



The Pfam and MEROPS databases

EMBO course 2004

Robert Finn (rdf@sanger.ac.uk)



Organisation of Tutorial

Part 1 – Background and Practical on Pfam

**Part 2 - Background and Practical on
MEROPS**

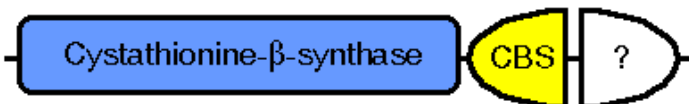
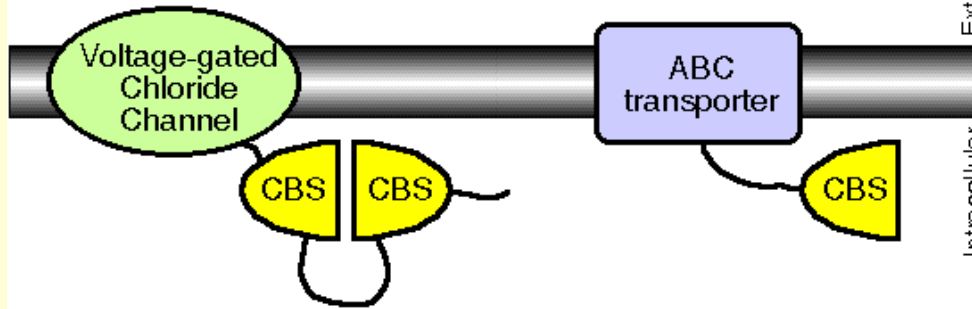


Summary

- **Introduction to Pfam**
 - What is Pfam?
 - Sequence Coverage
 - Using Pfam
- **More Advanced Topics**
 - Pfam and Protein Structures
 - Pfam Clans
 - *i*Pfam

What is Pfam ?

Proteins containing CBS domains



Domains can be considered as building blocks of proteins.

Some domains can be found in many proteins with different functions, while others are only found in proteins with a certain function.

The presence of a particular domain can be indicative of the function of the protein.

Pfam is a domain database.

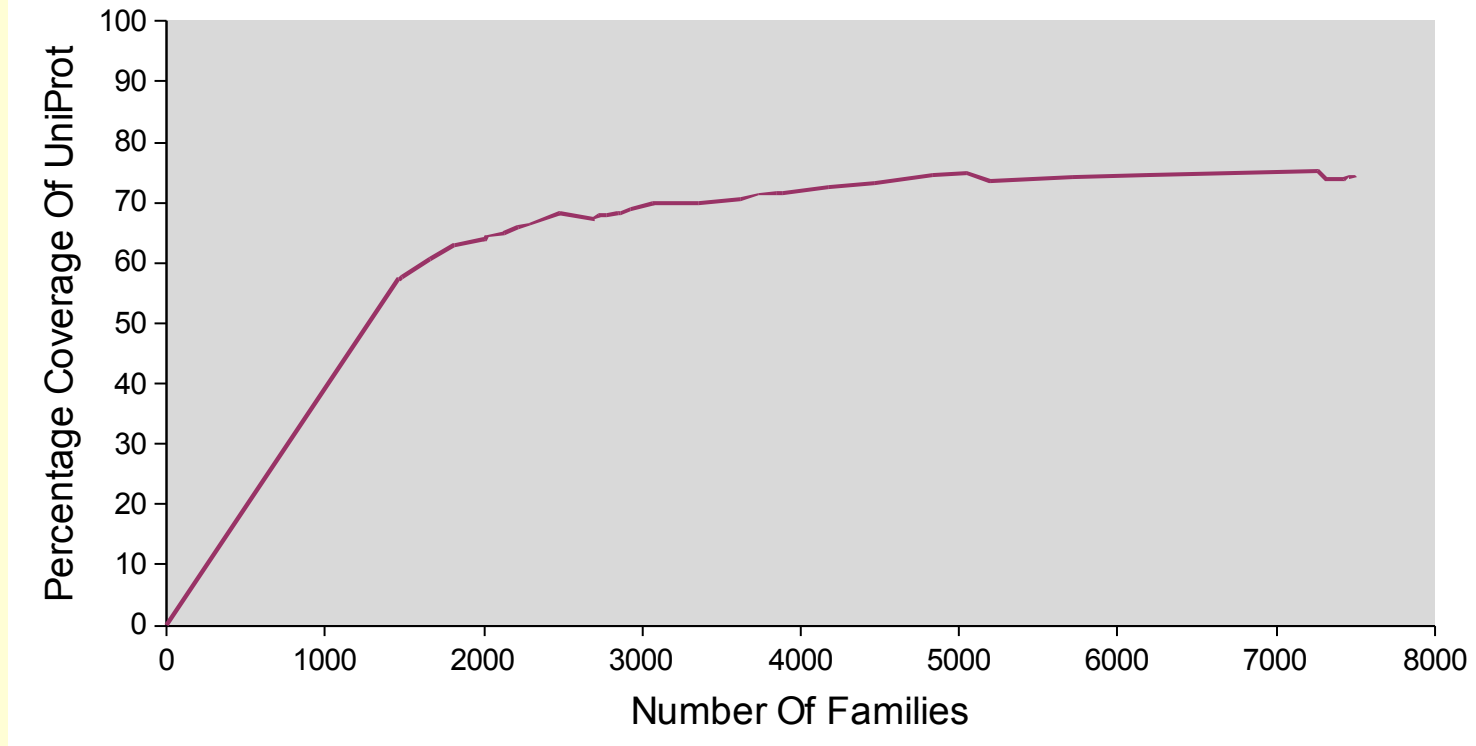
Comprised of two parts – Pfam-A and Pfam-B.

