- Structural data in alignments
- No heirarchy

http://www.sanger.ac.uk/Software/Pfam/

- Domain collection by Ponting and Bork.
- Specialises in
  - Signaling domains
  - Extracellular domains
  - Nuclear domains
- Excellent quality families.
- Really nice graphics
- Coiled-coil, TM, low-complexity

http://www.sanger.ac.uk/Software/Pfam/

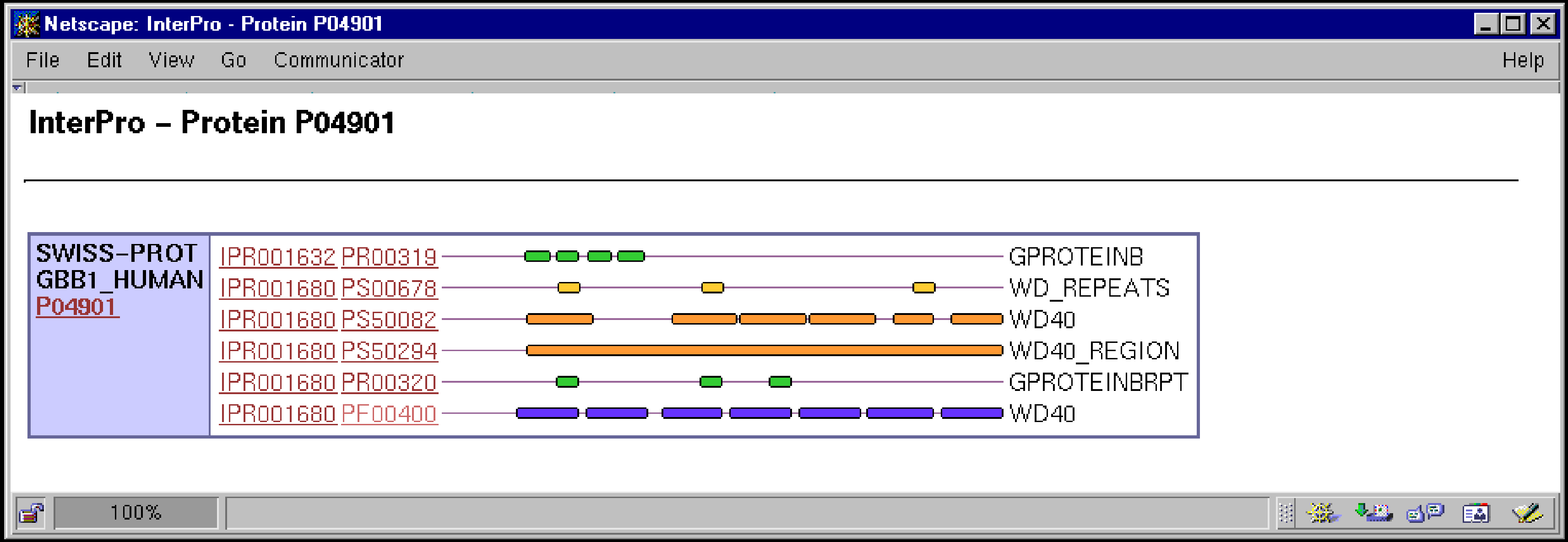
prosite

- Profiles
  - – Sensitive
  - – Low coverage (Good for signalling)
- Patterns
  - – e.g.   N-{P}-[ST]-{P}
  - – less sensitive
  - – many false positives

http://www.sanger.ac.uk/Software/Pfam/

Interpro

- Interpro is a database that presents Prosite, Prints, Prodom and Pfam domain.
- Annotation is a strong point



InterPro – Protein P04901

SWISS-PROT GBB1_HUMAN P04901	IPR001632	PR00319	GPROTEINB
	IPR001680	PS00678	WD_REPEATS
	IPR001680	PS50082	WD40
	IPR001680	PS50294	WD40_REGION
	IPR001680	PR00320	GPROTEINBRPT
	IPR001680	PF00400	WD40

http://www.sanger.ac.uk/Software/Pfam/

Conclusions

- Domains are the common currency of protein function
- Understanding the domain structure helps to understand the biology
- Domain databases are key labour saving tools

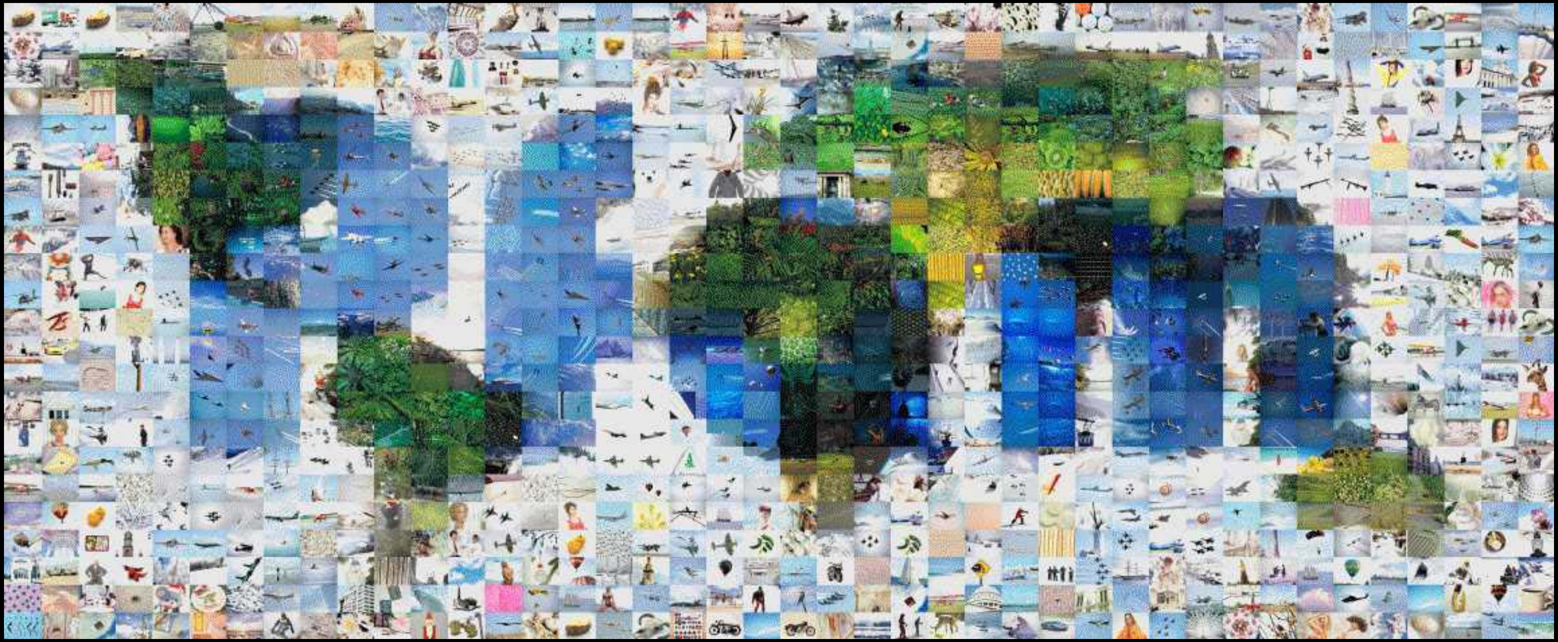
http://www.sanger.ac.uk/Software/Pfam/

II - Summary

- <u>Introduction to Pfam</u>
- Protein Interactions
- Pfam Clans

http://www.sanger.ac.uk/Software/Pfam/

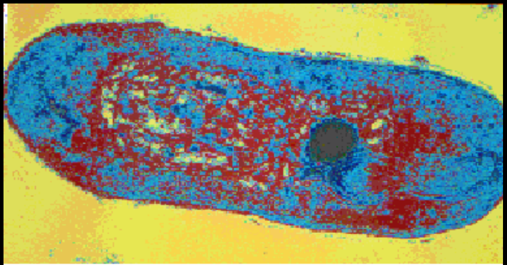
Pfam: 8,000 families for the molecular biologist

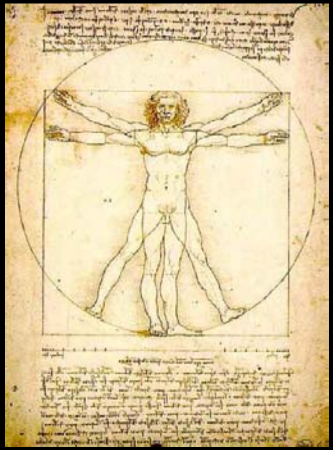


Alex Bateman, Richard Durbin, Sean Eddy, Ajay Khanna, Rob Finn, Sam Griffiths-Jones, Jaina Mistry, John Tate, Volker Hollich and Erik Sonnhammer.

http://www.sanger.ac.uk/Software/Pfam/

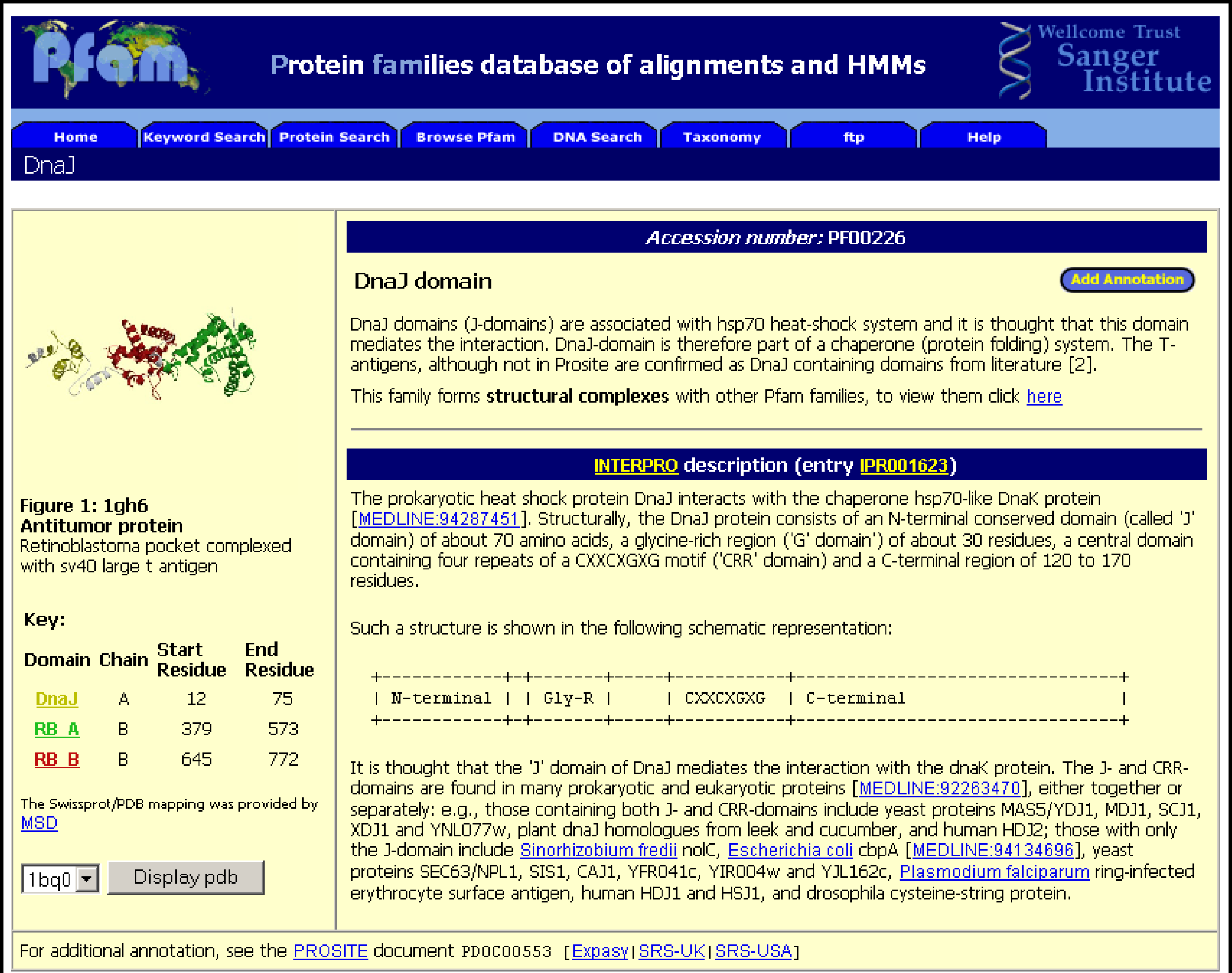
Annotating genomes





http://www.sanger.ac.uk/Software/Pfam/

Family Pages



DnaJ

Accession number: PF00226

DnaJ domain

DnaJ domains (J-domains) are associated with hsp70 heat-shock system and it is thought that this domain mediates the interaction. DnaJ-domain is therefore part of a chaperone (protein folding) system. The T-antigens, although not in Prosite are confirmed as DnaJ containing domains from literature [2].

This family forms **structural complexes** with other Pfam families, to view them click here

INTERPRO description (entry IPR001623)

The prokaryotic heat shock protein DnaJ interacts with the chaperone hsp70-like DnaK protein [MEDLINE:94287451]. Structurally, the DnaJ protein consists of an N-terminal conserved domain (called 'J' domain) of about 70 amino acids, a glycine-rich region ('G' domain') of about 30 residues, a central domain containing four repeats of a CXXCXGXG motif ('CRR' domain) and a C-terminal region of 120 to 170 residues.

Such a structure is shown in the following schematic representation:

```
+------------+-+-------+-----+-----------+--------------------------------+
| N-terminal | | Gly-R |     | CXXCXGXG  | C-terminal                     |
+------------+-+-------+-----+-----------+--------------------------------+
```

It is thought that the 'J' domain of DnaJ mediates the interaction with the dnaK protein. The J- and CRR-domains are found in many prokaryotic and eukaryotic proteins [MEDLINE:92263470], either together or separately: e.g., those containing both J- and CRR-domains include yeast proteins MAS5/YDJ1, MDJ1, SCJ1, XDJ1 and YNL077w, plant dnaJ homologues from leek and cucumber, and human HDJ2; those with only the J-domain include Sinorhizobium fredii nolC, Escherichia coli cbpA [MEDLINE:94134696], yeast proteins SEC63/NPL1, SIS1, CAJ1, YFR041c, YIR004w and YJL162c, Plasmodium falciparum ring-infected erythrocyte surface antigen, human HDJ1 and HSJ1, and drosophila cysteine-string protein.

Figure 1: 1gh6
Antitumor protein
Retinoblastoma pocket complexed with sv40 large t antigen

Key:

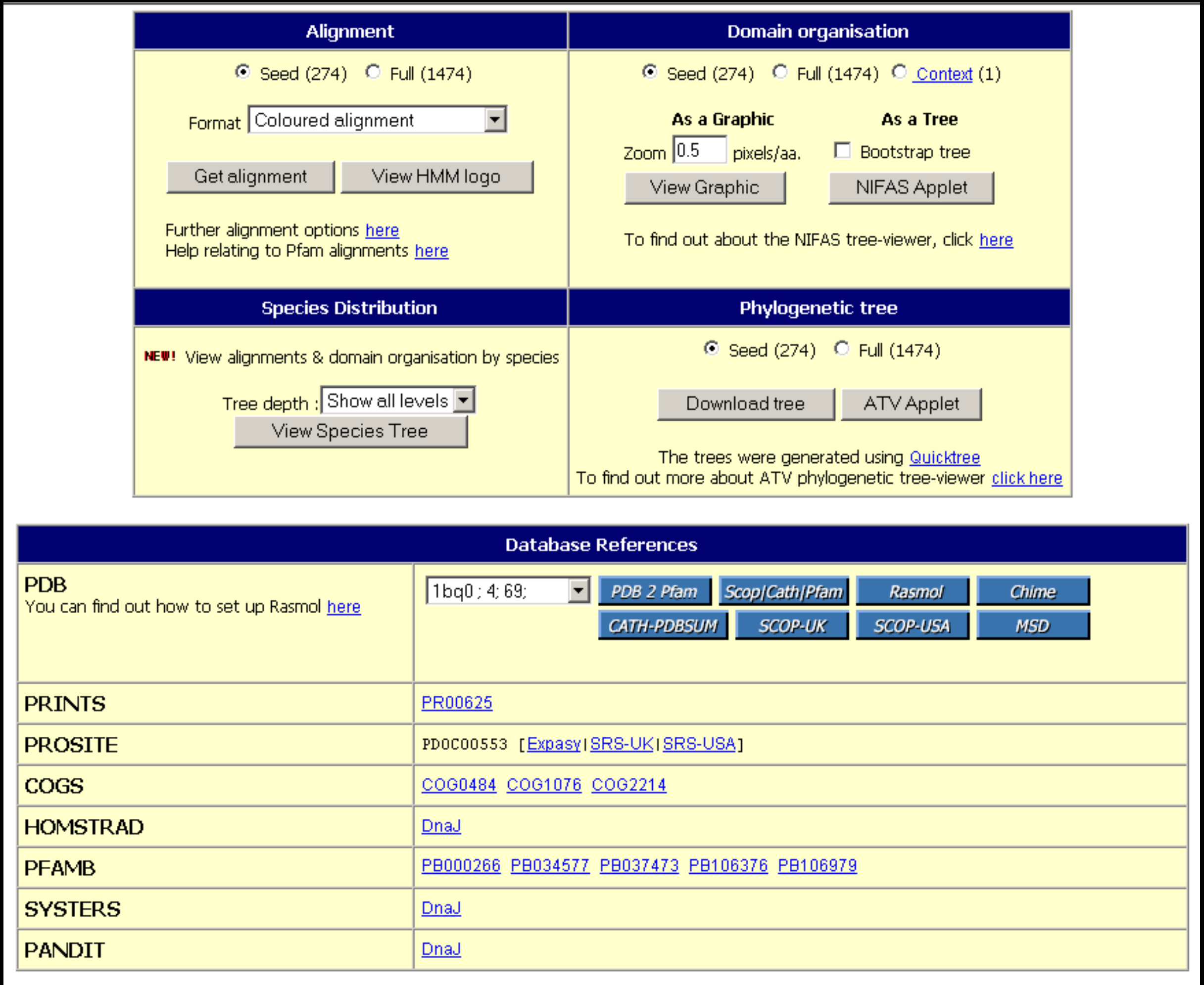
Domain	Chain	Start Residue	End Residue
DnaJ	A	12	75
RB_A	B	379	573
RB_B	B	645	772

The Swissprot/PDB mapping was provided by MSD

For additional annotation, see the PROSITE document PDOC00553 [Expasy|SRS-UK|SRS-USA]

http://www.sanger.ac.uk/Software/Pfam/

Family Pages



Alignment

● Seed (274) ○ Full (1474)

Format Coloured alignment

Further alignment options here
Help relating to Pfam alignments here

Domain organisation

● Seed (274) ○ Full (1474) ○ Context (1)

As a Graphic
Zoom 0.5 pixels/aa.

