Structural biology

From sequence to structure
Proteins form into distinct shapes
Aims of structural biology

- Predict the confirmation of a poly-peptide chain
- Predict and analyze the function of a protein
Protein structures

• Elements of structure
  – The amino acids
  – Levels
  – Databases
  – Folds and families

• Prediction

• How to check for correct assignment
  – Ramachandran plot
Structures of biomolecules

- **Primary structure**
  - Amino acid sequence

- **Secondary structure**
  - Local elements
    - Helices
    - Sheets

- **Tertiary structure (3D)**
  - Fold
  - Classification

- **Quaternary structure**
  - Interactions between chains
  - Protein-protein interaction
AMINO ACIDS
Visualizing Proteins

- High complexity
- Multiple levels of structure
- Important properties are “distributed throughout the 3D structure
- No single/simple “point” at which to look
Surface
Kühner S et al. (2009) Nature
PDB/RCSB database

• Protein Data Bank – One of the oldest databases on molecular biology
• Repository of all known structures
  – All published structures must be deposited
• Four-character identifier
Classification of protein structures

CATH database
- Fold
- Superfamily – Secondary structure contacts
- Sequence families
- Domains
- Rule based on secondary structure content, contacts and domain boundaries

SCOP database
- Class
  - All α, all β
  - α / β – Parallel sheets
  - α + β – Antiparallel sheets
  - Multi-domain proteins
  - Membrane
  - Unstructured proteins

- Folds
- Superfamilies
- Families
SRC kinase
FOLDS
Hemoglobin
α/β TIM barrel
All beta Immunglobulin Light chain
PREDICTION OF PROTEIN STRUCTURES
Anfinsen’s dogma (1961)

• Denatured proteins can refold *in vitro*
• No folding machinery required
• All information about the structure resides in the sequence
• Native structure: minimum free energy
  – Unique
  – Stable
  – Kinetically accessible
Levinthal’s Paradox

• Consider a protein with 101 residues
  – 100 $\Psi$ and 100$\varphi$ angels
  – If we assume only three stable positions and none for $\omega$
    – $3^{200}$ or $10^{95}$ confirmations
  – Sampling all confirmations exceeds the life time of the universe

• Proteins fold in milliseconds
Secondary structure

• Single sequence methods
  – Chou-Fasman
  – GOR
• Neural networks
  – PHD
• HMMs
Chou-Fasman

<table>
<thead>
<tr>
<th>Name etc</th>
<th>P(a)</th>
<th>P(b)</th>
<th>P(t)</th>
<th>f(i)</th>
<th>f(i+1)</th>
<th>f(i+2)</th>
<th>f(i+3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanin</td>
<td>142</td>
<td>83</td>
<td>66</td>
<td>0.06</td>
<td>0.076</td>
<td>0.035</td>
<td>0.058</td>
</tr>
<tr>
<td>Threonie</td>
<td>83</td>
<td>119</td>
<td>96</td>
<td>0.086</td>
<td>0.108</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculate if $P(a) > 100$ for 4 out 6 AA, assign helix
Calculate if $P(b) > 100$ for 3 out 5 AA assign sheet
Calculate $p(t) = f(i)$ … assign turn
Further rules to resolve clashes

Chou and Fasman (1974) Biochemistry
Single sequence methods

• Prediction based on propensity of an AA to occur in helix, sheet or turn
• Chou-Fasman
  – Empirical, rule based
• GOR
  – Log-odds score, Bayesian statistics
Neural network

Machine learning technique inspired by neuronal structures
TMHMM
Tertiary structure

• Homology modeling

• Threading
  – Fold recognition

• Ab initio modeling
RMSD

• Root-mean-square deviation

• Distance of backbone atoms
  – Usually $\alpha$

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i^2}$$
Some chemistry

• Intramolecular forces
  – Covalent bonds (400 kcal)
  – Strong but only relevant for cystin

• Intermolecular forces
  – Hydrogen bonds (12 – 16 kcal)
  – Van der Waals forces
    • Dipole-dipole (0.5 -2 kcal)
    • London (<1 kcal)
  – Buried hydrophobic faces
Lennard-Jones potential

- Summarizes the repulsion of atoms and attraction by van der Waals forces

\[ V_{LJ} = 4\varepsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] \]

\[ = \varepsilon \left[ \left( \frac{r_m}{r} \right)^{12} - 2 \left( \frac{r_m}{r} \right)^6 \right] \]

Structure prediction

1. Find backbone structure
   1. Homology modeling
   2. Threading
   3. Ab initio prediction
2. Loop modeling
3. Sidechain packing
4. Refinement
Homology modeling

• Find homologous sequence (BLAST etc)
• Multiple alignment (Muscle etc)
• Replace backbone in defined, conserved parts
• Check core model and re-align
• Model side chain
• Model loop regions
• Energy minimization
Homology modeling

• Simple procedure for ID >40% over 50 AA (typical values, check for plausibility)
• Difficult if ID <25% over reasonable range
• Automated, SWISSMODEL available for all suitable targets
• If no template can be found:
  – Search template with sensitive methods: threading
  – Build from scratch: *ab initio*
Threading

• Naïve approach: Perform Homology Modeling for many/all templates, score the best
• Alignments at low %ID become problematic
• Fold recognition occasionally works, models often fail
Ab initio prediction

• Library of k-mers from known structures
• Build „random“ structures of k-mers
• Optimize in cycles, using a custom scoring function
• Analyze the top structures according to protein-like appearance and/or expectations from the literature.
• ROSETTA (Baker et al. (1998) outperformed contestants in CASP3.)
Problem solved?

• Great improvements for globular proteins

• Open issues
  – Membrane proteins
  – Unstructured regions
  – Large assemblies