Overview

- Proteins Structure (Ch 2)
  - Primary Structure
  - Secondary structure
  - Tertiary structure
  - Quaternary structure

- Structure → Function

- For all levels of above:
  Bioinformatics tools
  (structure & sequence based)

- PDB
  - Sources
  - Bias
  - Structural Genomics

- Structure alignment

- Structural Classification
  - SCOP/ CATH
  - Domains
  - Convergence vs divergence
  - Disorder

- Structure prediction (Part 5, 6)
  - Preview: homology vs 'ab initio'

Levels of protein structure

- Primary: amino acid sequence
- Secondary: helix, loop, beta strand
- Tertiary: 3D orientation of secondary structure (topology)
- Quaternary: Interaction with other molecules, and peptide chains

Primary Protein Structure

- How could the codon table still be interesting for Bioinformatics research?

Special Amino acids

- Cysteine:
  - Can create a disulphide bridge
  - Post-translational modification
  - Covalent bond

- Proline
  - Forms ring with backbone,
  - Different phi/psi angles

- Glycine
  - No side chain
  - Different phi/psi angles
In 1951 Pauling, Robert Corey, and Herman Branson correctly proposed the alpha helix and beta sheet as the primary structural motifs in protein secondary structure based on the (1) planarity of the peptide bond, (2) favourable hydrogen bonding patterns and (3) steric hindrance.

- In what respect are beta strands “less local” than alpha-helices?
  - Can you think of any consequences?
Loops & flexible regions

- Loops tend to be more flexible
- Hydrogen bonds may be satisfied by
  - backbone
  - side chain
  - water

Secondary Structure prediction

- How would you go about this?

Super-secondary Structure

Tertiary structure

- How are the secondary structure elements connected
  - Topology (SCOP)
  - Architecture (CATH)

Why is water so important?

- How would water affect the preferred peptide configuration?
- The two major effects that stabilize protein folds depend heavily on the solvent... how?

Hydrophobic Collapse

- Click to add an outline
- Click to add an outline
**Protein folding**

- **Favorable energy vs Favorable entropy**

**Quaternary Structure**

Bovine 11-ATPase (1E79). This ATP synthase unit is responsible for the production of ATP within the cell.

Each chain is displayed in a different colour. Together these chains make up the quaternary structure.

**Structure & Function**

Classic Picture:
Sequence → Structure → Function

or

Once we have a structure, the function will follow

**Structure & Function**

- Protein binding sites (enzymes)
- Protein interaction surface
- How could one recognize the functional site?
  - Given structure?
  - Given sequence?
- Functional sites often located on 'structural rule' exceptions

**Protein – Protein interactions (PPI)**

**PDB**

- Protein DataBank
  - X-ray structures
  - NMR structures
  - Cryo-electron microscopy
- Biases in PDB
  - Proteins that we can:
    - Purify
    - Crystalize
    - Stabilize in solution
    - Sequence bias
- TM proteins
**Question:** Bioinformatics & Protein Structure

Which computational methods, do you know, could you imagine, or do you think would be useful to have:

think about: validation, comparison, function, prediction, searches, targets for research, testing of methods etc.

<table>
<thead>
<tr>
<th>Output / Goal</th>
<th>Input: Structure</th>
<th>Input: Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary structure</td>
<td>Lipid structure</td>
<td>Lipid structure</td>
</tr>
<tr>
<td>Secondary Structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tertiary Structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quaternary Structure</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PDB**

- Protein DataBank
  - X-ray structures
  - NMR structures
  - cryo-electron microscopy
- Biases in PDB
  - proteins that we can:
    - purify
    - crystallize
    - stabilize in solution
  - Sequence bias
- TM proteins

**Structural Genomics**

- So, we have the ability to sequence full genomes. What is next?
- Let’s determine the structure of entire genomes (2000)
- Do we now have complete structural coverage? NO!
- Has there been a noticeable effect? → see SCOP / CATH later

There is a good reason that it takes most crystallography/NMR labs nearly 2 years to establish a protein structure.

**Transmembrane proteins**

- How important is the solvent for protein folding?
- So what about membrane proteins? How do they fold?

Lipid tails: hydrophobic
Lipid heads: hydrophilic

- Would membrane proteins show different evolutionary patterns?

**Transmembrane proteins in PDB**

- TM proteins are difficult to:
  - Express
  - Purify
  - Crystallize
  - Stabilize in solution
Disordered regions missing in PDB

Disordered regions:
- Missing in X-ray structures
- Typically removed for crystallization
- 33% of eukaryotic proteins contain large disordered segment
- Associated functions:
  - Signaling
  - Regulatory
- Disordered flanks found next to binding motifs:
  - Hydrophobic binding motif
  - Hydrophilic flanks

Structure is more conserved than sequence

SCOP

class
fold
superfamily
family

secondary structure
similar fold structure
evolutionary related
sequence similarity

SCOP hierarchy

Click to add title
Structural Classification

**SCOP**

- **Family**
  - Proteins are clustered together into families on the basis of one of two criteria that imply having a common evolutionary origin:
  - All proteins that have sequence similarities.
  - Proteins with whose functions and structures are extremely similar.