As a Tree
□ Bootstrap tree

To find out about the NIFAS tree-viewer, click here

Species Distribution

NEW! View alignments & domain organisation by species

Tree depth : Show all levels

Phylogenetic tree

● Seed (274) ○ Full (1474)

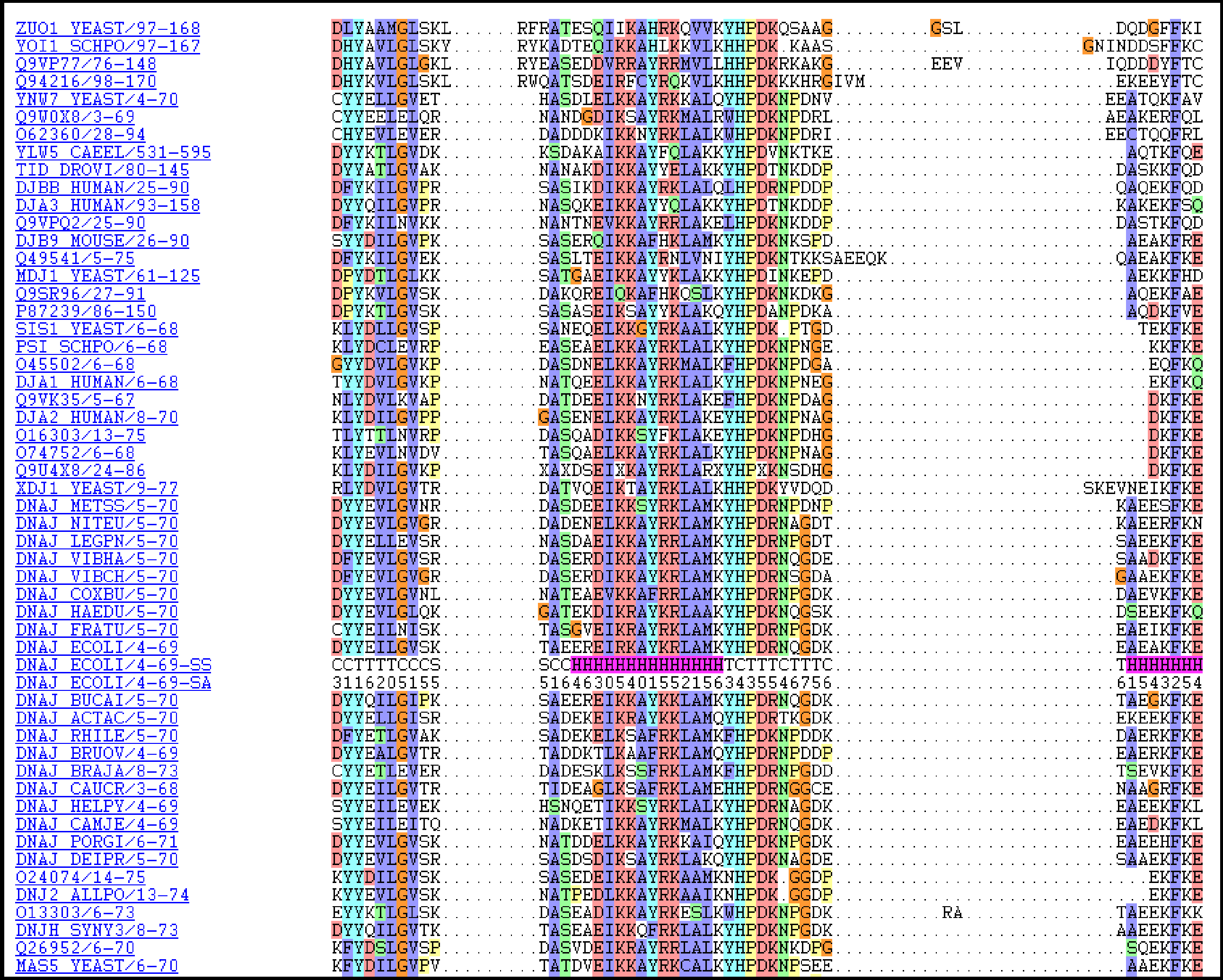
The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here

Database References

PDB You can find out how to set up Rasmol here	1bq0 ; 4; 69; PDB 2 Pfam · Scop/Cath/Pfam · Rasmol · Chime CATH-PDBSUM · SCOP-UK · SCOP-USA · MSD
PRINTS	PR00625
PROSITE	PDOC00553 [Expasy|SRS-UK|SRS-USA]
COGS	COG0484 COG1076 COG2214
HOMSTRAD	DnaJ
PFAMB	PB000266 PB034577 PB037473 PB106376 PB106979
SYSTERS	DnaJ
PANDIT	DnaJ

http://www.sanger.ac.uk/Software/Pfam/

Pfam contains Alignments



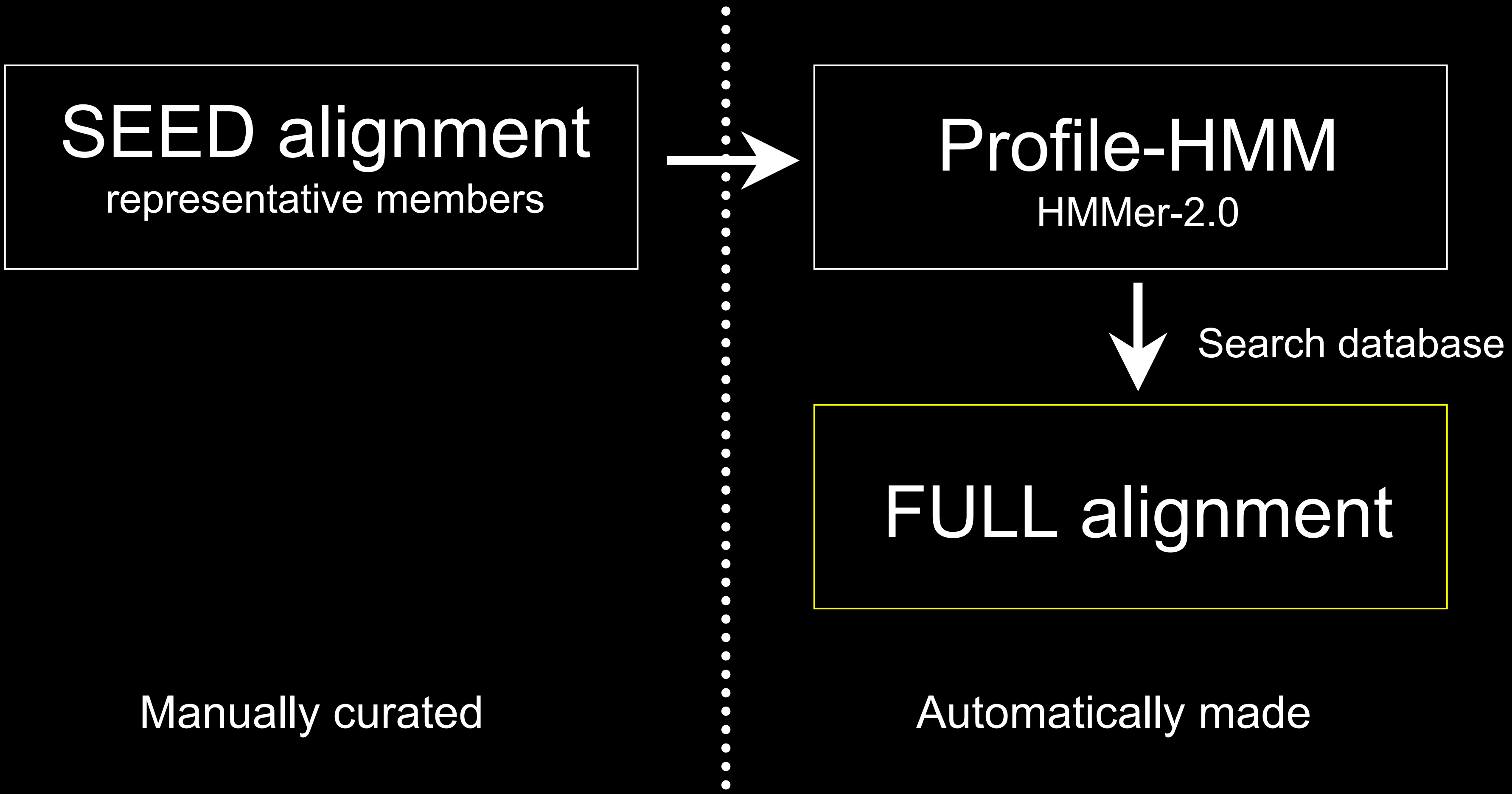
```
ZUO1_YEAST/97-168     DLYAAMGLSKL......RFRATESQIIKAHRKQVVKYHPDKQSAAG.........GSL..............DQDGFFKI
YOI1_SCHPO/97-167     DHYAVLGLSKY......RYKADTEQIKKAHLKKVLKHHPDK.KAAS.......................GNINDDSFFKC
Q9VP77/76-148         DHYAVLGLGKL......RYEASEDDVRRAYRRMVLLHHPDKRKAKG.........EEV.............IQDDDYFTC
Q94216/98-170         DHYKVLGLSKL......RWQATSDEIRFCYRQKVLKHHPDKKKHRGIVM.......................EKEEYFTC
YNW7_YEAST/4-70       CYYELLGVET.........HASDLELKKAYRKKALQYHPDKNPDNV.........................EEATQKFAV
Q9W0X8/3-69           CYYEELELQR.........NANDGDIKSAYRKMALRWHPDKNPDRL.........................AEAKERFQL
O62360/28-94          CHYEVLEVER.........DADDDKIKKNYRKLALKWHPDKNPDRI.........................EECTQQFRL
YLW5_CAEEL/531-595    DYYKTLGVDK.........KSDAKAIKKAYFQLAKKYHPDVNKTKE...........................AQTKFQE
TID_DROVI/80-145      DYYATLGVAK.........NANAKDIKKAYYELAKKYHPDTNKDDP..........................DASKKFQD
DJBB_HUMAN/25-90      DFYKILGVPR.........SASIKDIKKAYRKLALQLHPDRNPDDP..........................QAQEKFQD
DJA3_HUMAN/93-158     DYYQILGVPR.........NASQKEIKKAYYQLAKKYHPDTNKDDP..........................KAKEKFSQ
Q9VPQ2/25-90          DFYKILNVKK.........NANTNEVKKAYRRLAKELHPDKNKDDP..........................DASTKFQD
DJB9_MOUSE/26-90      SYYDILGVPK.........SASERQIKKAFHKLAMKYHPDKNKSPD...........................AEAKFRE
Q49541/5-75           DFYKILGVEK.........SASLTEIKKAYRNLVNIYHPDKNTKKSAEEQK.....................QAEAKFKE
MDJ1_YEAST/61-125     DPYDTLGLKK.........SATGAEIKKAYYKLAKKYHPDINKEPD...........................AEKKFHD
Q9SR96/27-91          DPYKVLGVSK.........DAKQREIQKAFHKQSLKYHPDKNKDKG...........................AQEKFAE
P87239/86-150         DPYKTLGVSK.........SASASEIKSAYYKLAKQYHPDANPDKA...........................AQDKFVE
SIS1_YEAST/6-68       KLYDLLGVSP.........SANEQELKKGYRKAALKYHPDK.PTGD............................TEKFKE
PSI_SCHPO/6-68        KLYDCLEVRP.........EASEAELKKAYRKLALKYHPDKNPNGE.............................KKFKE
O45502/6-68           GYYDVLGVKP.........DASDNELKKAYRKMALKFHPDKNPDGA.............................EQFKQ
DJA1_HUMAN/6-68       TYYDVLGVKP.........NATQEELKKAYRKLALKYHPDKNPNEG.............................EKFKQ
Q9VK35/5-67           NLYDVLKVAP.........DATDEEIKKNYRKLAKEFHPDKNPDAG.............................DKFKE
DJA2_HUMAN/8-70       KLYDILGVPP.........GASENELKKAYRKLAKEYHPDKNPNAG.............................DKFKE
O16303/13-75          TLYTTLNVRP.........DASQADIKKSYFKLAKEYHPDKNPDHG.............................DKFKE
O74752/6-68           KLYEVLNVDV.........TASQAELKKAYRKLALKYHPDKNPNAG.............................DKFKE
Q9U4X8/24-86          KLYDILGVKP.........XAXDSEIXKAYRKLARXYHPXKNSDHG.............................DKFKE
XDJ1_YEAST/9-77       RLYDVLGVTR.........DATVQEIKTAYRKLALKHHPDKYVDQD.......................SKEVNEIKFKE
DNAJ_METSS/5-70       DYYEVLGVNR.........DASDEEIKKSYRKLAMKYHPDRNPDNP..........................KAEESFKE
DNAJ_NITEU/5-70       DYYEVLGVGR.........DADENELKKAYRKLAMKYHPDRNAGDT..........................KAEERFKN
DNAJ_LEGPN/5-70       DYYELLEVSR.........NASDAEIKKAYRRLAMKYHPDRNPGDT..........................SAEEKFKE
DNAJ_VIBHA/5-70       DFYEVLGVSR.........DASERDIKKAYKRLAMKYHPDRNQGDE..........................SAADKFKE
DNAJ_VIBCH/5-70       DFYEVLGVGR.........DASERDIKKAYKRLAMKYHPDRNSGDA..........................GAAEKFKE
DNAJ_COXBU/5-70       DYYEVLGVNL.........NATEAEVKKAFRRLAMKYHPDRNPGDK..........................DAEVKFKE
DNAJ_HAEDU/5-70       DYYEVLGLQK.........GATEKDIKRAYKRLAAKYHPDKNQGSK..........................DSEEKFKQ
DNAJ_FRATU/5-70       CYYEILNISK.........TASGVEIKRAYRKLAMKYHPDRNPGDK..........................EAEIKFKE
DNAJ_ECOLI/4-69       DYYEILGVSK.........TAEEREIRKAYKRLAMKYHPDRNQGDK..........................EAEAKFKE
DNAJ_ECOLI/4-69-SS    CCTTTTCCCS.........SCCHHHHHHHHHHHHHHTCTTTCTTTC..........................THHHHHHH
DNAJ_ECOLI/4-69-SA    3116205155.........516463054015521563435546756..........................61543254
DNAJ_BUCAI/5-70       DYYQILGIPK.........SAEEREIKKAYKKLAMKYHPDRNQGDK..........................TAEGKFKE
DNAJ_ACTAC/5-70       DYYELLGISR.........SADEKEIKRAYKKLAMQYHPDRTKGDK..........................EKEEKFKE
DNAJ_RHILE/5-70       DFYETLGVAK.........SADEKELKSAFRKLAMKFHPDKNPDDK..........................DAERKFKE
DNAJ_BRUOV/4-69       DYYEALGVTR.........TADDKTLKAAFRKLAMQYHPDRNPDDP..........................EAERKFKE
DNAJ_BRAJA/8-73       CYYETLEVER.........DADESKLKSSFRKLAMKFHPDRNPGDD..........................TSEVKFKE
DNAJ_CAUCR/3-68       DYYEILGVTR.........TIDEAGLKSAFRKLAMEHHPDRNGGCE..........................NAAGRFKE
DNAJ_HELPY/4-69       SYYEILEVEK.........HSNQETIKKSYRKLALKYHPDRNAGDK..........................EAEEKFKL
DNAJ_CAMJE/4-69       SYYEILEITQ.........NADKETIKKAYRKMALKYHPDRNQGDK..........................EAEDKFKL
DNAJ_PORGI/6-71       DYYEVLGVSK.........NATDDELKKAYRKKAIQYHPDKNPGDK..........................EAEEHFKE
DNAJ_DEIPR/5-70       DYYEVLGVSR.........SASDSDIKSAYRKLAKQYHPDKNAGDE..........................SAAEKFKE
O24074/14-75          KYYDILGVSK.........SASEDEIKKAYRKAAMKNHPDK.GGDP.............................EKFKE
DNJ2_ALLPO/13-74      KYYEVLGVSK.........NATPEDLKKAYRKAAIKNHPDK.GGDP.............................EKFKE
O13303/6-73           EYYKTLGLSK.........DASEADIKKAYRKESLKWHPDKNPGDK..........RA..............TAEEKFKK
DNJH_SYNY3/8-73       DYYQILGVTK.........TASEAEIKKQFRKLALKYHPDKNPGDK..........................AAEEKFKE
Q26952/6-70           KFYDSLGVSP.........DASVDEIKRAYRRLALKYHPDKNKDPG...........................SQEKFKE
MAS5_YEAST/6-70       KFYDILGVPV.........TATDVEIKKAYRKCALKYHPDKNPSEE...........................AAEKFKE
```

http://www.sanger.ac.uk/Software/Pfam/

The data deluge

http://www.sanger.ac.uk/Software/Pfam/

Pfam contains:



SEED alignment

representative members

Profile-HMM

HMMer-2.0

Search database

FULL alignment

Manually curated

Automatically made

http://www.sanger.ac.uk/Software/Pfam/

Profiles, HMMs and PSSMs

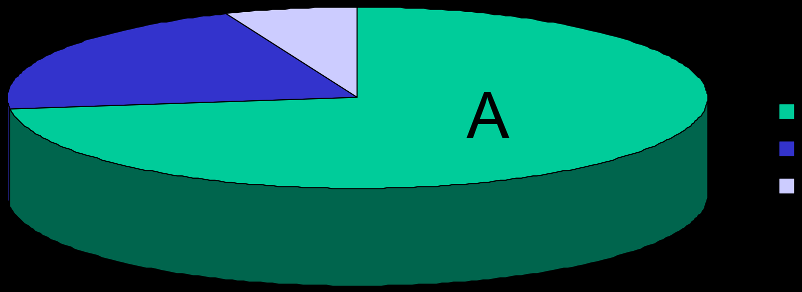
- Complicated names - Simple idea

```
RU1A_HUMAN rrm1   SSATNAL
RU1A_HUMAN rrm2   VQAGAAR
SFR1_HUMAN rrm1   RDAEDAV
SXLF_DROME rrm1   MDSQRAI
                  |||||||
PABP_DROME rrm3   EAAEAAV
```

http://www.sanger.ac.uk/Software/Pfam/

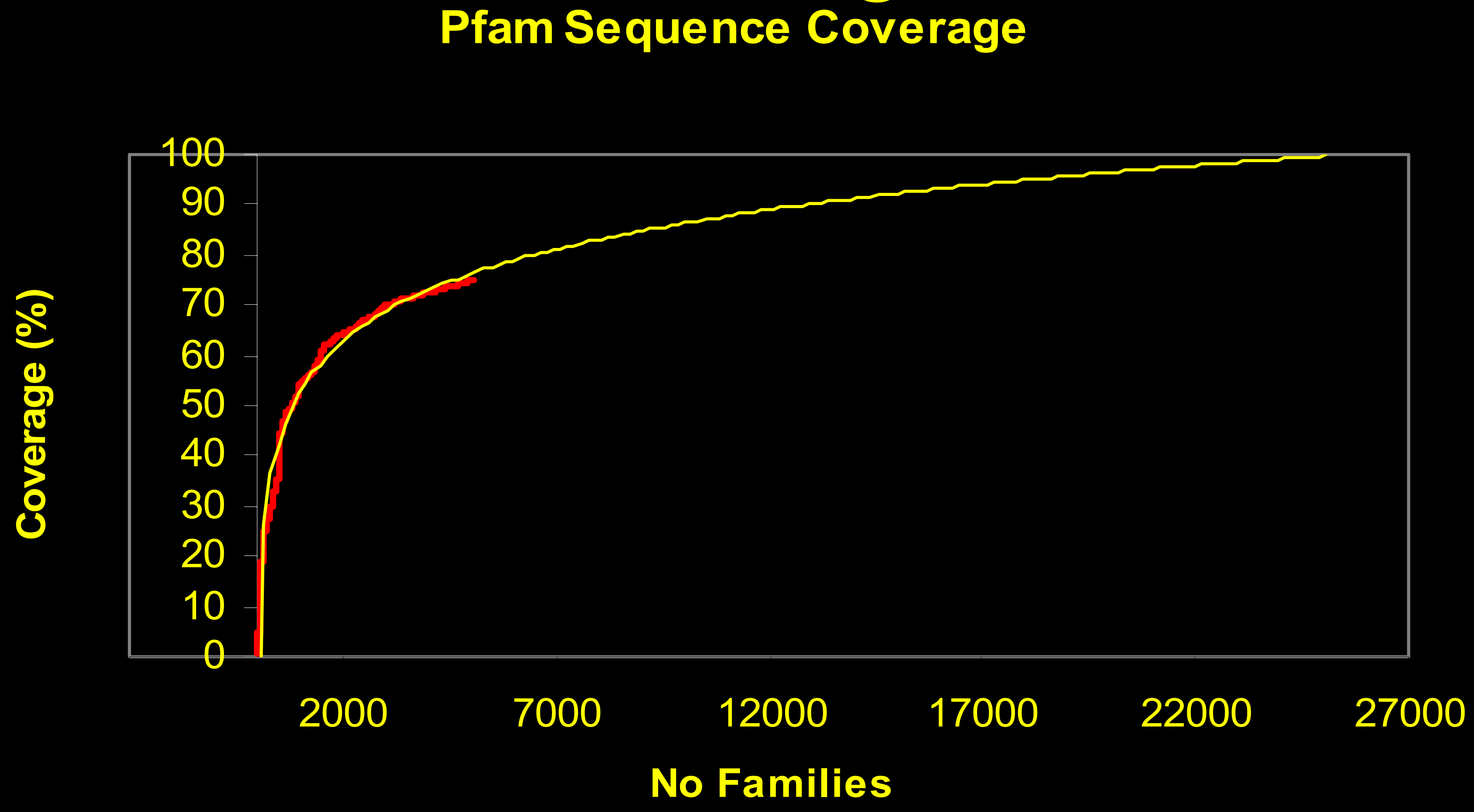
Pfam 18.0

- Pfam-A
  - 7,973 Curated families with annotation.
- Pfam-B
  - 100,000 families derived from Prodom.



http://www.sanger.ac.uk/Software/Pfam/

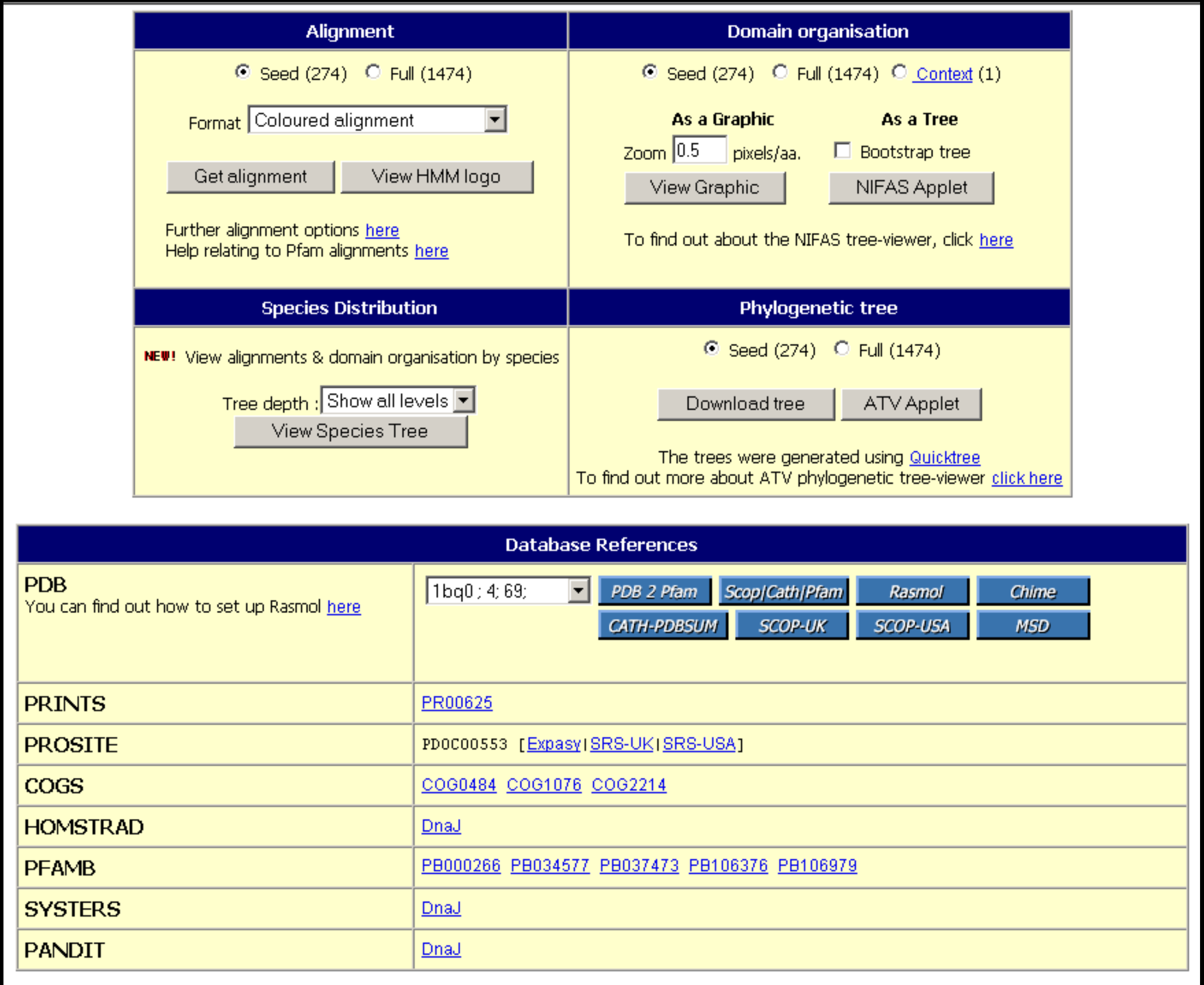
Coverage



- Retire sometime between Sept 2012 and May 2033!

http://www.sanger.ac.uk/Software/Pfam/

Family Pages



Alignment

● Seed (274) ○ Full (1474)

Format: Coloured alignment

Further alignment options here
Help relating to Pfam alignments here

Domain organisation

● Seed (274) ○ Full (1474) ○ Context (1)

As a Graphic

Zoom 0.5 pixels/aa.

As a Tree

□ Bootstrap tree

To find out about the NIFAS tree-viewer, click here

Species Distribution

NEW! View alignments & domain organisation by species

Tree depth : Show all levels

Phylogenetic tree

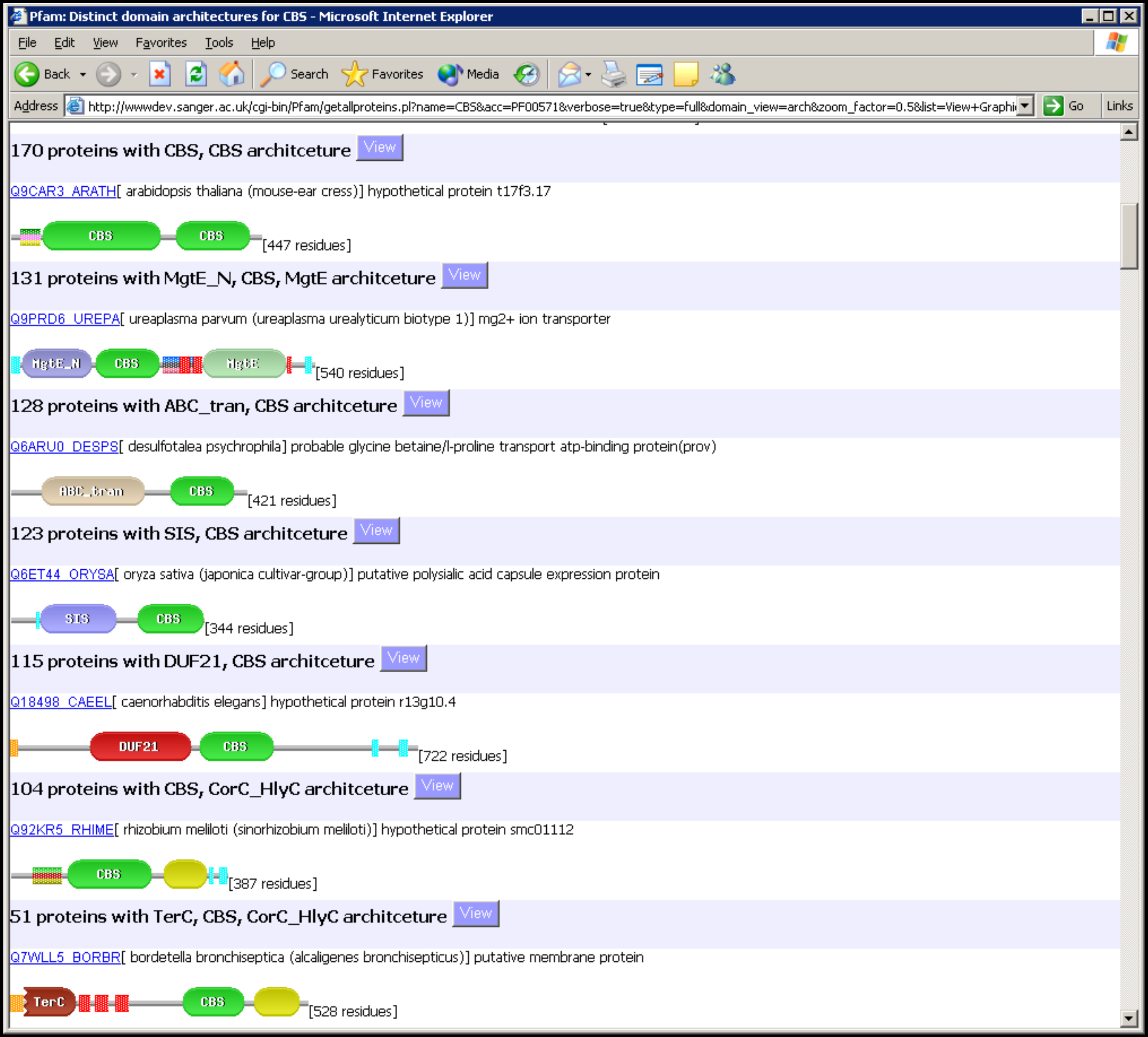
● Seed (274) ○ Full (1474)

The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here

Database References

PDB You can find out how to set up Rasmol here	1bq0 ; 4; 69;
PRINTS	PR00625
PROSITE	PDOC00553 [Expasy | SRS-UK | SRS-USA]
COGS	COG0484 COG1076 COG2214
HOMSTRAD	DnaJ
PFAMB	PB000266 PB034577 PB037473 PB106376 PB106979
SYSTERS	DnaJ
PANDIT	DnaJ

http://www.sanger.ac.uk/Software/Pfam/


Pfam: Distinct domain architectures for CBS - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
Go
Links

http://wwwdev.sanger.ac.uk/cgi-bin/Pfam/getallproteins.pl?name=CBS&acc=PF00571&verbose=true&type=full&domain_view=arch&zoom_factor=0.5&list=View+Graphi

170 proteins with CBS, CBS architceture View

Q9CAR3_ARATH[arabidopsis thaliana (mouse-ear cress)] hypothetical protein t17f3.17

[447 residues]

131 proteins with MgtE_N, CBS, MgtE architceture View

Q9PRD6_UREPA[ureaplasma parvum (ureaplasma urealyticum biotype 1)] mg2+ ion transporter

[540 residues]

128 proteins with ABC_tran, CBS architceture View

Q6ARU0_DESPS[desulfotalea psychrophila] probable glycine betaine/l-proline transport atp-binding protein(prov)

[421 residues]

123 proteins with SIS, CBS architceture View

Q6ET44_ORYSA[oryza sativa (japonica cultivar-group)] putative polysialic acid capsule expression protein

[344 residues]

115 proteins with DUF21, CBS architceture View

Q18498_CAEEL[caenorhabditis elegans] hypothetical protein r13g10.4

[722 residues]

104 proteins with CBS, CorC_HlyC architceture View

Q92KR5_RHIME[rhizobium meliloti (sinorhizobium meliloti)] hypothetical protein smc01112

[387 residues]

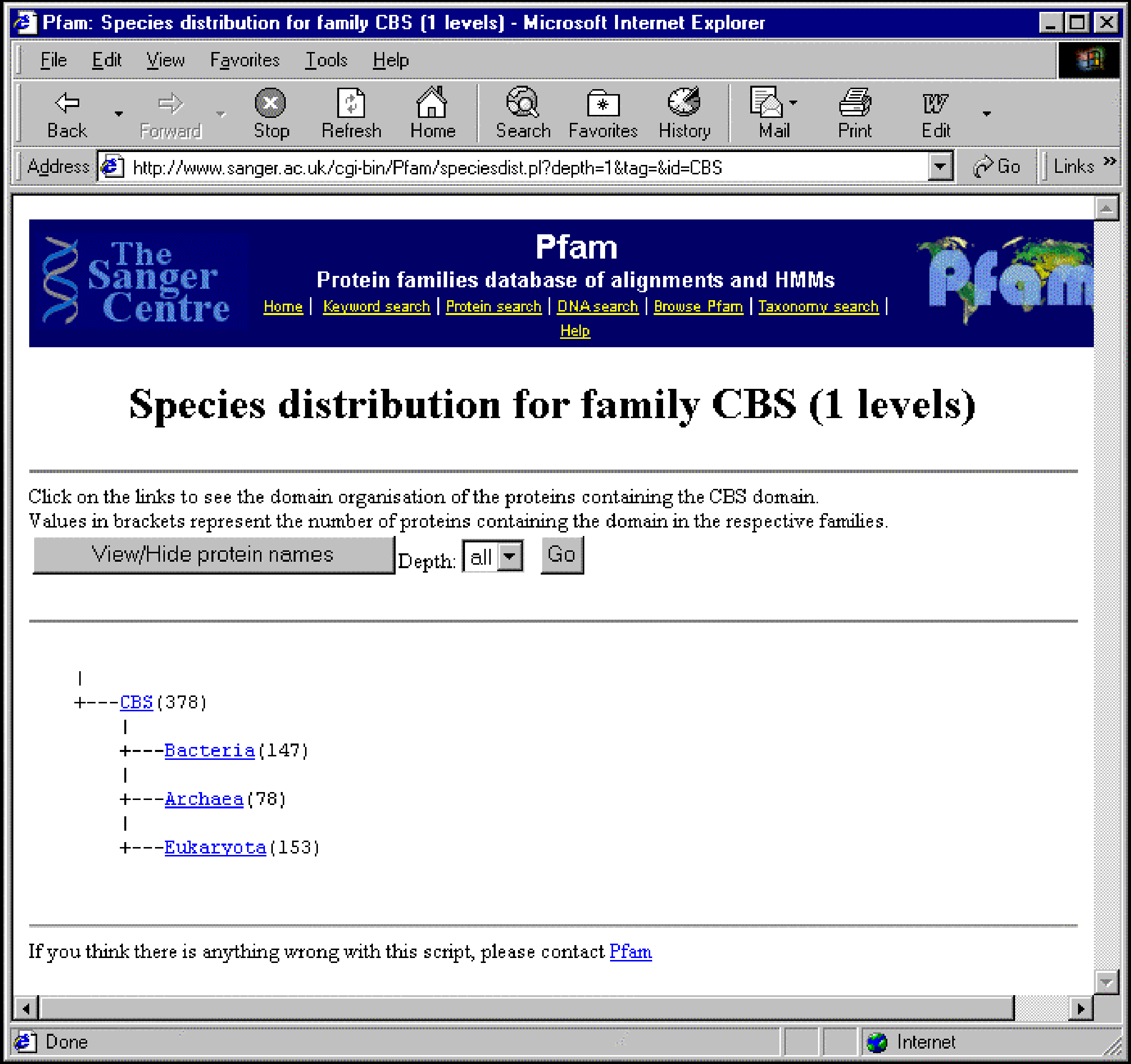
51 proteins with TerC, CBS, CorC_HlyC architceture View

Q7WLL5_BORBR[bordetella bronchiseptica (alcaligenes bronchisepticus)] putative membrane protein

[528 residues]

http://www.sanger.ac.uk/Software/Pfam/

- Taxonomy information
  - Does your favourite thermophile have a member?