Pfam is use by many different groups in many different ways. Originally set up to aid the annotation the *C. elegans* genomes.

What is a Pfam-A Entry?

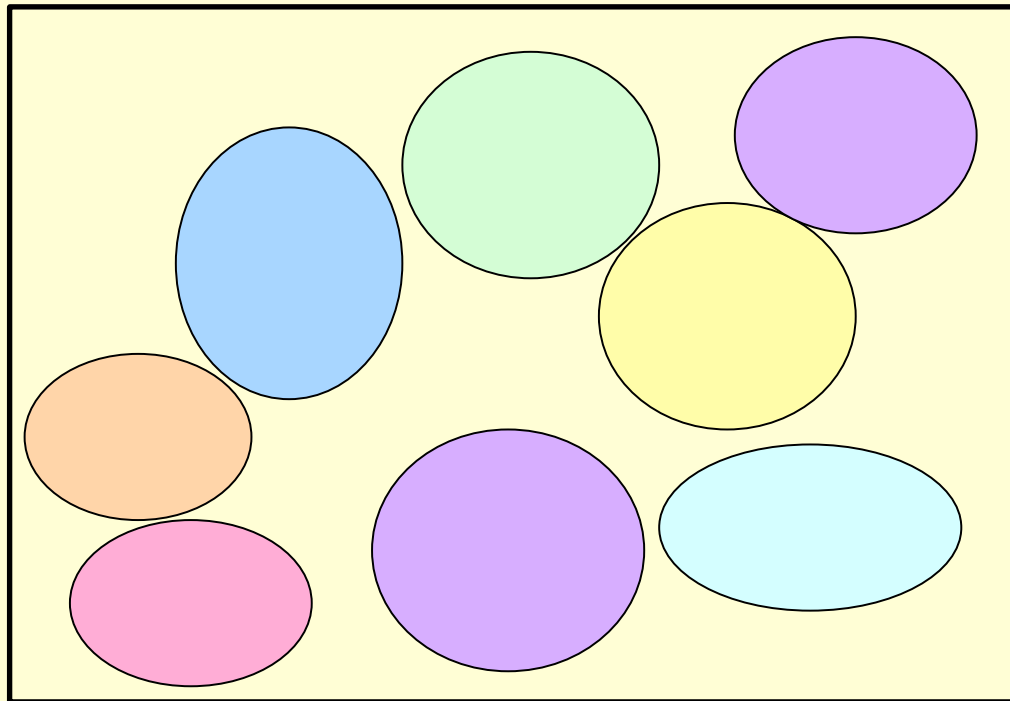
- A SEED alignment – contains a set or representative sequences
- HMM – built using the SEED alignment
- A full alignment – contains all (detectable) sequences in the family
- A description of the family, includes thresholds you to create the full alignment
- Rules – No false positives. A family is not allowed to overlap with any other family

