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
  - iPfam
What is Pfam?

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 used by many different groups in many different ways. Originally set up to aid the annotation of 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
- First 2000 families covered ~ 65% of UniProt
- Currently, 7503 families cover 74% of UniProt
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 | Chain Start
  - SecA_DEAD | A
  - 1 | 382
  - SecA_PP_bind | A
  - 338 | 226
  - Helicase_C | A
  - 530 | 448
  - SecA_SW | A
  - 780 | 568

- This is also applied to structures
Alignment Markup

- **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 practical
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**

<table>
<thead>
<tr>
<th>Author</th>
<th>Finn RD, Bateman A.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comment</strong></td>
<td>Members of this clan all belong to the EGF superfamily. This particular superfamily is characterised as having least 5 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, these domains are arranged a tandem repeats in extracellular proteins.</td>
</tr>
<tr>
<td><strong>Member families</strong></td>
<td>Laminin, EGF, EGF CA, EGF</td>
</tr>
<tr>
<td><strong>Database references</strong></td>
<td>CATH; SCOP;</td>
</tr>
</tbody>
</table>
What is iPfam?

- 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 viewpoint of PDB structure or UniProt Sequence
Further Reading


- 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.


- Efamily - http://www.efamily.org.uk
Pfam Practical

Now go to the following page:

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