Pfam: Species distribution for family CBS (1 levels) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Edit
Address
Go
Links
The Sanger Centre
Home | Keyword search | Protein search | DNA search | Browse Pfam | Taxonomy search | Help
Done
Internet

http://www.sanger.ac.uk/cgi-bin/Pfam/speciesdist.pl?depth=1&tag=&id=CBS

Pfam

Protein families database of alignments and HMMs

Species distribution for family CBS (1 levels)

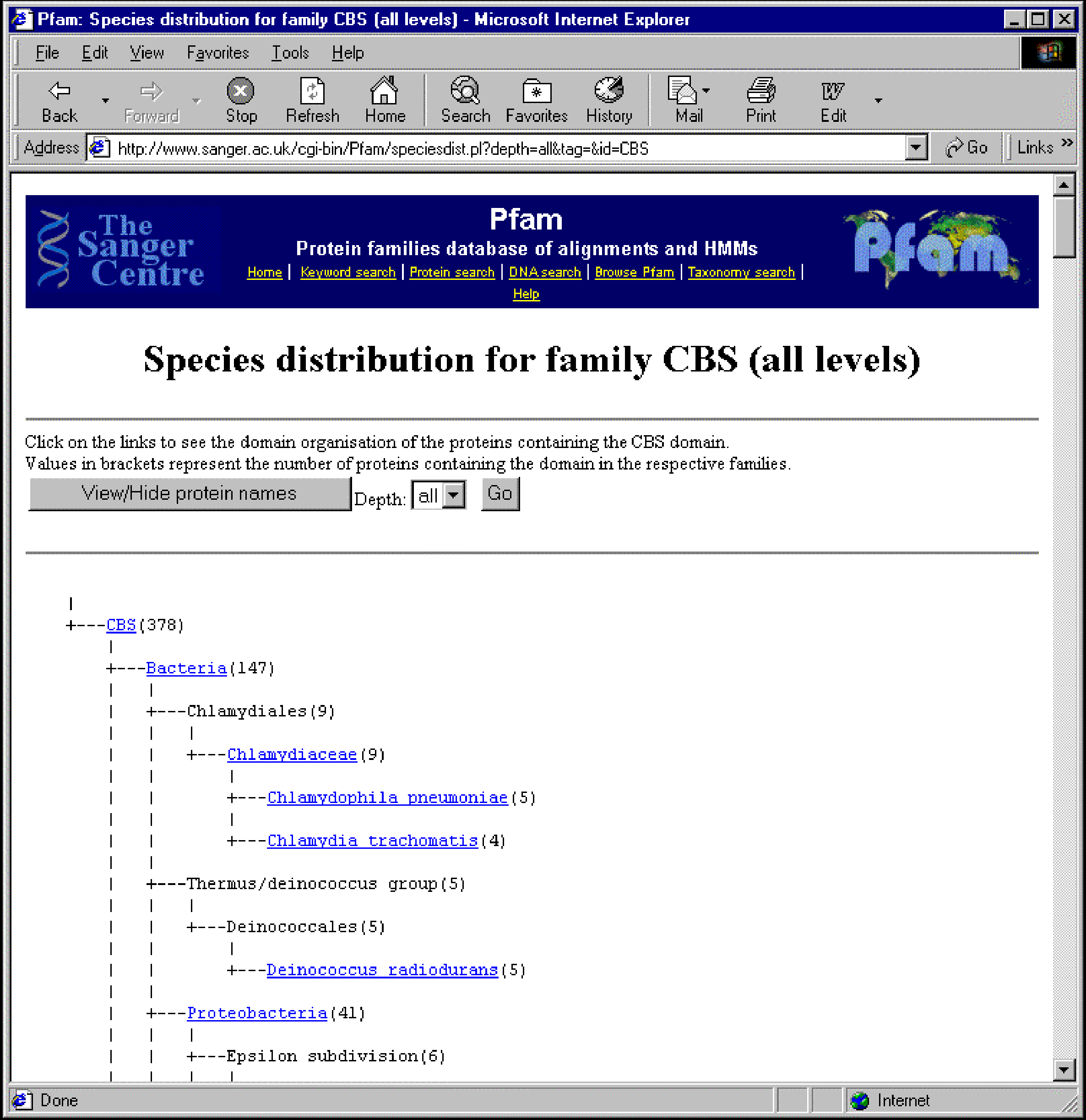
Click on the links to see the domain organisation of the proteins containing the CBS domain.
Values in brackets represent the number of proteins containing the domain in the respective families.

View/Hide protein names Depth: all Go

```
|
+---CBS(378)
     |
     +---Bacteria(147)
     |
     +---Archaea(78)
     |
     +---Eukaryota(153)
```

If you think there is anything wrong with this script, please contact Pfam

http://www.sanger.ac.uk/Software/Pfam/


Pfam: Species distribution for family CBS (all levels) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address http://www.sanger.ac.uk/cgi-bin/Pfam/speciesdist.pl?depth=all&tag=&id=CBS
Go Links
The Sanger Centre
Done
Internet

Pfam

Protein families database of alignments and HMMs

Home | Keyword search | Protein search | DNA search | Browse Pfam | Taxonomy search | Help

Species distribution for family CBS (all levels)

Click on the links to see the domain organisation of the proteins containing the CBS domain.
Values in brackets represent the number of proteins containing the domain in the respective families.

View/Hide protein names Depth: all Go

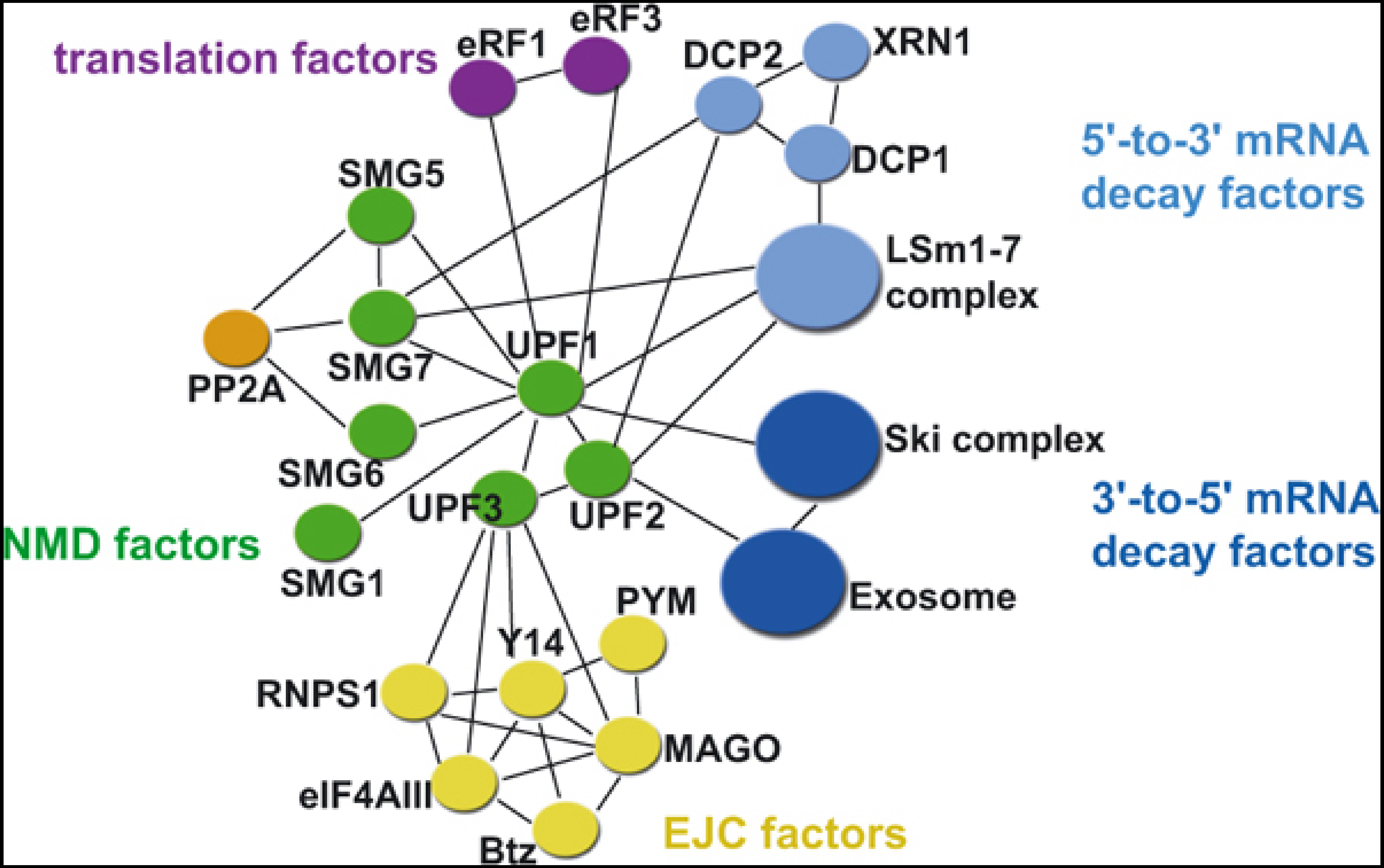
```
 |
 +---CBS(378)
       |
       +---Bacteria(147)
       |     |
       |     +---Chlamydiales(9)
       |     |     |
       |     |     +---Chlamydiaceae(9)
       |     |           |
       |     |           +---Chlamydophila pneumoniae(5)
       |     |           |
       |     |           +---Chlamydia trachomatis(4)
       |     |
       |     +---Thermus/deinococcus group(5)
       |     |     |
       |     |     +---Deinococcales(5)
       |     |           |
       |     |           +---Deinococcus radiodurans(5)
       |     |
       |     +---Proteobacteria(41)
       |     |     |
       |     |     +---Epsilon subdivision(6)
       |     |     |     |
```

http://www.sanger.ac.uk/Software/Pfam/

II - Summary

- Introduction to Pfam
- Protein Interactions
- Pfam Clans

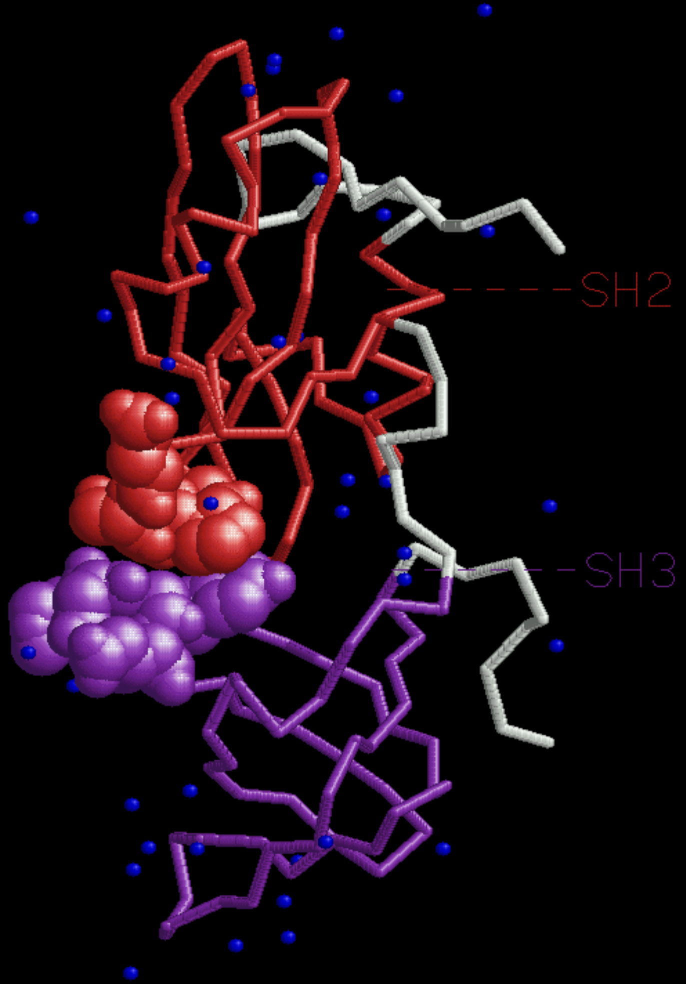
http://www.sanger.ac.uk/Software/Pfam/



http://www.sanger.ac.uk/Software/Pfam/

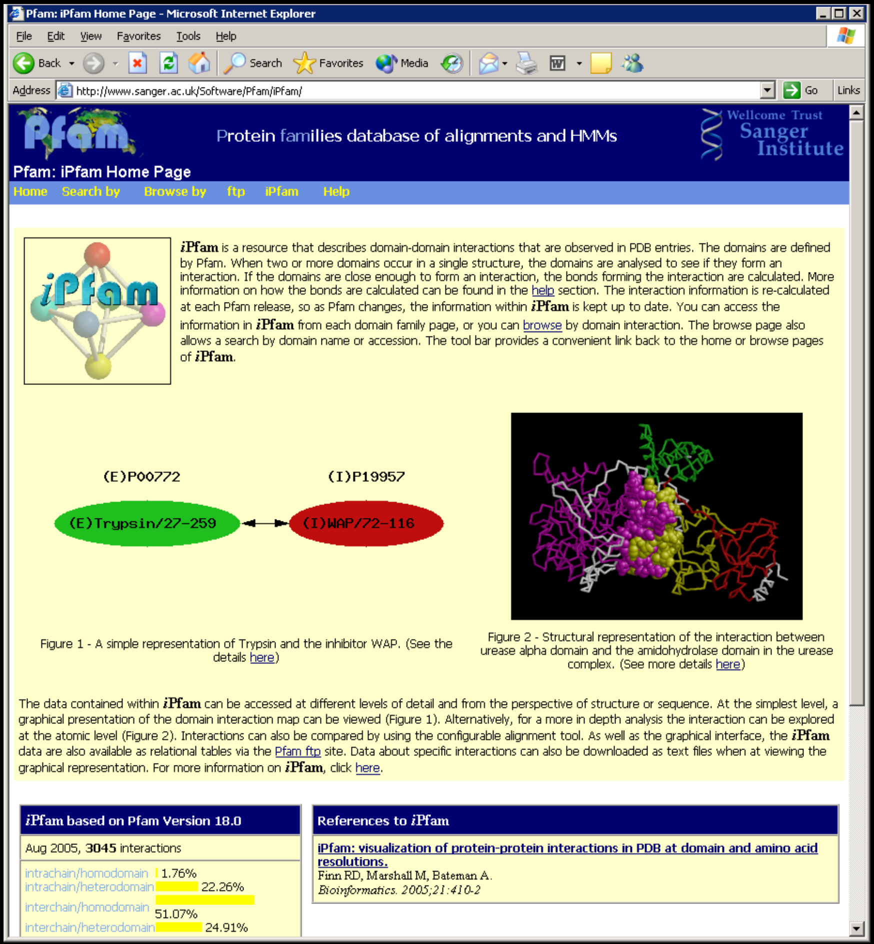
From Izaurralde group home page

Protein Interactions



http://www.sanger.ac.uk/Software/Pfam/

iPfam


Pfam: iPfam Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://www.sanger.ac.uk/Software/Pfam/iPfam/
Go
Links
Pfam
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Pfam: iPfam Home Page
Home
Search by
Browse by
ftp
iPfam
Help
iPfam
(E)P00772
(I)P19957
(E)Trypsin/27-259
(I)WAP/72-116

iPfam is a resource that describes domain-domain interactions that are observed in PDB entries. The domains are defined by Pfam. When two or more domains occur in a single structure, the domains are analysed to see if they form an interaction. If the domains are close enough to form an interaction, the bonds forming the interaction are calculated. More information on how the bonds are calculated can be found in the help section. The interaction information is re-calculated at each Pfam release, so as Pfam changes, the information within ***iPfam*** is kept up to date. You can access the information in ***iPfam*** from each domain family page, or you can browse by domain interaction. The browse page also allows a search by domain name or accession. The tool bar provides a convenient link back to the home or browse pages of ***iPfam***.