Pfam Coverage



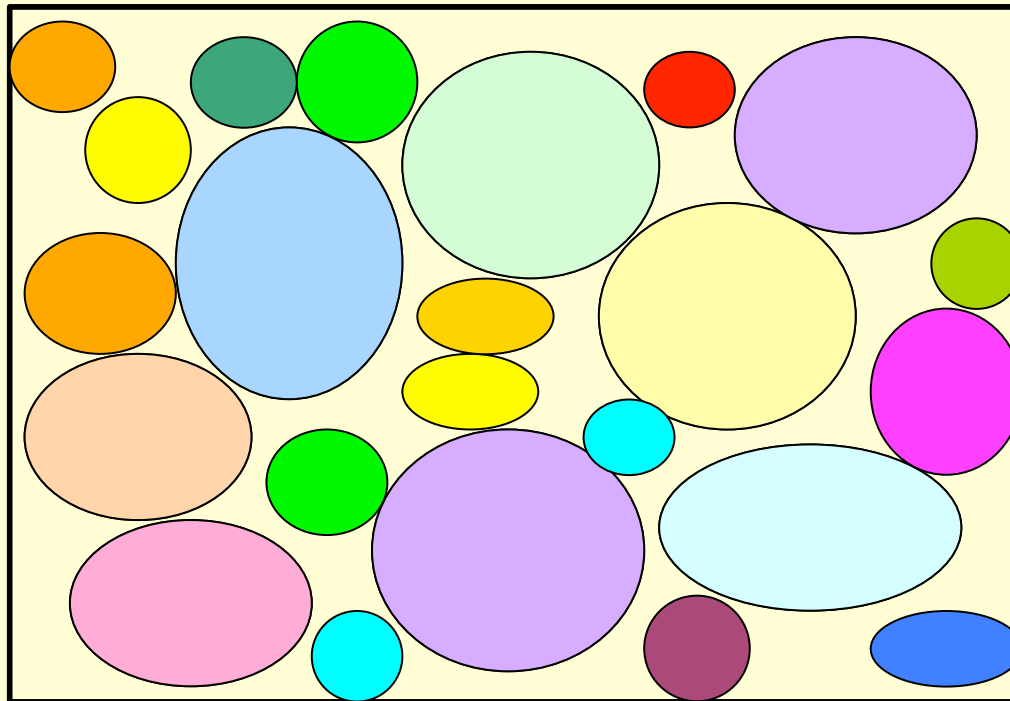
- First 2000 families covered ~ 65% of UniProt
- Currently, 7503 families cover 74% of UniProt

Pfam Sequence Coverage



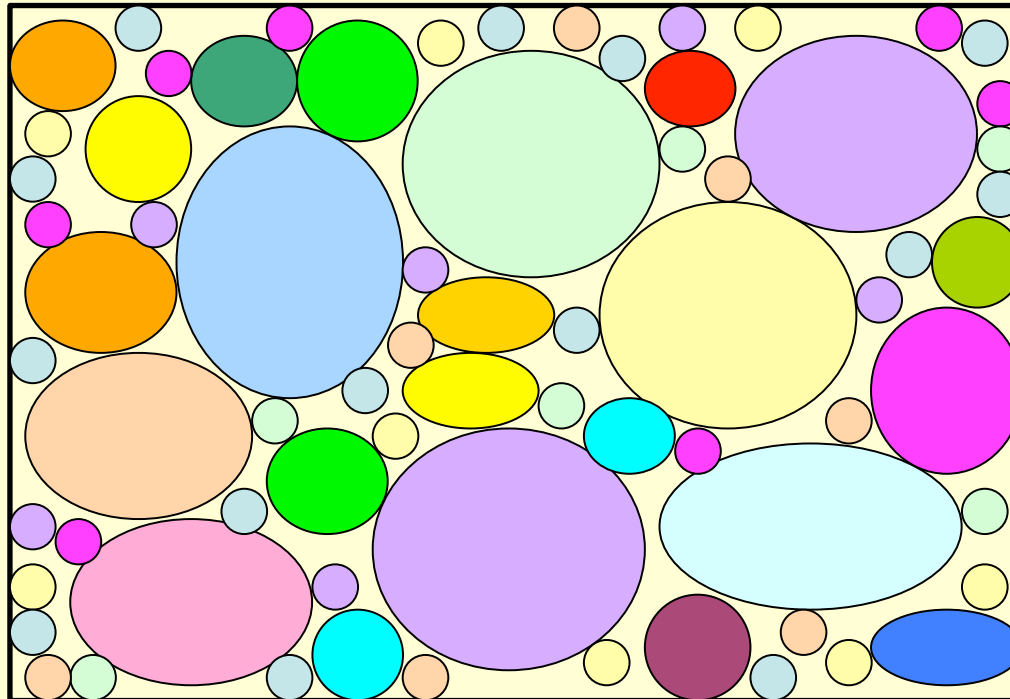
So why does the curve look logarithmic ?

