**Protein structures**

- Proteins are linear polypeptide chains (one or more).
- Building blocks: 20 types of amino acids.
- Range from a few 10s to 1000s.
- They “fold” into varying three-dimensional shapes.

**The basics of protein**

- Certain level of function can be found without structure. But a structure is a key to understand the detailed mechanism.
- A predicted structure is a powerful tool for function inference.

**Relevance of Protein Structure in the Post-Genome Era**

**Structure-Function Relationship**

- Trp repressor as a function switch.

**Structure-Based Drug Design**

- Structure-based rational drug design is a major method for drug discovery.
- HIV protease inhibitor.

**Formation of polypeptide chain**

- Partial double bond character.
Levels of protein structure

- Primary structure
- Secondary structure
- Tertiary structure
- Quaternary structure

Primary structure

- This is simply the amino acid sequences of polypeptides chains (proteins).

Secondary structure

- Local organization of protein backbone: α-helix, β-strand (groups of β-strands assemble into β-sheet), turn and interconnecting loop.

Secondary structure

- Secondary structure is also a linear information and represented as string similar to amino acid sequence of proteins

Tertiary structure

- Three-dimensional coordinates of the atoms of a chain is its tertiary structure (the structure of a single chain of a protein)
- Quaternary structure: describes how different changes are positioned relatively (the overall protein structure)

Tertiary structure

Packing the secondary structure elements into a compact spatial unit
Quaternary structure

Assembly of homo or heteromeric protein chains.

Usually the functional unit of a protein, especially for enzymes.

---

Structural Motifs

- Four helix bundle
- Helix-loop-helix
- Coiled coil

---

PDB Files: the "header"

- Primary and secondary structure are ONE-dimensional; Tertiary and quaternary structure are THREE-dimensional.
- "structure" usually refers to 3-D structure of protein.
### PDB Files: the coordinates

<table>
<thead>
<tr>
<th>Atom &amp; Residue</th>
<th>XYZ Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM 1  O PHE A</td>
<td>12.063 25.960 -55.390</td>
</tr>
<tr>
<td>ATOM 2  CA PHE A</td>
<td>12.063 25.960 -55.390</td>
</tr>
<tr>
<td>ATOM 3  C PHE A</td>
<td>12.063 25.960 -55.390</td>
</tr>
<tr>
<td>ATOM 4  O PHE A</td>
<td>12.063 25.960 -55.390</td>
</tr>
</tbody>
</table>

### Experimental techniques for structure determination

- X-ray Crystallography
- Nuclear Magnetic Resonance spectroscopy (NMR)
- Electron Microscopy/Diffraction

### X-ray Crystallography

- From small molecules to viruses
- Information about the positions of individual atoms
- Limited information about dynamics
- Requires crystals

### NMR

- Limited to molecules up to ~50kDa (good quality up to 30 kDa)
- Information about distances between pairs of atoms
- A 2-d resonance spectrum with off-diagonal peaks
- Requires soluble, non-aggregating material

### What does resolution mean for the structure?

- Low resolution, 6Å
- Medium resolution, 3Å
- High resolution, 1.5Å
Secondary structure prediction

- Given a protein sequence (primary structure)

  GHMIATRGQLEAYEDYRHFSRCPFIP

- Predict its secondary structure content
  (C=coils  H=Alpha Helix  E=Beta Strands)
  CEEEEECHHHHHHHHHHHCCCHCC

Why Secondary Structure Prediction?

- Easier problem than 3D structure prediction (more than 40 years of history).
- Accurate secondary structure prediction can be an important information for the tertiary structure prediction
- Improving sequence alignment accuracy
- Protein function prediction/classification

Prediction Methods

- Statistical methods
  - Chou-Fasman method, GOR I-IV
  - Nearest neighbors
    - NNSSP, SSPAL
  - Neural network
    - PHD, Psi-Pred, J-Pred
  - Support vector machines
  - Hidden Markov Models

Chou-Fasman method

- Compute parameters for amino acids
  - Preference to be in
    - alpha helix: $P(a)$
    - beta sheet: $P(b)$
    - Turn: $P(t)$
  - Frequencies with which the amino acid is in the 1st, 2nd, 3rd, and 4th position of a turn: $f(i), f(i+1), f(i+2), f(i+3)$.
- Use a sliding window

SSE prediction by Chou-Fasman

- Alpha-helix prediction
  - Find all regions where 4 of the 6 amino acids in window have $P(a) > 100$.
  - Extend the region in both directions unless 4 consecutive residues have $P(a) < 100$.
  - If $\sum P(a) > \sum P(b)$ then the region is predicted to be alpha-helix.
- Beta-sheet prediction is analogous.
- Turn prediction
  - Compute $P(t) = f(i) \cdot f(i+1) \cdot f(i+2) \cdot f(i+3)$ for 4 consecutive residues.
  - Predict a turn if
    - $P(t) > 0.000075$ (check)
    - The average $P(t) > 100$
  - $\sum P(t) > \sum P(a)$ and $\sum P(t) > \sum P(b)$

GOR method

- Use a sliding window of 17 residues
- Compute the frequencies with which each amino acid occupies the 17 positions in helix, sheet, and turn.
- Use this to predict the SSE probability of each residue.
A Simple and Fast Secondary Structure Prediction Method using Hidden Neural Networks

Kuang Lin, Victor A. Simossis, Willam R. Taylor, Jaap Heringa
Bioinformatics Advance Access published September 17, 2004

A protein folds into a unique 3D structure under the physiological condition: determine this structure.

Lysozyme sequence:

KVFGRCELAA AMKRHGLDNY
RGYSLGNWVC AAKFESNFNT
QATNRNTDGS TDYGILQINS
RWWCNDGRTP GSRNLCNIPC
SALLSSDITA SVNCAKKIVS
DGNGMNAWVA WRNRCKGTDV
QAWIRGCRL

Consider a 100 residue protein. If each residue can have only 3 conformations, there are $3^{100} \times 5 \times 10^{47}$ possible conformations.

If it takes $10^{-13}$s to convert from 1 structure to another, exhaustive search would take $1.6 \times 10^{27}$ years!

Folding must proceed by progressive stabilization of intermediates.

It is believed that hydrophobic collapse is a key driving force for protein folding.

- Hydrophobic core
- Polar surface interacting with solvent
- Minimum volume (no cavities)
- Disulfide bond formation stabilizes structure
- Hydrogen bonds
- Polar and electrostatic interactions

Proteins are, in fact, only marginally stable.

Native state is little more stable than the unfolded form

Many proteins help in folding

- Protein disulfide isomerase – catalyzes shuffling of disulfide bonds
- Chaperones – alter protein structures and (in theory) unfold misfolded proteins

Hemoglobin is the protein in red blood cells (erythrocytes) responsible for binding oxygen.

The mutation E→V in the β chain replaces a charged Glu by a hydrophobic Val on the surface of hemoglobin

The resulting “sticky patch” causes hemoglobin to stick together and form fibers which deform the red blood cell and do not carry oxygen efficiently

Sickle cell anemia was the first identified molecular disease
Sequestering hydrophobic residues in the protein core protects proteins from hydrophobic agglutination (sticking together).

### Protein Structure Prediction

- *Ab-initio* techniques
- Homology modeling
  - Sequence-sequence comparison
- Protein threading
  - Sequence-structure comparison