Figure 1 - A simple representation of Trypsin and the inhibitor WAP. (See the details here)

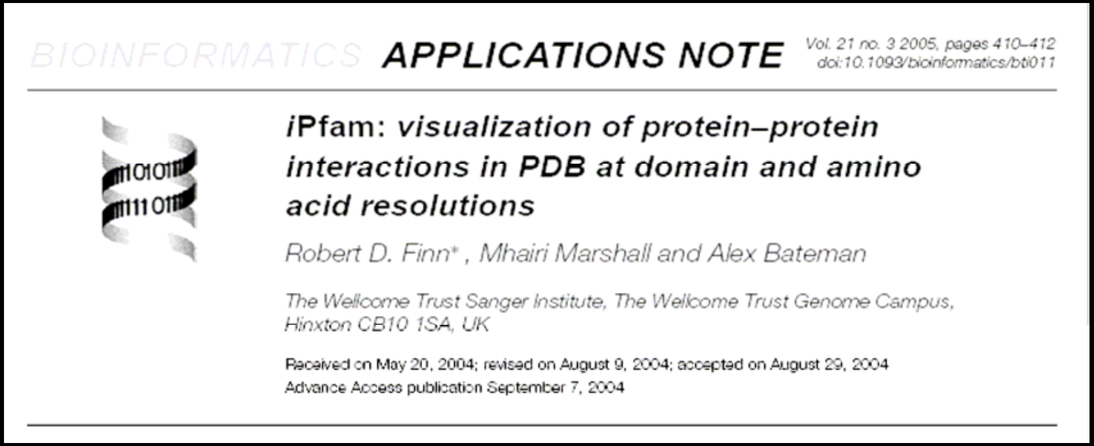
Figure 2 - Structural representation of the interaction between urease alpha domain and the amidohydrolase domain in the urease complex. (See more details here)

The data contained within ***iPfam*** can be accessed at different levels of detail and from the perspective of structure or sequence. At the simplest level, a graphical presentation of the domain interaction map can be viewed (Figure 1). Alternatively, for a more in depth analysis the interaction can be explored at the atomic level (Figure 2). Interactions can also be compared by using the configurable alignment tool. As well as the graphical interface, the ***iPfam*** data are also available as relational tables via the Pfam ftp site. Data about specific interactions can also be downloaded as text files when at viewing the graphical representation. For more information on ***iPfam***, click here.

*i*Pfam based on Pfam Version 18.0	
Aug 2005, **3045** interactions	
intrachain/homodomain	1.76%
intrachain/heterodomain	22.26%
interchain/homodomain	51.07%
interchain/heterodomain	24.91%

References to *i*Pfam

iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions.
Finn RD, Marshall M, Bateman A.
Bioinformatics. 2005;21:410-2



BIOINFORMATICS **APPLICATIONS NOTE**

Vol. 21 no. 3 2005, pages 410–412
doi:10.1093/bioinformatics/bti011

iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions

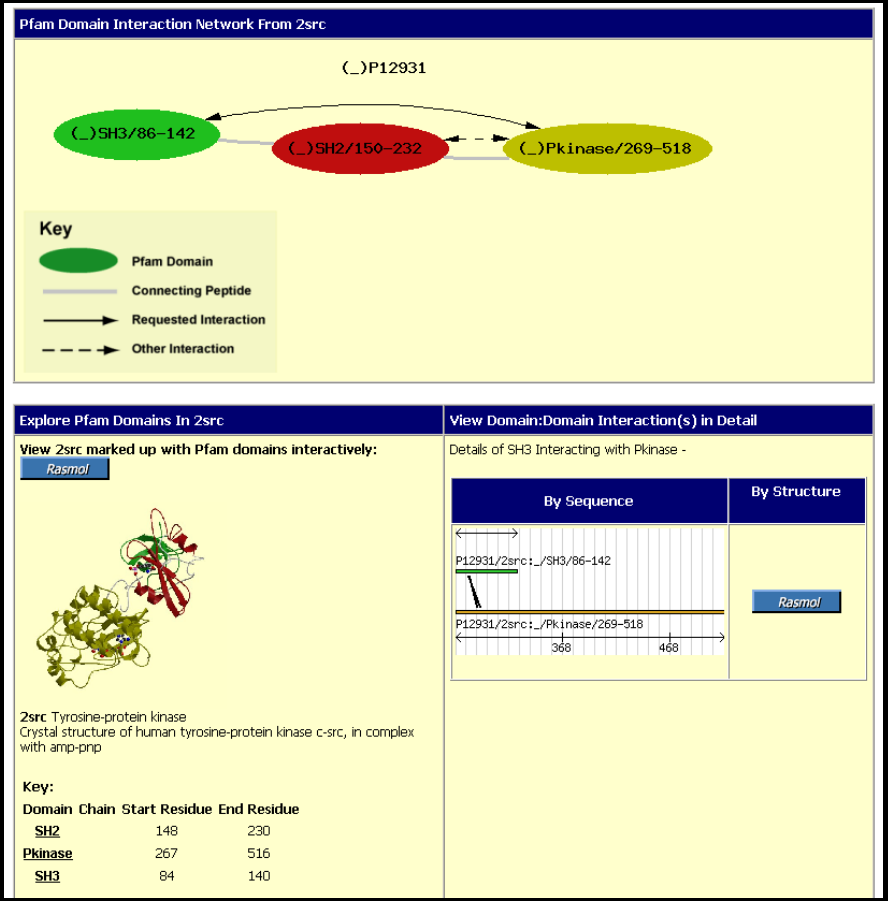
Robert D. Finn, Mhairi Marshall and Alex Bateman*

The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

Received on May 20, 2004; revised on August 9, 2004; accepted on August 29, 2004
Advance Access publication September 7, 2004

http://www.sanger.ac.uk/Software/Pfam/

Protein Interactions



Pfam Domain Interaction Network From 2src

Key

- 🟩 Pfam Domain
- ━ Connecting Peptide
- ⟶ Requested Interaction
- - - → Other Interaction

Explore Pfam Domains In 2src

View 2src marked up with Pfam domains interactively:

2src Tyrosine-protein kinase
Crystal structure of human tyrosine-protein kinase c-src, in complex with amp-pnp

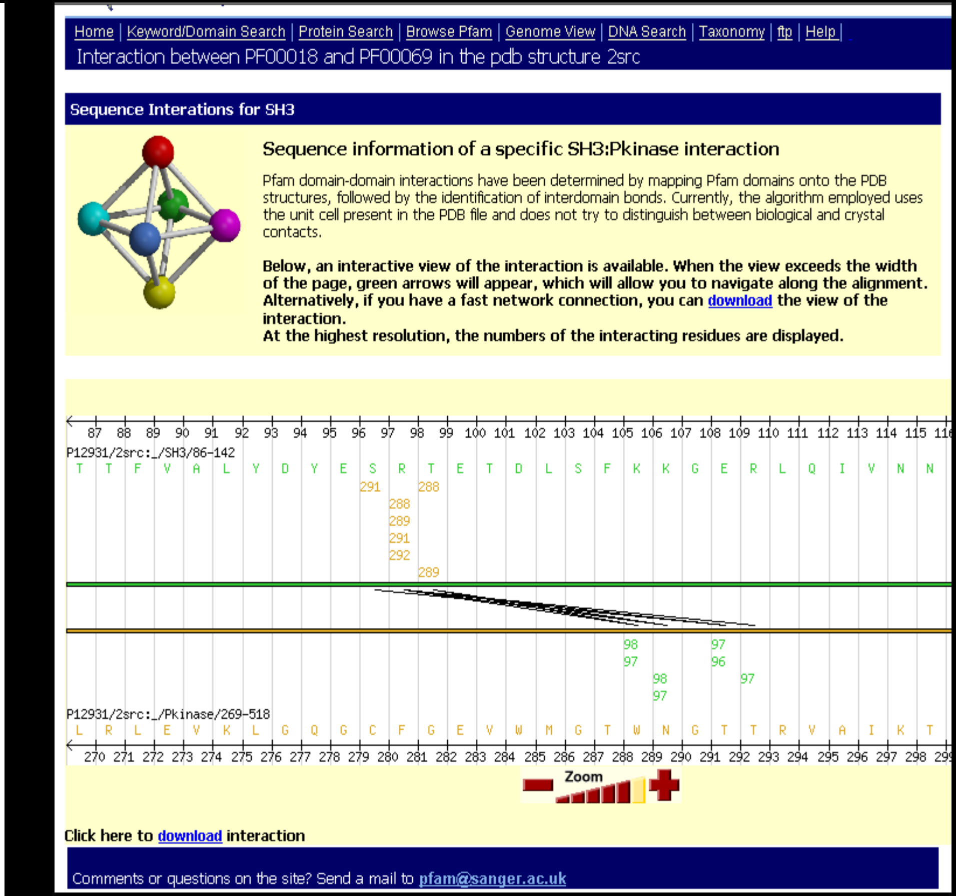
Key:

Domain	Chain	Start Residue	End Residue
SH2		148	230
Pkinase		267	516
SH3		84	140

View Domain:Domain Interaction(s) in Detail

Details of SH3 Interacting with Pkinase -

By Sequence	By Structure
P12931/2src:_/SH3/86-142 P12931/2src:_/Pkinase/269-518 368 468	Rasmol



Home | Keyword/Domain Search | Protein Search | Browse Pfam | Genome View | DNA Search | Taxonomy | ftp | Help |

Interaction between PF00018 and PF00069 in the pdb structure 2src

Sequence Interations for SH3

Sequence information of a specific SH3:Pkinase interaction

Pfam domain-domain interactions have been determined by mapping Pfam domains onto the PDB structures, followed by the identification of interdomain bonds. Currently, the algorithm employed uses the unit cell present in the PDB file and does not try to distinguish between biological and crystal contacts.

Below, an interactive view of the interaction is available. When the view exceeds the width of the page, green arrows will appear, which will allow you to navigate along the alignment. Alternatively, if you have a fast network connection, you can download the view of the interaction.
At the highest resolution, the numbers of the interacting residues are displayed.

Click here to download interaction

Comments or questions on the site? Send a mail to pfam@sanger.ac.uk

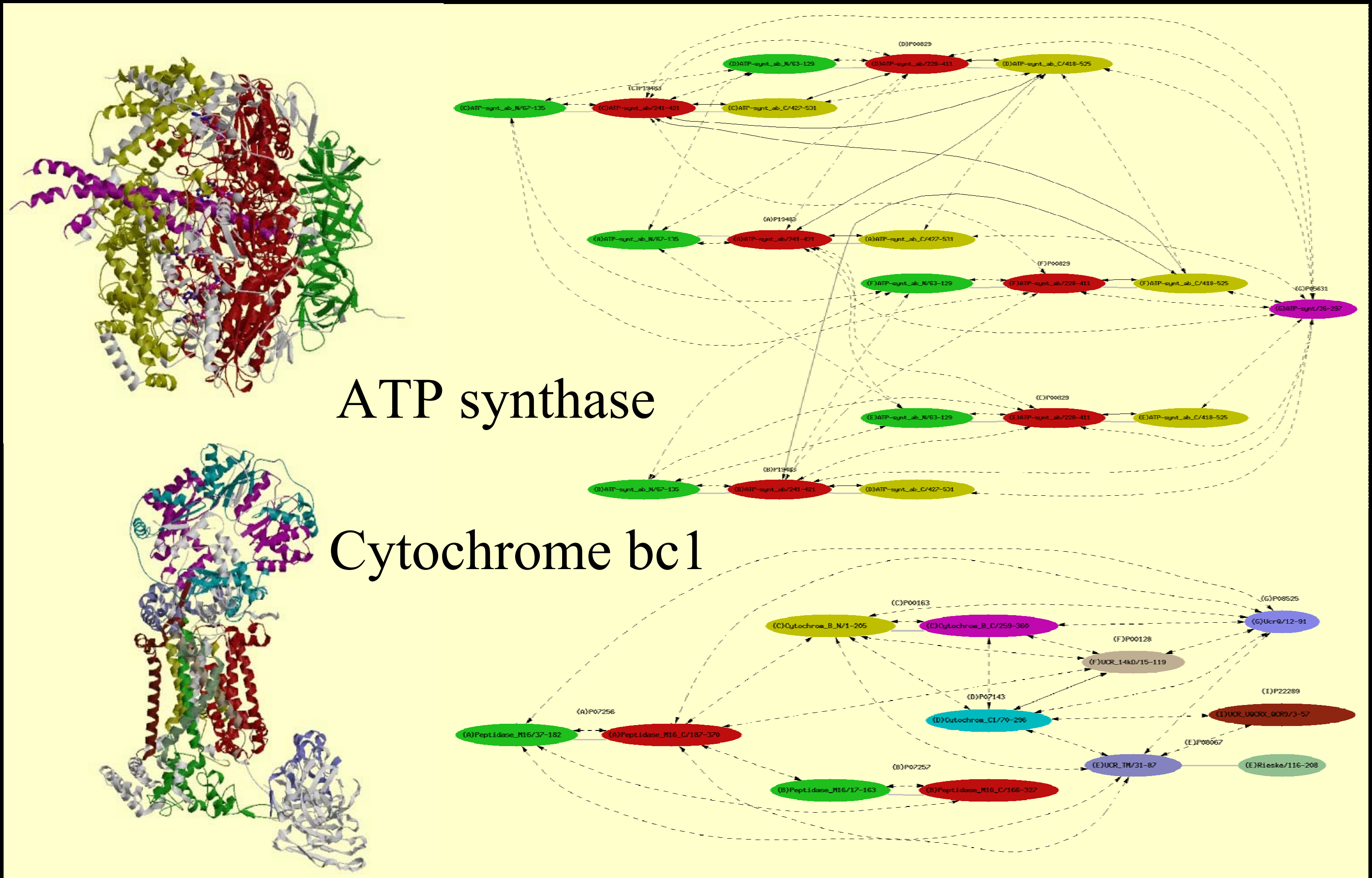
http://www.sanger.ac.uk/Software/Pfam/

The Pfam Supercomputer

A typical computer console.

http://www.sanger.ac.uk/Software/Pfam/

Complex Complexes



http://www.sanger.ac.uk/Software/Pfam/

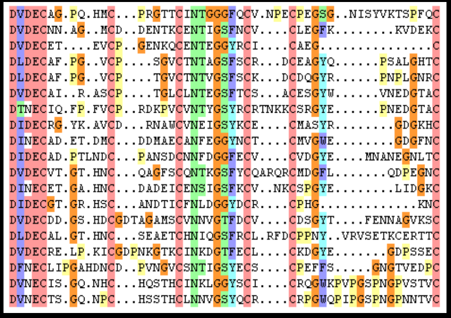
II - Summary

- Introduction to Pfam
- Protein Interactions
- Pfam Clans

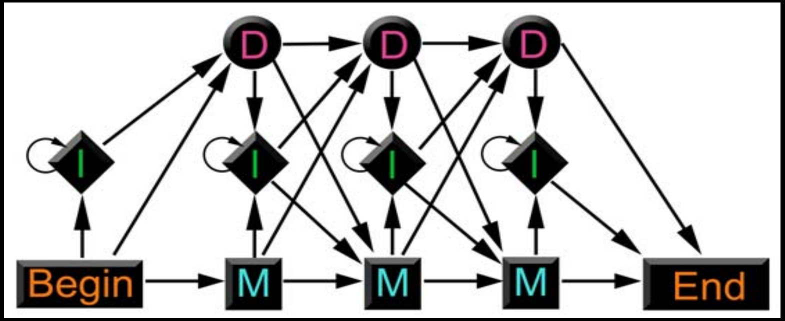
http://www.sanger.ac.uk/Software/Pfam/

What is Pfam?