Pfam Sequence Coverage



So why does the curve look logarithmic ?

Pfam Sequence Coverage



So why does the curve look logarithmic ?



Pfam-B

- Pfam-A covers about 74% of sequences
- To be comprehensive we have Pfam-B
- There are over 140,000 Pfam-B
- They cover 24% of UniProt (not covered by Pfam-A)
- Automatically generated clusters that are derived from Prodom



Pfam – Nuts and Bolts

- **Collection of sequence alignments and profile hidden Markov models (HMMs)**
- **Over 7,500 families**
- **mySQL database**
- **Bi-Monthly Releases - flatfiles and relational tables**
- **Current Release – 15.0**
- **Mirrored around the World**



Searching Pfam

- **Two Fundamental Ways of Searching Pfam**
 - **By Sequence**
 - Website – Demonstrated in the practical
 - Download HMM libraries and Run Locally
 - **By Domain**
 - Website – Demonstrated in the practical
 - Flatfiles & RDB



YFD is absent from Pfam.....

- **Send us an Alignment and Some Annotation and we will, in most cases, add it to Pfam.**
- **Build Your Own HMM and use of to search a sequence database.**



More Advanced Topics.....



Pfam & Structure

- **Part of a collaborative Project called eFamily**
 - **Structural Markups**
 - **Alignment Markup**
 - **Domain Comparison**

Structural Markup

- 1m6n – SecA Translocation ATPase



Domain	End	Chain Start
SecA_DEAD	1	A
SecA_PP_bind	338	226
Helicase_C	530	448
SecA_SW	780	568

- This is also applied to structures

Alignment Markup

```

MASY NEOCR/11-531-AS
MASY CANTR/15-540
MASY CANTR/15-540-AS
MASY PICAN/21-543
MASY PICAN/21-543-AS
MASY ECOLI/11-533
MASY ECOLI/11-533-AS
Q9U3Q5/416-923
GCP CAEEL/443-956
GCP CAEEL/443-956-AS
MASY STRCL/10-540
MASY STRCL/10-540-AS
MASY STRAE/10-542
MASZ PSEFL/17-697
MASZ PSEFL/17-697-AS
MASZ MYCLE/17-704
MASZ MYCLE/17-704-AS
MASZ CORGL/34-712
MASZ CORGL/34-712-AS
MASZ ECOLI/16-696
MASZ ECOLI/16-696-SS
MASZ ECOLI/16-696-SA
MASZ ECOLI/16-696-AS

```

```

..VPINY...LMEDAATAEVSRTQIQWVTHGAKTDT.G
..VPINN...LMEDAATAEVSRLQLYSWCKHAVKMDDTG
..VPIYG...LMEDAATAEISRTSIWQWIHHQKTLN.G
..VPLDN...LMEDLATAEISRQLWQWLHHEAKLED.G
..VPLYN...LMEDAATAEISRQLWQWLHHDAKLED.G
..VAIFH...LMEDAATAEISRSQIQWINAGVVLEN.G
..VAIFN...LMEDAATAEISRSQIQWINAGVEFEH.G
SKVPDINDVGLMEDRATLRIS SQHIANWLRHGVVTQDQ.
SKVPDIHNVALMEDRATLRISQQLANWLRHGVITSED.
SKVPDIHDIDL MEDRATLRISQMLANWIRHDVVSKEQ.
SKVPDIHNVALMEDRATLRIS SQHIANWLRHGILTKEQ.
EEEECTTCEEECHHHHHHHHHHHHHHHHHHTTSSCHHH.
130213663220100010000000000001325204442.

