Protein Modeling Methods

- Ab initio methods:
  - solution of a protein folding problem
  - search in conformational space

- Energy-based methods:
  - energy minimization
  - molecular simulation

- Knowledge-based methods:
  - homology modeling
  - fold recognition

Knowledge

Knowledge is a pattern that exceeds a certain threshold of interestingness.

Factors that contribute to interestingness:
- coverage
- confidence
- statistical significance
- simplicity
- unexpectedness
- actionability

Knowledge-based methods

Finding patterns in known structures
Deriving rules (usually in the form of PMF)
Applying the rules

Fold Recognition

Pattern searching
- sequence patterns
- structure patterns
- residue composition patterns

Threading
- sequence-structure compatibility
- structure-sequence compatibility
**Threading**

- Only the local environment is taken into account
- Non-local contacts are assumed with generic peptide
- No gaps are allowed in the alignment

**Segmental Threading**

- Only the local environment is taken into account
- Non-local contacts are assumed with generic peptide
- No gaps are allowed in the alignment

**Homology Modeling**

- Identification of structurally conserved regions (using multiple alignment)
- Backbone construction (based on SCR)
- Loop construction (KB or conformational search)
- Side-chain restoration (KB, rotamer, or MM)
- Structure verification and evaluation
- Structure refinement (energy minimization)

**Homology Modeling Programs**

- Swiss-Model (http://swissmodel.expasy.org)
- Modeller (http://salilab.org/modeller)
- CPHmodels (http://www.cbs.dtu.dk/services/CPHmodels)
- Protein Model Portal (PMP) (http://www.proteinmodelportal.org)
Swiss-Model

- **Method:**
  Knowledge-based approach.

- **Requirements:**
  At least one known 3D-structure of a related protein.
  Good quality sequence alignments.

- **Procedures:**
  Superposition of related 3D-structures.
  Generation of a multiple alignment.
  Generation of a framework for the new sequence.
  Rebuild lacking loops.
  Complete and correct backbone.
  Correct and rebuild side chains.
  Verify model structure quality and check packing.
  Refine structure by energy minimisation and molecular dynamics.

Methods and Programs used by Swiss-Model

- **Sequence Alignment**
  ProModII (Peitsch, M.C. Unpublished, Server-specific tool)

- **Knowledge Based Protein Modelling**

- **Energy Minimisation**
  Gromos96 (van Gunsteren W.F. http://igc.ethz.ch/gromos/)

- **Model evaluation**
  Swiss-PdbViewer (http://www.expasy.ch/spdbv/mainpage.html)

Swiss-Model Request Types

- First Approach mode.
- Optimise mode.
- Combine mode.
- GPCR mode.

Model Confidence Factors

The Model B-factors are determined as follows:

- The number of template structures used for model building.
- The deviation of the model from the template structures.
- The Distance trap value used for framework building.

The Model B-factor is computed as:

\[ 85.0 \times \left( \frac{1}{\# \text{selected template str.}} \right) \times \left( \frac{\text{Distance trap}}{2.5} \right) \]

and 99.9 for all atoms added during loop and side-chain building.

Structure verification and validation

**Bond lengths (Procheck)**

<table>
<thead>
<tr>
<th>Bond</th>
<th>Labeling</th>
<th>Value</th>
<th>sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-N</td>
<td>C-NH1 (except Pro)</td>
<td>1.329</td>
<td>0.014</td>
</tr>
<tr>
<td>C-N</td>
<td>C-N (Pro)</td>
<td>1.341</td>
<td>0.016</td>
</tr>
<tr>
<td>C-O</td>
<td>C-O</td>
<td>1.231</td>
<td>0.020</td>
</tr>
<tr>
<td>Calpha-C</td>
<td>C1E-C (except Gly)</td>
<td>1.525</td>
<td>0.021</td>
</tr>
<tr>
<td>Calpha-C</td>
<td>C2G-C (Gly)</td>
<td>1.516</td>
<td>0.018</td>
</tr>
<tr>
<td>Calpha-Cbeta</td>
<td>C1E-C1H3E (Ala)</td>
<td>1.521</td>
<td>0.033</td>
</tr>
<tr>
<td>Calpha-Cbeta</td>
<td>C1E-C1H1E (Ile, Thr, Val)</td>
<td>1.540</td>
<td>0.027</td>
</tr>
<tr>
<td>Calpha-Cbeta</td>
<td>C1E-C1H2E (the rest)</td>
<td>1.530</td>
<td>0.020</td>
</tr>
<tr>
<td>N-Calpa</td>
<td>NH1-C1H1E (except Gly, Pro)</td>
<td>1.458</td>
<td>0.019</td>
</tr>
<tr>
<td>N-Calpa</td>
<td>NH1-C1H2E (Gly)</td>
<td>1.451</td>
<td>0.016</td>
</tr>
<tr>
<td>N-Calpa</td>
<td>N-C1H1E (Pro)</td>
<td>1.466</td>
<td>0.015</td>
</tr>
</tbody>
</table>

**Bond angles (Procheck)**

<table>
<thead>
<tr>
<th>Angle</th>
<th>Labeling</th>
<th>Value</th>
<th>sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-N-Calpha</td>
<td>C-NH1-C1E (except Gly, Pro)</td>
<td>121.7</td>
<td>1.8</td>
</tr>
<tr>
<td>C-N-Calpha</td>
<td>C-NH1-C2G (Gly)</td>
<td>120.6</td>
<td>1.7</td>
</tr>
<tr>
<td>C-N-Calpha</td>
<td>C-N-C1H1E (Pro)</td>
<td>122.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Calpha-C-N</td>
<td>C1E-C-NH1 (except Gly, Pro)</td>
<td>116.2</td>
<td>2.0</td>
</tr>
<tr>
<td>Calpha-C-N</td>
<td>C1E-C-C1H1E (Gly)</td>
<td>116.4</td>
<td>2.1</td>
</tr>
<tr>
<td>Calpha-C-N</td>
<td>C1E-C-C1H2E (Pro)</td>
<td>116.9</td>
<td>1.5</td>
</tr>
<tr>
<td>Calpha-C-O</td>
<td>C1E-C-O (except Gly)</td>
<td>120.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Calpha-C-O</td>
<td>C1E-C-O (Gly)</td>
<td>120.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>
a. Ramachandran plot quality - percentage of the protein's residues that are in the core regions of the Ramachandran plot.

b. Peptide bond planarity - standard deviation of the protein structure's omega torsion angles.

c. Bad non-bonded interactions - number of bad contacts per 100 residues.

d. $\text{C}_{\alpha}$ tetrahedral distortion - standard deviation of the $\zeta$ torsion angle ($\text{C}_{\alpha}$, N, C, and $\text{C}_{\beta}$).

e. Main-chain hydrogen bond energy - standard deviation of the hydrogen bond energies for main-chain hydrogen bonds.

f. Overall G-factor - average of different G-factors for each residue in the structure.