- **Superfamily**
  - If proteins have low sequence identities but share common evolutionary origins, they are placed together in superfamilies; for example, the variable and constant domains of immunoglobulins.

- **Fold**
  - Superfamilies and families are defined as having a common fold if their proteins have the same major secondary structures in the same arrangement and with the same topological connections [recent reviews see \((3,4)\)]. The structural similarities of proteins in the same fold category, probably arise from the physics and chemistry of proteins favouring certain packing arrangements and chain topologies.

SCOP - website

Let's have a look:
http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html

CATH: Architecture

- **Definitions**
  - Independent folding unit (with evidence it folds independently)
  - Recurring set of secondary structure elements (through structural alignments)
  - Compact substructure of protein

Fold Classes

**SCOP**

- all alpha
- all beta
- alpha + beta

CATH

- C: Class
- A: Architecture
- T: Topology
- H: Homology

Structural Domains

- Definitions
  - Independent folding unit (with evidence it folds independently)
  - Recurring set of secondary structure elements (through structural alignments)
  - Compact substructure of protein

CATH: Architecture
Domain Definitions

- What different domain definitions you come up with considering:
  - Sequence
  - Structure
  - Physics
  - Evolution

Problems in Structure Classification

- Fold and domain definitions: what is the right answer?
  - Manual assignment
    - Delays
  - Domain definitions

Convergent vs Divergent evolution

- Convergent evolution:
  There exists two similar structured (e.g. same fold) proteins, that are not homologous. This would imply the same structure has been found twice independently by nature

- Divergent evolution:
  There exists an ancestral protein domain that is homologous to two current protein domains that have a similar structure (e.g. fold)

Importance of Structure Classification

- Homology gold standard
  - Using structure information
    - Structure more conserved than sequence

- Functional site comparison

- On domain level, easier to predict structure, function etc

Midnight zone (sequence only)

- Click to add an outline

Rost et al. 1998
**Evolutionary signal**

Sequence profile of protein family (Ig family – pfam)

---

**PSI-BLAST profiles**

PSI-BLAST uses the power of databases AND evolutionary signal

---

**Pfam Database**

Each Pfam-A family consists of a curated seed alignment containing a small set of representative members of the family, profile hidden Markov models (profile HMMs) built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences belonging to the family, as defined by profile HMM searches of primary sequence databases.

---

**Pfam Ig Family Alignment**

---

**Clan pages in Pfam**

Clan pages in Pfam (A) A screen shot of a clan summary page, containing the description, annotation and membership of the clan. From this page, the user can view the family relationship diagram (B). Each family in the clan is represented by a blue box and its relationship to other families is represented by solid lines (significant profile–profile comparison score) or dashed lines (non-significant profile–profile comparison score). Below each box, the profile–profile comparison E-value score is presented. This score is also linked to a visualization of the profile–profile comparison alignment (C). The clan summary page also provides a link to the clan alignment (D). The clan alignment is a multiple sequence alignment of all of the clan members (seed alignments and set of seed sequences are separated by the alternate background shading). The alignments are colored using Jalview.

---
Pfam Coverage

- Click to add an outline

<table>
<thead>
<tr>
<th>PfamID</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF0001</td>
<td>63.2%</td>
</tr>
<tr>
<td>PF0002</td>
<td>64.4%</td>
</tr>
<tr>
<td>PF0003</td>
<td>65.5%</td>
</tr>
<tr>
<td>PF0004</td>
<td>66.6%</td>
</tr>
<tr>
<td>PF0005</td>
<td>67.7%</td>
</tr>
<tr>
<td>PF0006</td>
<td>68.8%</td>
</tr>
</tbody>
</table>

Sequence Structure 'Gap'

- Against 65,000 protein structures
- Can we do better?

Structure Prediction

How would you go about this?

Key concepts

- Proteins Structure (Ch 2)
  - Primary Structure
  - Secondary structure
  - Tertiary structure
  - Quaternary structure

- Structure → Function

- Structure more conserved than sequence

- For all levels of above:
  - Bioinformatics tools (structure & sequence based)
    - Structure prediction (Part 5, 6)
      - Preview: homology vs 'ab initio'

- PDB
  - Sources
  - Bias
  - Structural Genomics

- Structure alignment

- Structural Classification
  - SCOP/CATH
  - Domains
  - Convergence vs divergence
  - disorder