```

- AS – active site
- SS – secondary structure
- SA – solvent accessibility
- DSSP is used to calculate SS and SA
- MSD-UniProt Mapping used for the markup

Domain Comparison

- Often it is useful to compare Pfam domains to other domain databases
- Pfam provides a convenient tool for comparing domains between Pfam, CATH and SCOP
- Domains can be compared in 2D or 3D
- Explored Further in parctical

MASZ_ECOLI matches PDB identifier : [1d8c:A:17-697](#)

Pfam Domain Organisation of MASZ_ECOLI:



Malate_synthase 16-696

SCOP Domain Organisation of 1d8c (chain A) :



SCOP:Malate synthase G 1-722

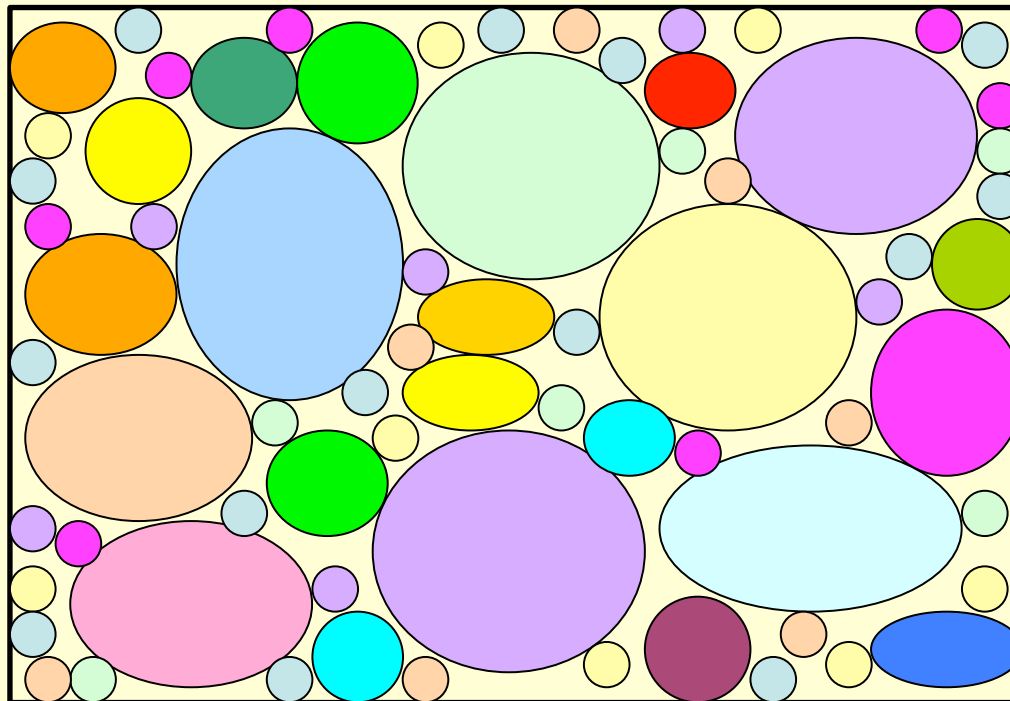
CATH Domain Organisation of 1d8c (chain A) :