- Database of Protein families
- Pfam defined by alignments & HMMs



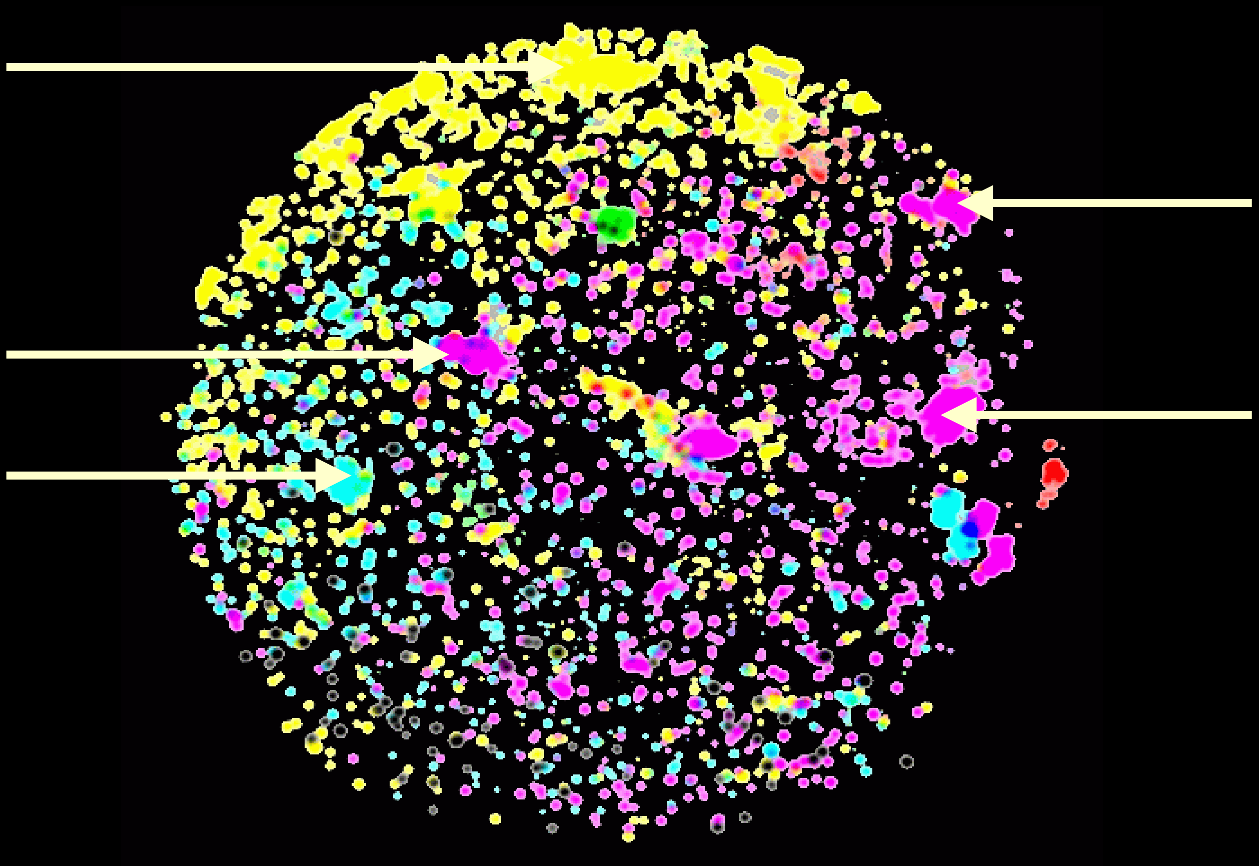
```
DVDECAG.PQ.HMC...PRGTTCINTGGGFQCV.NPECPEGSG..NISYVKTSPFQC
DVDECNN.AG..MCD..DENTKCENTIGSFNCV....CLEGFK........KVDEKC
DVDECET....EVCP..GENKQCENTEGGYRCI....CAEG..............C
DLDECAF.PG..VCP....SGVCTNTAGSFSCR...DCEAGYQ......PSALGHTC
DLDECAF.PG..VCP....TGVCTNTVGSFSCK...DCDQGYR......PNPLGNRC
DVDECAI..R.ASCP....TGLCLNTEGSFTCS...ACESGYW......VNEDGTAC
DTNECIQ.FP.FVCP..RDKPVCVNTYGSYRCRTNKKCSRGYE......PNEDGTAC
DIDECRG.YK.AVCD...RNAWCVNEIGSYKCE....CMASYR........GDGKHC
DINECAD.ET.DMC...DDMAECANFEGGYNCT....CMVGWE........GDGFNC
DIDECAD.PTLNDC...PANSDCNNFDGGFECV....CVDGYE....MNANEGNLTC
DVDECVT.GT.HNC...QAGFSCQNTKGSFYCQARQRCMDGFL.......QDPEGNC
DINECET.GA.HNC...DADEICENSIGSFKCV..NKCSPGYE........LIDGKC
DIDECGT.GR.HSC...ANDTICFNLDGGYDCR....CPHG............KNC
DVDECDD.GS.HDCGDTAGAMSCVNNVGTFDCV....CDSGYT....FENNAGVKSC
DLDECAL.GT.HNC...SEAETCHNIQGSFRCL.RFDCPPNY..VRVSETKCERTTC
DVDECRE.LP.KICGDPNKGTKCINKDGTFECL....CKDGYE.......GDPSSEC
DFNECLIPGAHDNCD..PVNGVCSNTIGSYECS....CPEFFS.....GNGTVEDPC
DVNECIS.GQ.NHC...HQSTHCINKLGGYSCI....CRQGWKPVPGSPNGPVSTVC
DVNECTS.GQ.NPC...HSSTHCLNNVGSYQCR....CRPGWQPIPGSPNGPNNTVC
```



- Flat classification
- Many families are related

http://www.sanger.ac.uk/Software/Pfam/

Protein Space



Rossmann

Apartic Protease

Serine Protease

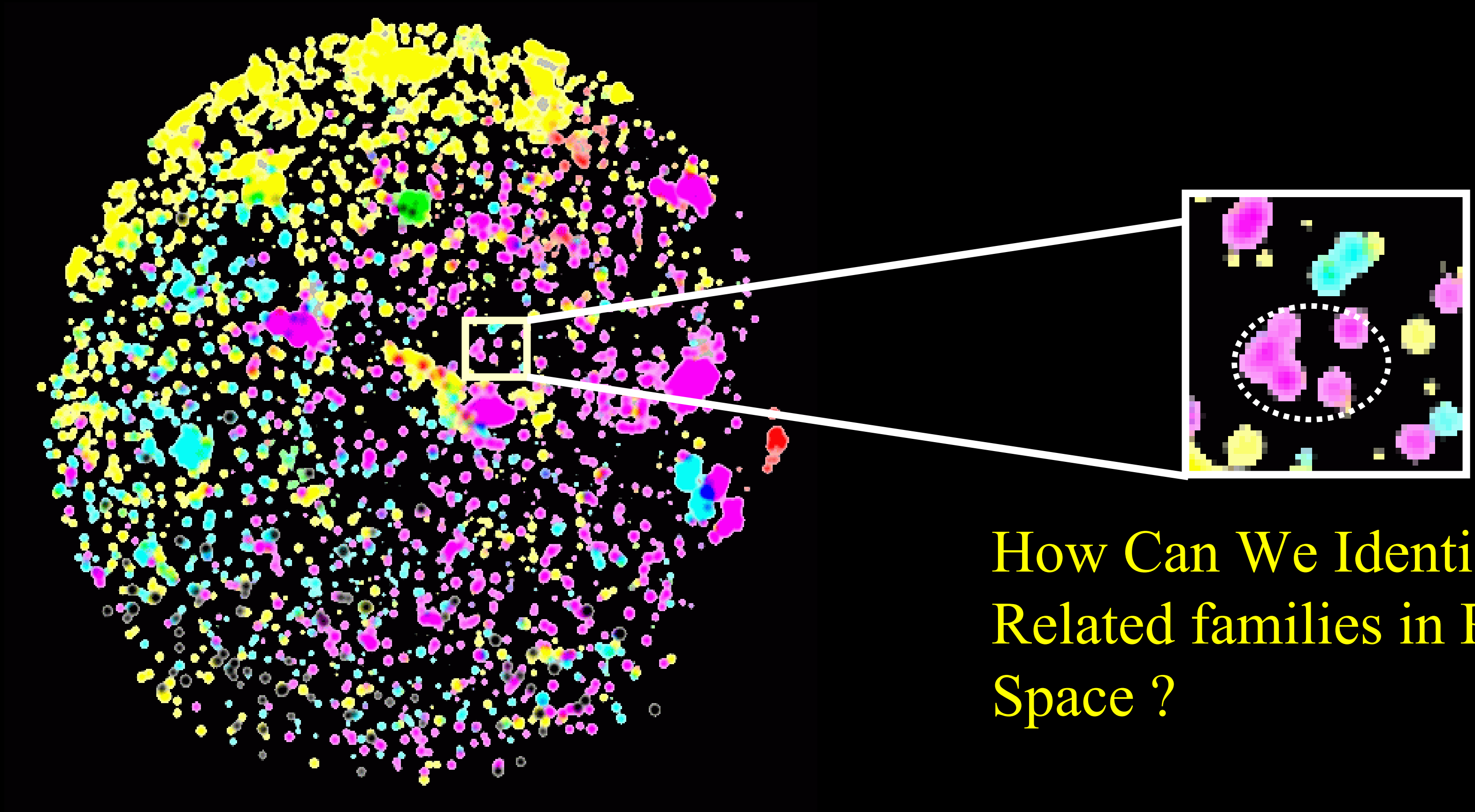
Immuno-globulins

Globins

http://www.zbh.uni-hamburg.de/wurst/protspace

http://www.sanger.ac.uk/Software/Pfam/

Protein Space



How Can We Identify Related families in Proteins Space ?

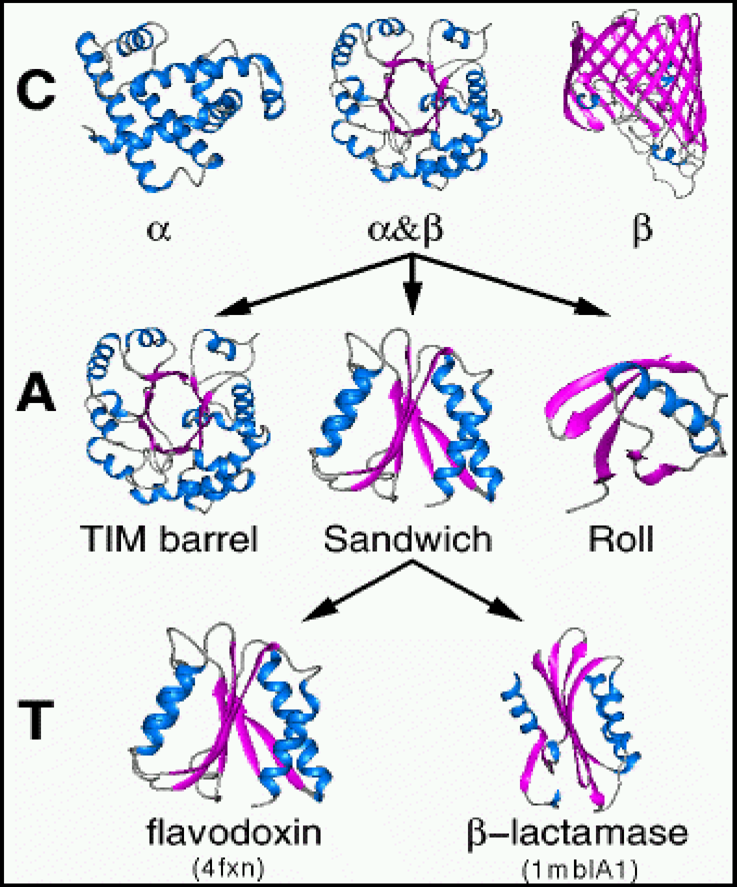
http://www.sanger.ac.uk/Software/Pfam/

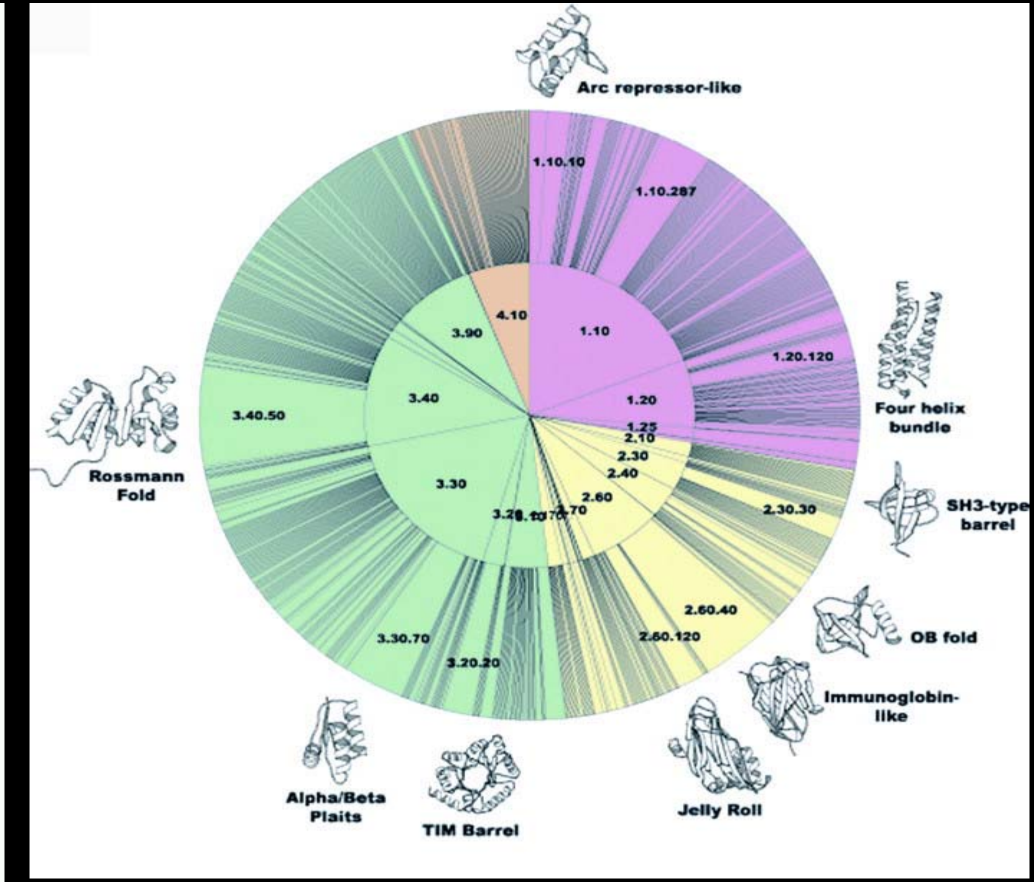
Pfam Clans

- Group together families
  - Structure databases
  - PRC
  - Overlaps
  - Literature