CATH:2.170.170.11.1.1 134-261

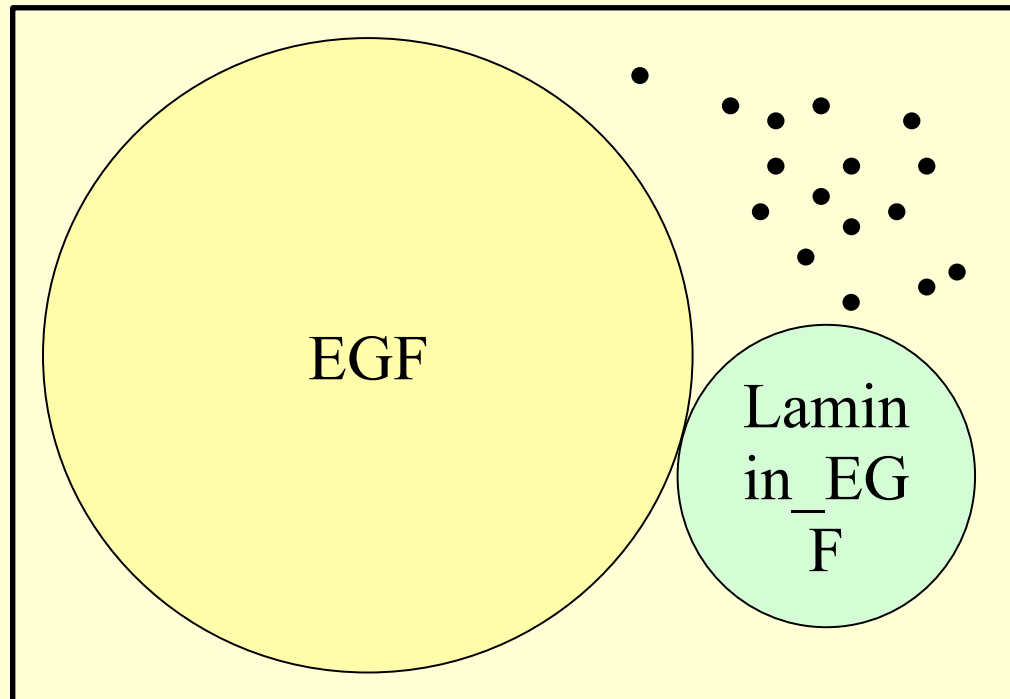
CATH:2.170.170.11.1.1 295-332

Pfam Sequence Coverage



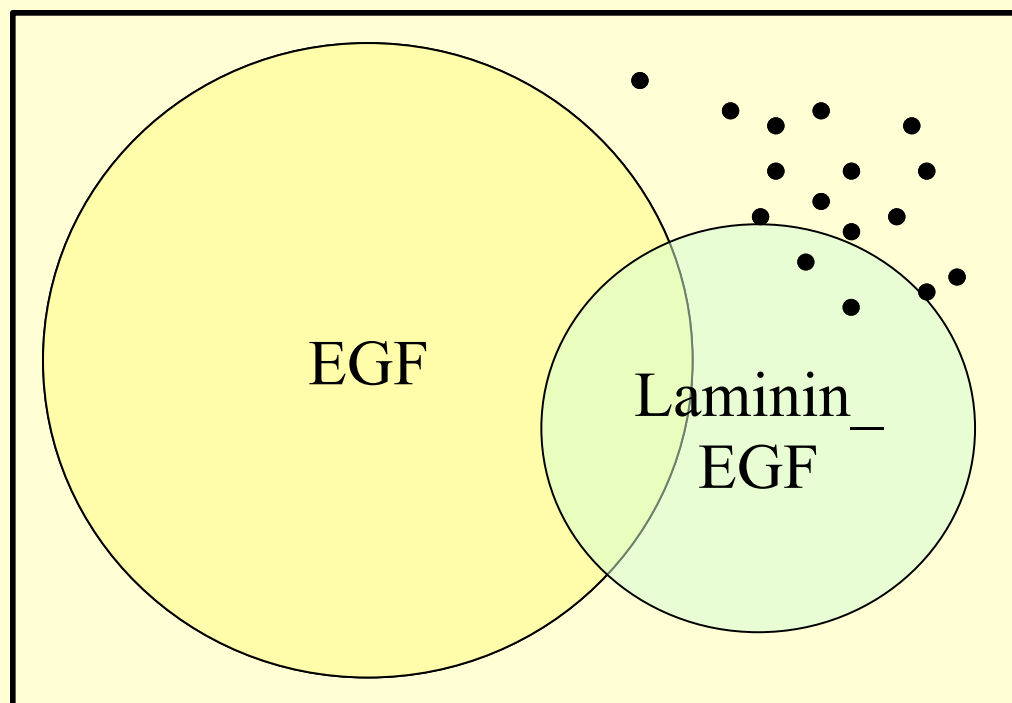
Pfam Clans

- Lets focus in.....
- Two related families in Pfam



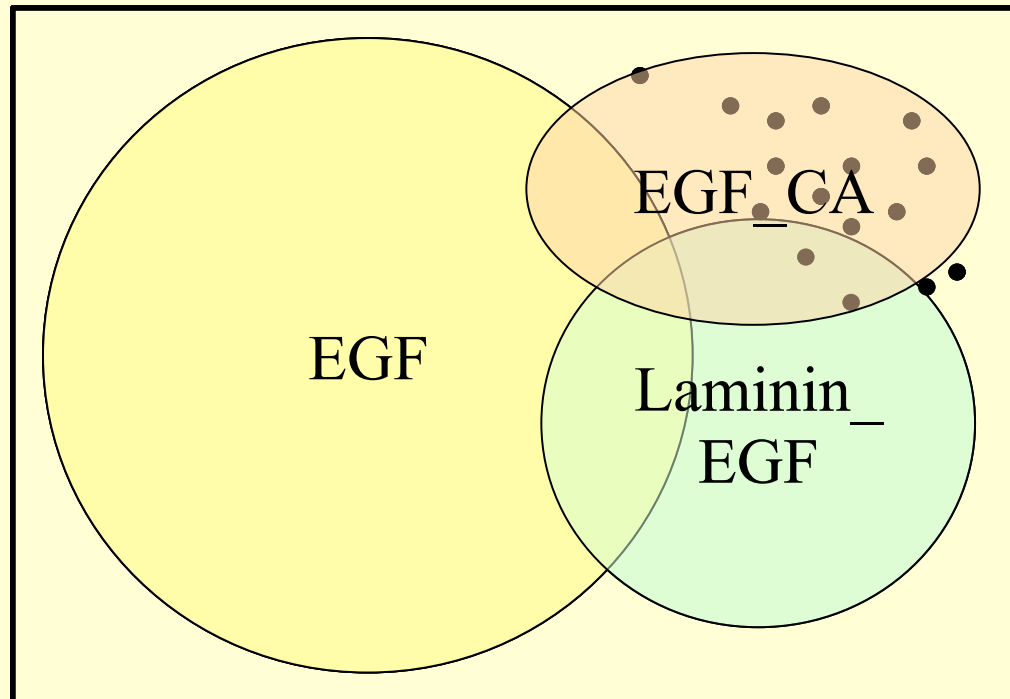
Pfam Clans

- Two related families in Pfam, but now they overlap



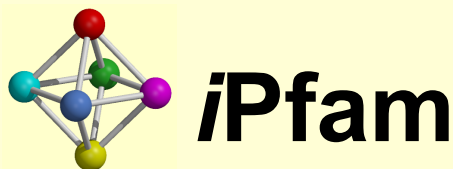
Pfam Clans

- Add a new family to the Clan to get missing sequences



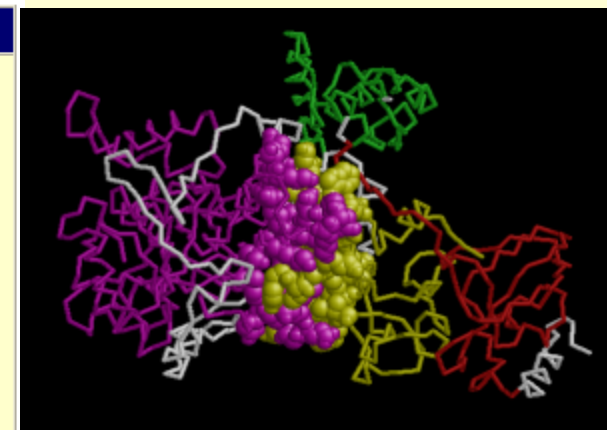
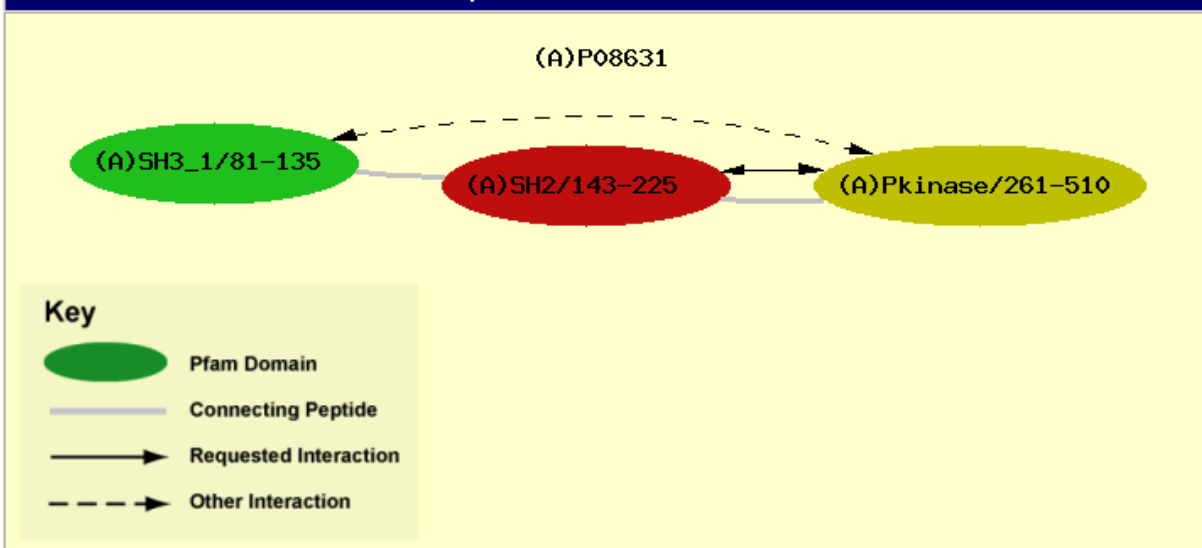
Clan Entry Page