http://www.sanger.ac.uk/Software/Pfam/

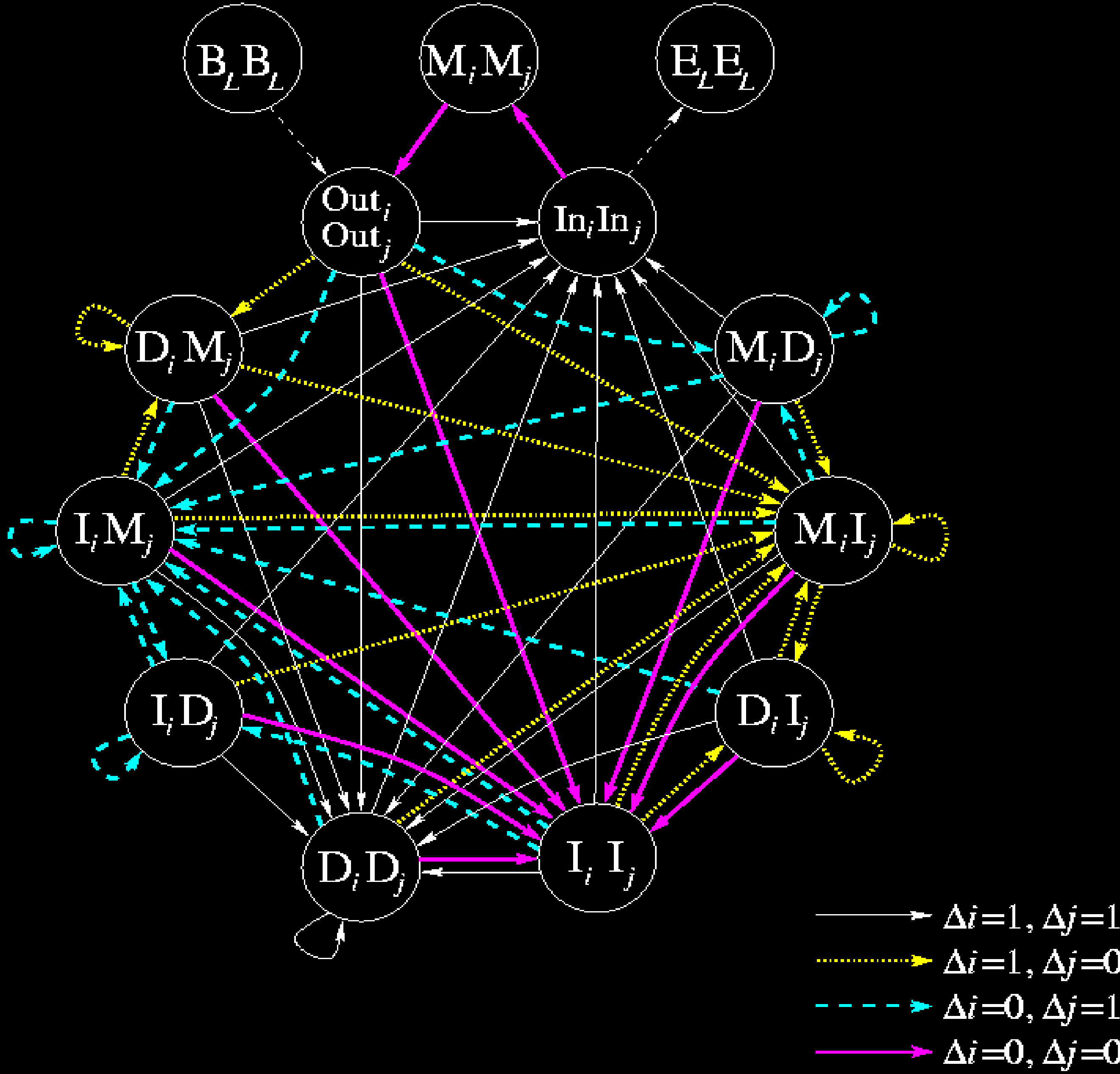
Structural Databases



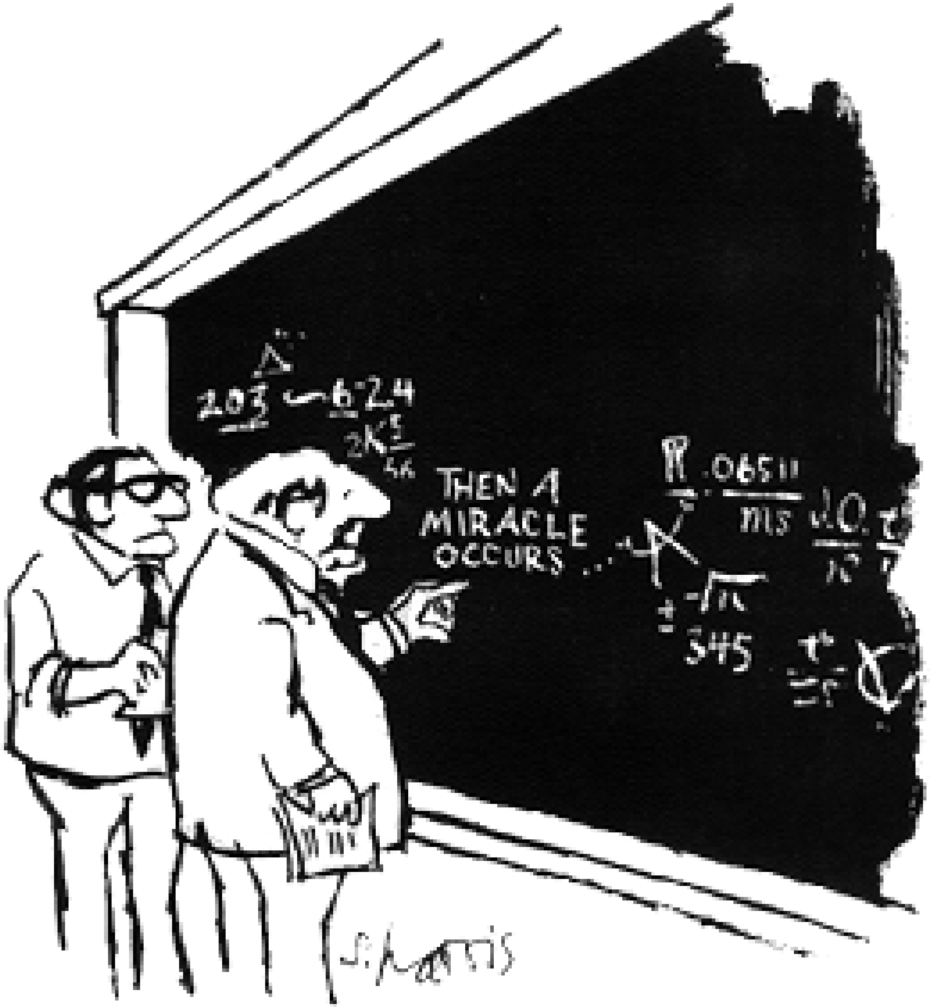


http://www.sanger.ac.uk/Software/Pfam/

Profile HMM Comparison

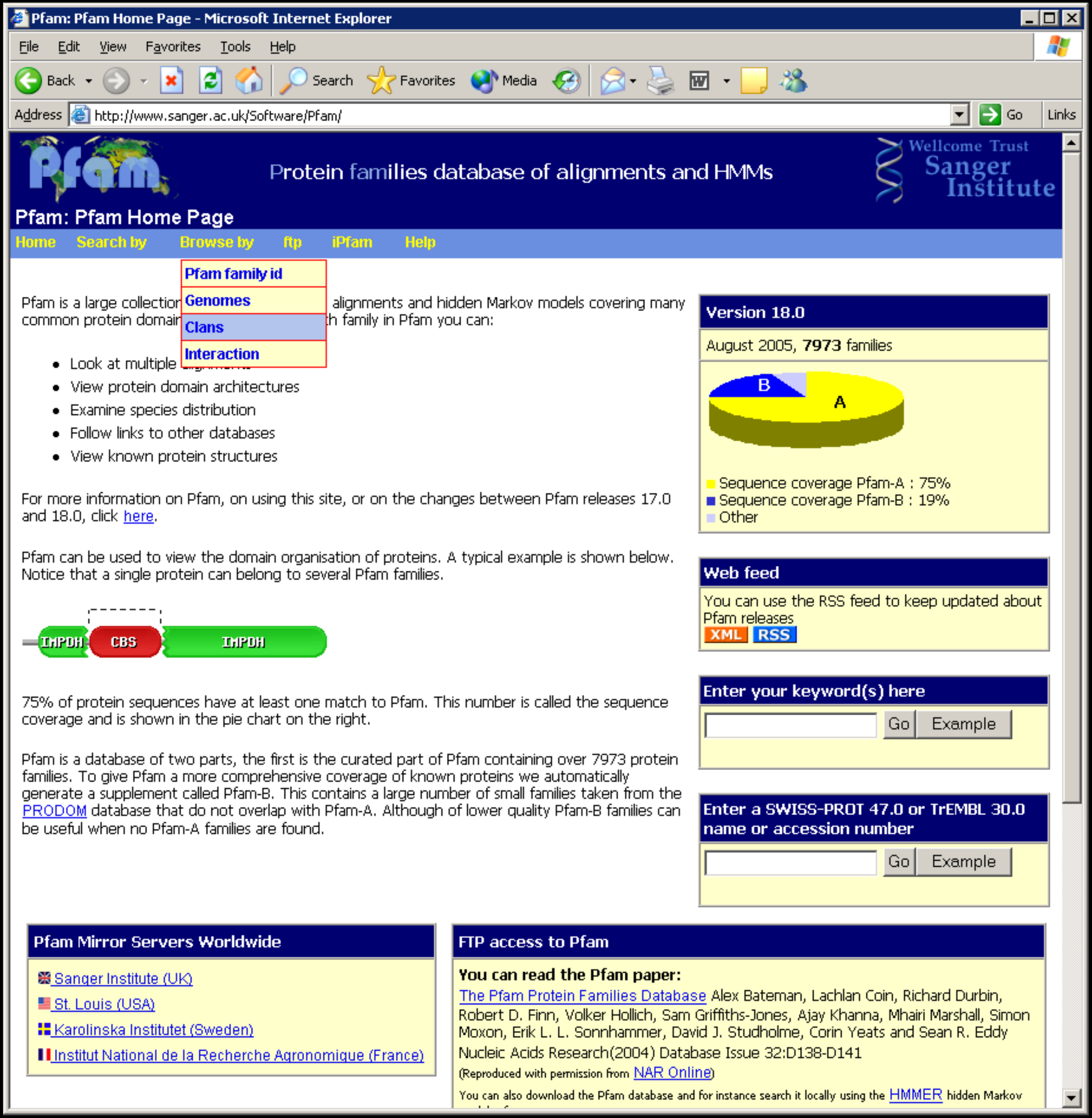


http://www.sanger.ac.uk/Software/Pfam/



"I think you should be more explicit here in step two."

http://www.sanger.ac.uk/Software/Pfam/


Pfam: Pfam Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://www.sanger.ac.uk/Software/Pfam/
Go
Links
Pfam
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Home
Search by
Browse by
ftp
iPfam
Help
Pfam family id
Genomes
Clans
Interaction
B
A
IMPDH
CBS
IMPDH
XML
RSS
Go
Example
Go
Example

Pfam: Pfam Home Page

Pfam is a large collectior alignments and hidden Markov models covering many common protein domair h family in Pfam you can:

- Look at multiple
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or on the changes between Pfam releases 17.0 and 18.0, click here.

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.

75% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 7973 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

Version 18.0

August 2005, **7973** families

- 🟨 Sequence coverage Pfam-A : 75%
- 🟦 Sequence coverage Pfam-B : 19%
- ▫ Other

Web feed

You can use the RSS feed to keep updated about Pfam releases

Enter your keyword(s) here

Enter a SWISS-PROT 47.0 or TrEMBL 30.0 name or accession number

Pfam Mirror Servers Worldwide

- Sanger Institute (UK)
- St. Louis (USA)
- Karolinska Institutet (Sweden)
- Institut National de la Recherche Agronomique (France)

FTP access to Pfam

You can read the Pfam paper:

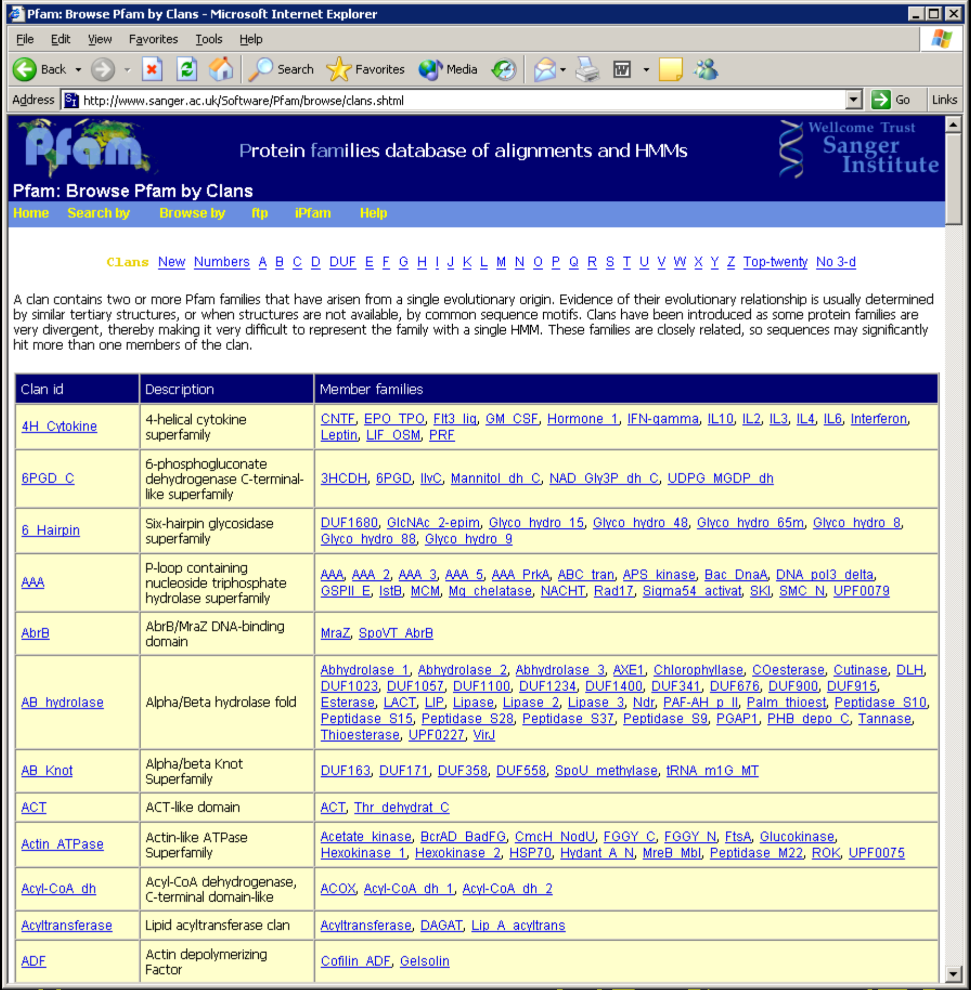
The Pfam Protein Families Database Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats and Sean R. Eddy Nucleic Acids Research(2004) Database Issue 32:D138-D141

(Reproduced with permission from NAR Online)

You can also download the Pfam database and for instance search it locally using the HMMER hidden Markov

http://www.sanger.ac.uk/Software/Pfam/

Browse clans


Pfam: Browse Pfam by Clans - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
Links
Pfam
Wellcome Trust Sanger Institute
Home Search by Browse by ftp iPfam Help
Clans New Numbers A B C D DUF E F G H I J K L M N O P Q R S T U V W X Y Z Top-twenty No 3-d

http://www.sanger.ac.uk/Software/Pfam/browse/clans.shtml

Protein families database of alignments and HMMs

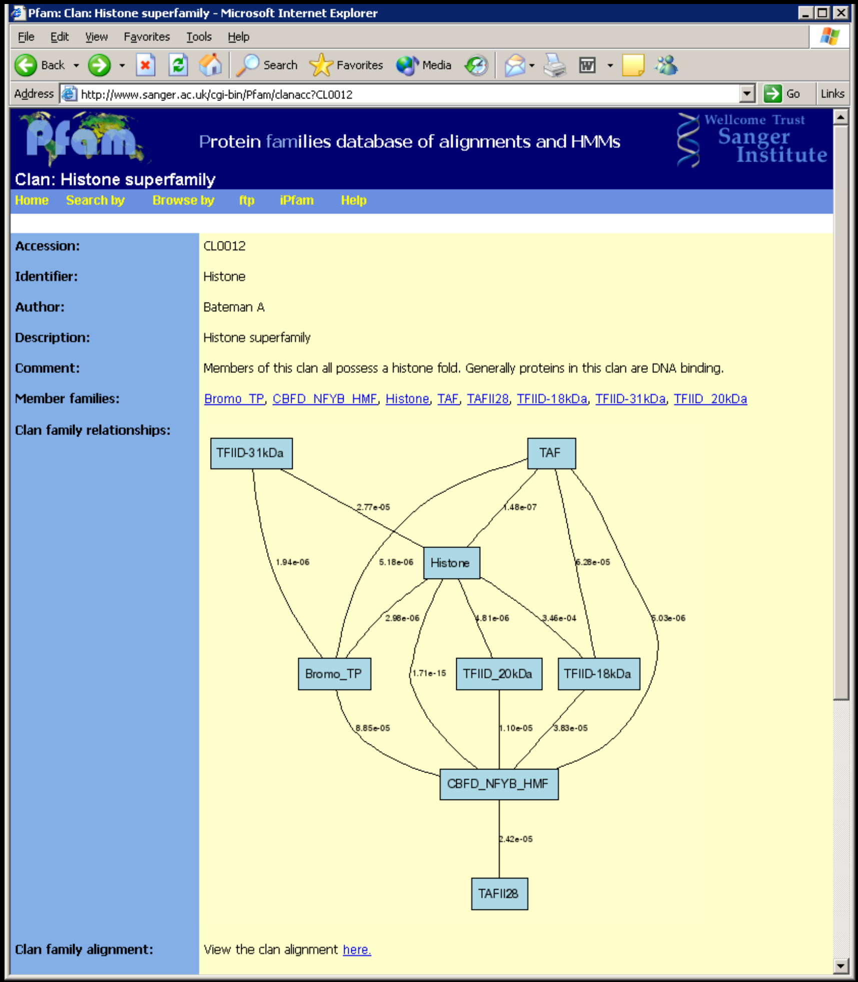
Pfam: Browse Pfam by Clans

A clan contains two or more Pfam families that have arisen from a single evolutionary origin. Evidence of their evolutionary relationship is usually determined by similar tertiary structures, or when structures are not available, by common sequence motifs. Clans have been introduced as some protein families are very divergent, thereby making it very difficult to represent the family with a single HMM. These families are closely related, so sequences may significantly hit more than one members of the clan.

Clan id	Description	Member families
4H_Cytokine	4-helical cytokine superfamily	CNTF, EPO_TPO, Flt3_lig, GM_CSF, Hormone_1, IFN-gamma, IL10, IL2, IL3, IL4, IL6, Interferon, Leptin, LIF_OSM, PRF
6PGD_C	6-phosphogluconate dehydrogenase C-terminal-like superfamily	3HCDH, 6PGD, IlvC, Mannitol_dh_C, NAD_Gly3P_dh_C, UDPG_MGDP_dh
6_Hairpin	Six-hairpin glycosidase superfamily	DUF1680, GlcNAc_2-epim, Glyco_hydro_15, Glyco_hydro_48, Glyco_hydro_65m, Glyco_hydro_8, Glyco_hydro_88, Glyco_hydro_9
AAA	P-loop containing nucleoside triphosphate hydrolase superfamily	AAA, AAA_2, AAA_3, AAA_5, AAA_PrkA, ABC_tran, APS_kinase, Bac_DnaA, DNA_pol3_delta, GSPII_E, IstB, MCM, Mg_chelatase, NACHT, Rad17, Sigma54_activat, SKI, SMC_N, UPF0079
AbrB	AbrB/MraZ DNA-binding domain	MraZ, SpoVT_AbrB
AB_hydrolase	Alpha/Beta hydrolase fold	Abhydrolase_1, Abhydrolase_2, Abhydrolase_3, AXE1, Chlorophyllase, COesterase, Cutinase, DLH, DUF1023, DUF1057, DUF1100, DUF1234, DUF1400, DUF341, DUF676, DUF900, DUF915, Esterase, LACT, LIP, Lipase, Lipase_2, Lipase_3, Ndr, PAF-AH_p_II, Palm_thioest, Peptidase_S10, Peptidase_S15, Peptidase_S28, Peptidase_S37, Peptidase_S9, PGAP1, PHB_depo_C, Tannase, Thioesterase, UPF0227, VirJ
AB_Knot	Alpha/beta Knot Superfamily	DUF163, DUF171, DUF358, DUF558, SpoU_methylase, tRNA_m1G_MT
ACT	ACT-like domain	ACT, Thr_dehydrat_C
Actin_ATPase	Actin-like ATPase Superfamily	Acetate_kinase, BcrAD_BadFG, CmcH_NodU, FGGY_C, FGGY_N, FtsA, Glucokinase, Hexokinase_1, Hexokinase_2, HSP70, Hydant_A_N, MreB_Mbl, Peptidase_M22, ROK, UPF0075
Acyl-CoA_dh	Acyl-CoA dehydrogenase, C-terminal domain-like	ACOX, Acyl-CoA_dh_1, Acyl-CoA_dh_2
Acyltransferase	Lipid acyltransferase clan	Acyltransferase, DAGAT, Lip_A_acyltrans
ADF	Actin depolymerizing Factor	Cofilin_ADF, Gelsolin

http://www.sanger.ac.uk/Software/Pfam/

Clan relationships



Pfam: Clan: Histone superfamily - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://www.sanger.ac.uk/cgi-bin/Pfam/clanacc?CL0012
Go
Links
Pfam
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Home
Search by
Browse by
ftp
iPfam
Help

Clan: Histone superfamily

Accession: CL0012

Identifier: Histone

Author: Bateman A

Description: Histone superfamily

Comment: Members of this clan all possess a histone fold. Generally proteins in this clan are DNA binding.

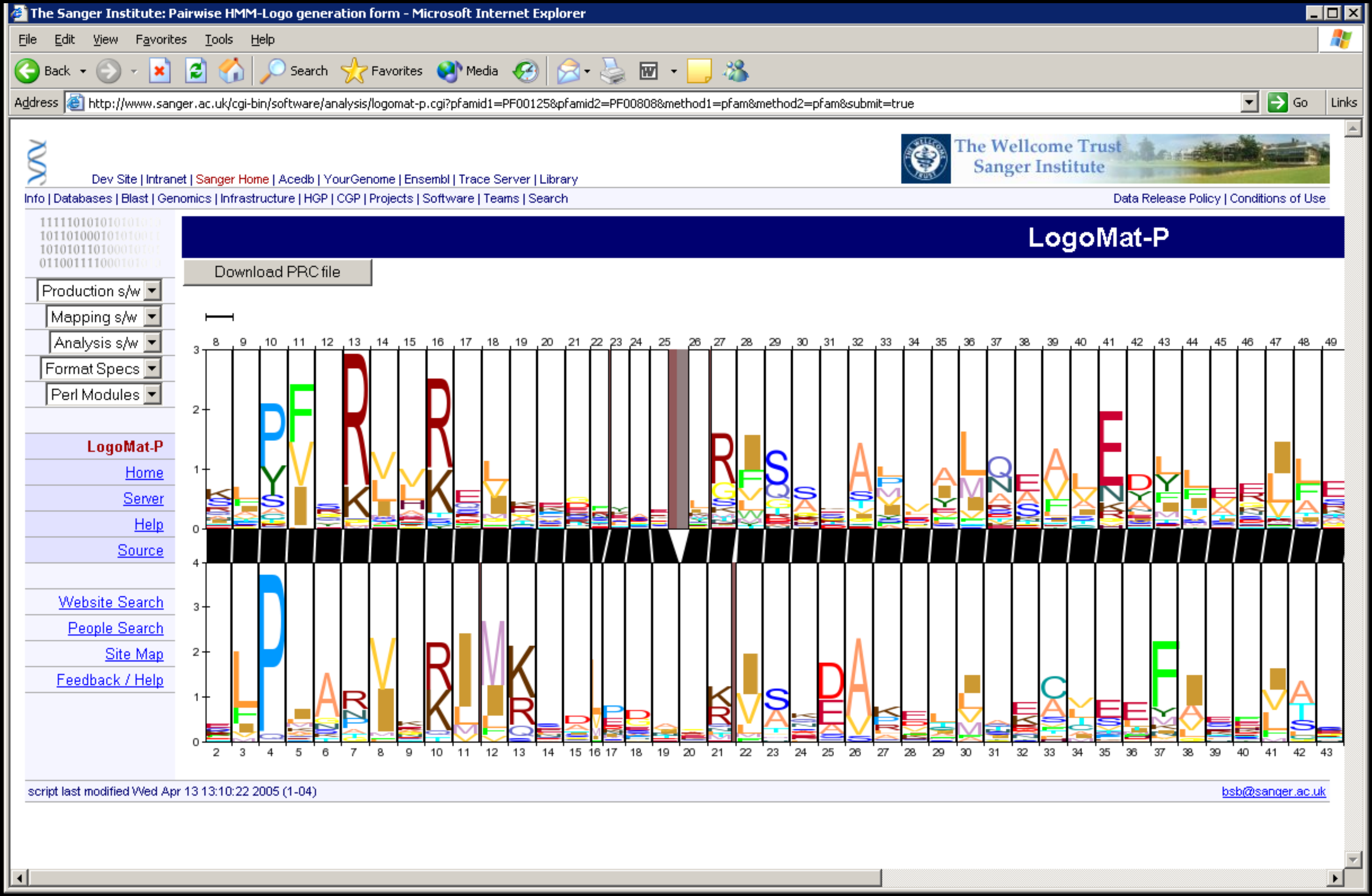
Member families: Bromo_TP, CBFD_NFYB_HMF, Histone, TAF, TAFII28, TFIID-18kDa, TFIID-31kDa, TFIID_20kDa

Clan family relationships:

Clan family alignment: View the clan alignment here.

http://www.sanger.ac.uk/Software/Pfam/

HMM comparisons



The Sanger Institute: Pairwise HMM-Logo generation form - Microsoft Internet Explorer

Address http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-p.cgi?pfamid1=PF00125&pfamid2=PF00808&method1=pfam&method2=pfam&submit=true

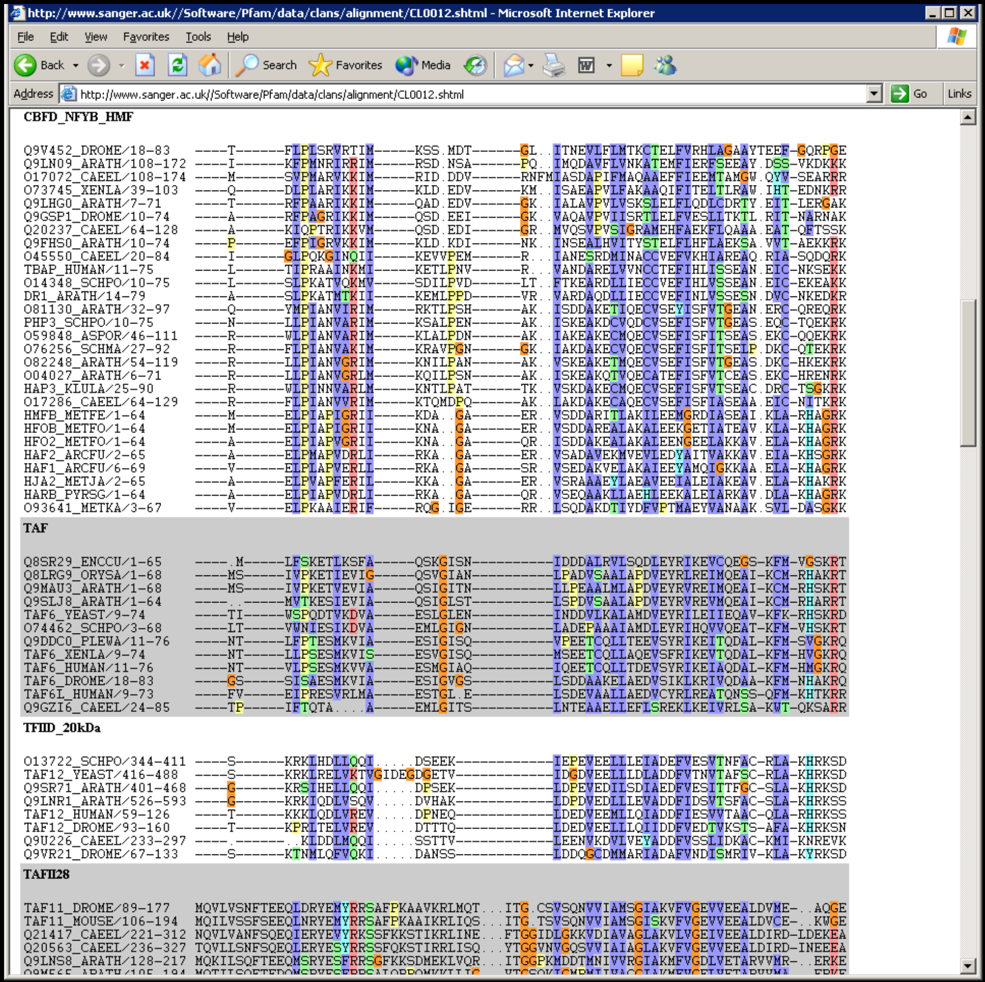
LogoMat-P

script last modified Wed Apr 13 13:10:22 2005 (1-04)

bsb@sanger.ac.uk

http://www.sanger.ac.uk/Software/Pfam/

Clan alignments


http://www.sanger.ac.uk//Software/Pfam/data/clans/alignment/CL0012.shtml - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
Address
http://www.sanger.ac.uk//Software/Pfam/data/clans/alignment/CL0012.shtml
Go
Links

CBFD_NFYB_HMF

```
Q9V452_DROME/18-83   ----T------FLPLSRVRTIM-----KSS.MDT------GL..ITNEVLFLMTKCTELFVRHLAGAAYTEEF-GQRPGE
Q9LN09_ARATH/108-172 ----I------KFPMNRIRRIM-----RSD.NSA------PQ..IMQDAVFLVNKATEMFIERFSEEAY.DSS-VKDKKK
O17072_CAEEL/108-174 ----M------SVPMARVKKIM-----RID.DDV------RNFMIASDAPIFMAQAAEFFIEEMTAMGW.QYV-SEARRR
O73745_XENLA/39-103  ----Q------DLPLARIKKIM-----KLD.EDV------KM..ISAEAPVLFAKAAQIFITELTLRAW.IHT-EDNKRR
Q9LHG0_ARATH/7-71    ----T------RFPAARIKKIM-----QAD.EDV------GK..IALAVPVLVSKSLELFLQDLCDRTY.EIT-LERGAK
Q9GSP1_DROME/10-74   ----A------RFPAGRIKKIM-----QSD.EEI------GK..VAQAVPVIISRTLELFVESLLTKTL.RIT-NARNAK
Q20237_CAEEL/64-128  ----A------KIQPTRIKKVM-----QSD.EDI------GR..MVQSVPVSIGRAMEHFAEKFLQAAA.EAT-QFTSSK
Q9FHS0_ARATH/10-74   ----P------EFPIGRVKKIM-----KLD.KDI------NK..INSEALHVITYSTELFLHFLAEKSA.VVT-AEKKRK
O45550_CAEEL/20-84   ----I------GLPQKGINQII-----KEVVPEM------R...IANESRDMINACCVEFVKHIAREAQ.RIA-SQDQRK
TBAP_HUMAN/11-75     ----L------TIPRAAINKMI-----KETLPNV------R...VANDARELVVNCCTEFIHLISSEAN.EIC-NKSEKK
O14348_SCHPO/10-75   ----L------SLPKATVQKMV-----SDILPVD------LT..FTKEARDLLIECCVEFIHLVSSEAN.EIC-EKEAKK
DR1_ARATH/14-79      ----A------SLPKATMTKII-----KEMLPPD------VR..VARDAQDLLIECCVEFINLVSSESN.DVC-NKEDKR
O81130_ARATH/32-97   ----Q------YMPIANVIRIM-----RKTLPSH------AK..ISDDAKETIQECVSEYISFVTGEAN.ERC-QREQRK
PHP3_SCHPO/10-75     ----N------LLPIANVARIM-----KSALPEN------AK..ISKEAKDCVQDCVSEFISFVTGEAS.EQC-TQEKRK
O59848_ASPOR/46-111  ----R------WLPIANVARIM-----KLALPDN------AK..IAKEAKECMQECVSEFISFITSEAS.EKC-QQEKRK
O76256_SCHMA/27-92   ----R------FLPIANVAKIM-----KRAVPGN------GK..IAKDAKECVQECVSEFISFITSELP.DKC-QTEKRK
O82248_ARATH/54-119  ----R------LLPIANVGRIM-----KNILPAN------AK..VSKEAKETMQECVSEFISFVTGEAS.DKC-HKEKRK
O04027_ARATH/6-71    ----R------LLPIANVGRLM-----KQILPSN------AK..ISKEAKQTVQECATEFISFVTCEAS.EKC-HRENRK
HAP3_KLULA/25-90     ----R------WLPINNVARLM-----KNTLPAT------TK..VSKDAKECMQECVSEFISFVTSEAC.DRC-TSGKRK
O17286_CAEEL/64-129  ----R------FLPIANVVRIM-----KTQMDPQ------AK..LAKDAKECAQECVSEFISFIASEAA.EIC-NITKRK
HMFB_METFE/1-64      ----M------ELPIAPIGRII-----KDA..GA------ER..VSDDARITLAKILEEMGRDIASEAI.KLA-RHAGRK
HFOB_METFO/1-64      ----M------ELPIAPIGRII-----KNA..GA------ER..VSDDAREALAKALEEKGETIATEAV.KLA-KHAGRK
HFO2_METFO/1-64      ----A------ELPIAPVGRII-----KNA..GA------QR..ISDDAKEALAKALEENGEELAKKAV.ELA-KHAGRK
HAF2_ARCFU/2-65      ----A------ELPMAPVDRLI-----RKA..GA------ER..VSADAVEKMVEVLEDYAITVAKKAV.EIA-KHSGRK
HAF1_ARCFU/6-69      ----V------ELPLAPVERLL-----RKA..GA------SR..VSEDAKVELAKAIEEYAMQIGKKAA.ELA-KHAGRK
HJA2_METJA/2-65      ----A------ELPVAPFERIL-----KKA..GA------ER..VSRAAAEYLAEAVEEIALEIAKEAV.ELA-KHAKRK
HARB_PYRSG/1-64      ----A------ELPIAPVDRLI-----RKA..GA------QR..VSEQAAKLLAEHLEEKALEIARKAV.DLA-KHAGRK
O93641_METKA/3-67    ----V------ELPKAAIERIF-----RQG.IGE------RR..LSQDAKDTIYDFVPTMAEYVANAAK.SVL-DASGKK
```

TAF

```
Q8SR29_ENCCU/1-65    ----.M-----LFSKETLKSFA-----QSKGISN----------IDDDALRVLSQDLEYRIKEVCQEGS-KFM-VGSKRT
Q8LRG9_ORYSA/1-68    ----MS-----IVPKETIEVIG-----QSVGIAN----------LPADVSAALAPDVEYRLREIMQEAI-KCM-RHAKRT
Q9MAU3_ARATH/1-68    ----MS-----IVPKETVEVIA-----QSIGITN----------LLPEAALMLAPDVEYRVREIMQEAI-KCM-RHSKRT
Q9SLJ8_ARATH/1-64    ----..-----MVTKESIEVIA-----QSIGLST----------LSPDVSAALAPDVEYRVREVMQEAI-KCM-RHARRT
TAF6_YEAST/9-74      ----TI-----WSPQDTVKDVA-----ESLGLEN----------INDDVLKALAMDVEYRILEIIEQAV-KFK-RHSKRD
O74462_SCHPO/3-68    ----LT-----VWNIESIKDVA-----EMLGIGN----------LADEPAAAIAMDLEYRIHQVVQEAT-KFM-VHSKRT
Q9DDC0_PLEWA/11-76   ----NT-----LFPTESMKVIA-----ESIGISQ----------VPEETCQLITEEVSYRIKEITQDAL-KFM-SVGKRQ
TAF6_XENLA/9-74      ----NT-----LLPSESMKVIS-----ESVGISQ----------MSEETCQLLAQEVSFRIKEVTQDAL-KFM-HVGKRQ
TAF6_HUMAN/11-76     ----NT-----VLPSESMKVVA-----ESMGIAQ----------IQEETCQLLTDEVSYRIKEIAQDAL-KFM-HMGKRQ
TAF6_DROME/18-83     ----GS-----SISAESMKVIA-----ESIGVGS----------LSDDAAKELAEDVSIKLKRIVQDAA-KFM-NHAKRQ
TAF6L_HUMAN/9-73     ----FV-----EIPRESVRLMA-----ESTGL.E----------LSDEVAALLAEDVCYRLREATQNSS-QFM-KHTKRR
Q9GZI6_CAEEL/24-85   ----TP-----IFTQTA....A-----EMLGITS----------LNTEAAELLEFLSREKLKEIVRLSA-KWT-QKSARR
```

TFIID_20kDa

```
O13722_SCHPO/344-411 ----S------KRKLHDLLQQI.....DSEEK------------IEPEVEELLLEIADEFVESVTNFAC-RLA-KHRKSD
TAF12_YEAST/416-488  ----S------KRKLRELVKTVGIDEGDGETV------------IDGDVEELLLDLADDFVTNVTAFSC-RLA-KHRKSD
Q9SR71_ARATH/401-468 ----G------KRSIHELLQQI.....DPSEK------------LDPEVEDIISDIAEDFVESITTFGC-SLA-KHRKSD
Q9LNR1_ARATH/526-593 ----G------KRKIQDLVSQV.....DVHAK------------LDPDVEDLLLEVADDFIDSVTSFAC-SLA-KHRKSS
TAF12_HUMAN/59-126   ----T------KKKLQDLVREV.....DPNEQ------------LDEDVEEMLLQIADDFIESVVTAAC-QLA-RHRKSS
TAF12_DROME/93-160   ----T------KPRLTELVREV.....DTTTQ------------LDEDVEELLLQIIDDFVEDTVKSTS-AFA-KHRKSN
Q9U226_CAEEL/233-297 ----.--------..KLDDLMQQI.....SSTTV------------LEENVKDVIVEYADDFVSSLIDKAC-KMI-KNREVK
Q9VR21_DROME/67-133  ----S------KTNMLQFVQKI.....DANSS------------LDDQGCDMMARIADAFVNDISMRIV-KLA-KYRKSD
```

TAFII28

```
TAF11_DROME/89-177   MQVLVSNFTEEQLDRYEMYRRSAFPKAAVKRLMQT...ITG.CSVSQNVVIAMSGIAKVFVGEVVEEALDVME-..AQGE
TAF11_MOUSE/106-194  MQILVSSFSEEQLNRYEMYRRSAFPKAAIKRLIQS...ITG.TSVSQNVVIAMSGISKVFVGEVVEEALDVCE-..KWGE
Q21417_CAEEL/221-312 NQVLVANFSQEQIERYEVYRKSSFKKSTIKRLINE...FTGGIDLGKKVDIAVAGLAKVLVGEIVEEALDIRD-LDEKEA
Q20563_CAEEL/236-327 TQVLLSNFSQEQLERYESYRRSSFQKSTIRRLISQ...YTGGVNVGQSVVIAIAGLAKVFVGEVVEEALDIRD-INEEEA
Q9LNS8_ARATH/128-217 MQKILSQFTEEQMSRYESFRRSGFKKSDMEKLVQR...ITGGPKMDDTMNIVVRGIAKMFVGDLVETARVVMR-..ERKE
```

http://www.sanger.ac.uk/Software/Pfam/

So How Are We Doing ?

- 172 Clans
  - Contains 1181 Pfams (15%)
  - Largest Clan is NADP_Rossmann
    - 53 families
    - Covers 4 SCOP sf
    - Added over 5000 domain hits to Pfam
  - Largest family without a structure is MFS
    - 19 families
    - Also added over 5000 domain hits to Pfam
  - 66% have a structure representative.
  - many families w/o structure can now be related to a structure

http://www.sanger.ac.uk/Software/Pfam/

Conclusions

- Majority of proteins have Pfam domains
- Pfam helps to understand protein interactions
- Pfam clans give heirarchy

http://www.sanger.ac.uk/Software/Pfam/