EGF superfamily	
Author:	Finn RD, Bateman A
Comment:	Members of this clan all belong to the EGF superfamily. This particular superfamily is characterised as having least 6 cysteines residues. These cysteine form disulphide bonds, in the order 1-3, 2-4, 5-6, which are essential for the stability of the EGF fold. These disulphide bonds are stacked in a ladder-like arrangement. The Laminin EGF family is distinguished by having an an additional disulphide bond. The function of the domains within this family remains unclear, but they are though to largely perform a structural role. More often than not, there domains are arranged a tandem repeats in extracellular proteins.
Member families:	Laminin EGF , EGF_CA , EGF
Literature references:	1 Structure and function of epidermal growth factor-like regions in proteins. FEBS Lett 1988;231:1-4. Appella E, Weber IT, Blasi F; 2 Domain structure and organisation in extracellular matrix proteins. Matrix Biol 2002;21:115-128. Hohenester E, Engel J;
Database references:	CATH ; SCOP ;



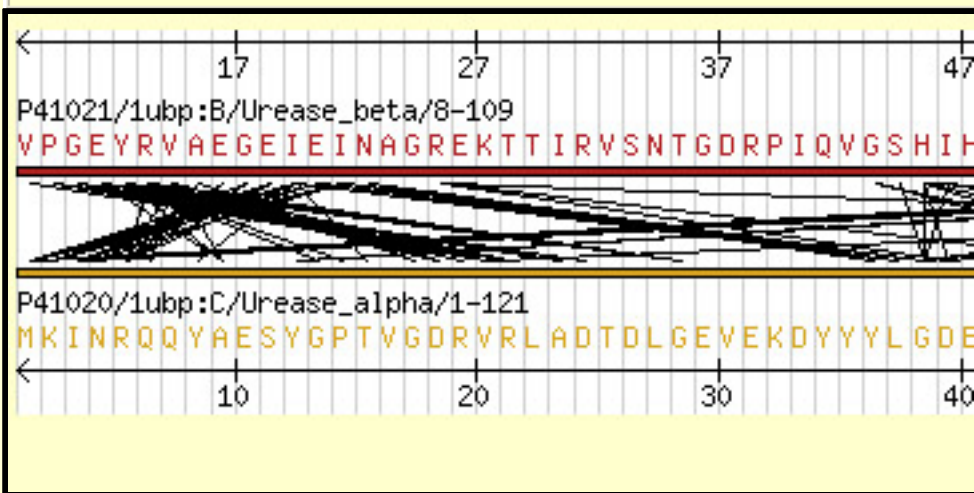
- **What is *i*Pfam?**
 - A database of Pfam domain interactions in known structures
 - Interaction information is contained at the level of domains, residues and atoms.
 - Information is available from the view point of PDB structure or UniProt Sequence

Pfam Domain Interaction Network From 1qcf



```

    .---RVYHYRINTASDG.....KLYVSSERFNTLAELVHH
    .---KVEHYRIMYHASK.....LSIDEEVYFENLMQLVEH
    1k9a_E
    1k9a_A
    SETTKGAYSLSIRDWDDmKGDHVKHYKIRKLDNG.....GYYITTRAQFETLQQLVQH
    1g83_A
    SESAPGDFSLSVKFGN----DVQHFKVLRDQAG.....KYFLMVVKFNSLNELVDY
    1gri_A
    SESTAGSFSLSVRDFDQhQGEVVKHYKIRNLDNG.....GFYISPRITFPGLHELVRH
    11ck_A
    SESVPGVYCLCVLYHG----YIYTYRVSQTETGswsaeTAPGVHKRYFRKIKNLISA
    1m27_A
  
```



Further Reading

- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, KhannaA, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D138-41. PMID: 14681378
- The Pfam website contains many help pages and answers to FAQ
- pfam@sanger.ac.uk - will answer specific queries
- There is a section in Current Protocols in bioinformatics that explains in detail how to use Pfam.
- Biological Sequence Analysis: Probablistic Models of Proteins and Nucleic Acids ~ Richard Durbin, et al
- Stockholm Format - <http://www.cgr.ki.se/cgb/groups/sonnhammer/Stockholm.html>
- Efamily - <http://www.efamily.org.uk>



Pfam Practical

Now go to the following page:

<http://www.sanger.ac.uk/Users/rdf/EMBO/section1.html>