<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Fold recognition</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.20</td>
<td>Exercise: Modelling remote homologues</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td>Summary &amp; discussion</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>
Feedback Persons

http://www.bio-evaluering.dk/
Homology Modelling

Revisited
Why Do We Need Homology Modelling?

• *Ab Initio* protein folding (random sampling):
  – 100 aa, 3 conf./residue gives approximately $10^{48}$ different overall conformations!

• Random sampling is *NOT feasible*, even if conformations can be sampled at picosecond ($10^{-12}$ sec) rates.
  – Levinthal’s paradox

• Do homology modelling instead.
How Is It Possible?

• The structure of a protein is uniquely determined by its amino acid sequence (but sequence is sometimes not enough):
  – prions
  – pH, ions, cofactors, chaperones

• Structure is conserved much longer than sequence in evolution.
  – Structure > Function > Sequence
How Is It Done?

• Identify template(s)
  – Initial alignment
• Improve alignment

• Backbone generation

• Loop modelling
• Side chains
• Refinement

• Validation ←
Improving the Alignment

PHE ASP ILE CYS ARG LEU PRO GLY SER ALA GLU ALA VAL CYS
PHE ASN VAL CYS ARG THR PRO --- --- --- GLU ALA ILE CYS
PHE ASN VAL CYS ARG --- --- --- THR PRO GLU ALA ILE CYS

From "Professional Gambling" by Gert Vriend
http://www.cmbi.kun.nl/gv/articles/text/gambling.html
Template Quality

- Selecting the best template is crucial!
- The best template may not be the one with the highest % id (best p-value…)
  - Template 1: 93% id, 3.5 Å resolution 😞
  - Template 2: 90% id, 1.5 Å resolution 😊
Error Recovery

• Errors in the model can NOT be recovered at a later step
  – The alignment can not make up for a bad choice of template.
  – Loop modeling can not make up for a poor alignment.

• The step where the errors were introduced should be redone.
Validation

• Most programs will get the bond lengths and angles right.

• Model Rama. plot ~ template Rama. plot.
  – select a high quality template!

• Inside/outside distributions of polar and apolar residues.
Summary

• Successful homology modelling depends on the following:
  – Template quality
  – Alignment (add biological information)
  – Modelling program/procedure (use more than one)

• Always validate your final model!
## Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Fold recognition</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.20</td>
<td>Exercise: Modelling remote homologues</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td>Summary &amp; discussion</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>
Fold recognition and \textit{ab initio} protein structure prediction

by Pernille Andersen
Outline

• Threading and pair potentials
• *Ab initio* structure prediction methods
• Human intervention (what kind of knowledge can be used for alignment and selection of templates?)
• Meta-servers (the principle, 3d jury)
• Summary of take-home messages
Threading and pair potentials

- Compares a given sequence against known structures (folds)
- Potentials that describe tendencies observed in known protein structures

Example: Pair potentials
How normal is it to observe a pair of an alanine and a valine separated by 20 residues in the sequence and 3Å in space? (X)

How normal is it to observe any pair of residues separated by 20 residues and 3Å in space? (Y)

Potential: \[ E = -\log \left( \frac{X}{Y} \right) \]
Potentials of mean force

Alignment score from structural fitness (pair potential)

How well does K fit environment at P6?
If P8 is acidic then fine, if P8 is basic then poor
Threading methods today

• Problem: No protein is average

• Interactions in proteins cannot only be described by **pairs** of amino acids

• The information in the potentials is partly captured with sequence profiles or HMMs

• Today mostly used in **HYBRID** approaches in combination with profile-profile based methods

• Potentials can be used to score models based on different templates or alignments

HMM alignment, hhpred
Two-high-scoring predictions by the top groups in FR/H (top) and FR/A (bottom). The assigned z-scores are given for the top predictions (center) as well as for two average predictions (right).

G. Wang  Assessment of fold recognition predictions in CASP6, Proteins 61, S7, Pages 46-66
Ab initio/ free modeling methods

• Aim is to find the fold of native protein by simulating the biological process of protein folding.

• A VERY DIFFICULT task because a protein chain can fold into millions of different conformations.

• Use it **only** when no detectable homologues can be found.

• Methods can also be useful for fold recognition in cases of extremely low homology (e.g. convergent evolution).
Fragment-based *ab initio* modelling

- Rosetta method of the Baker group:
  - Secondary structure prediction
  - Fragments library of 3 and 9 residues from known structures
  - Link fragments together, use only backbone and CB atoms
  - Contact/pair potential
  - Energy minimization techniques (Monte Carlo optimization) to calculate tertiary structure
  - Refine structure including side chains


http://robetta.bakerlab.org/
Energy minimization

The energy of the whole protein model is minimized to obtain the final model.
Potentials for finding good models

- Potentials should make models more “native-like”

  van der Waal’s attractive/repulsive forces

  Pair potentials

  Contact number potentials

  Back bone torsion angle potential

  Solvation potentials

  Hydrogen bond potentials

  Side chain rotamer potentials

  Uroplatus Fimbriatus (gecko)
Problems with empirical potentials

Fragments with correct local structure

Nature’s potential

Empirical potential
CASP6 & Ab Initio (new folds category)

Excellent modelling

The Baker group ( #100) was among the top scoring

Hardest target

The Baker group ( #100) was among the top scoring
Human intervention

- The best groups in CASP use maximum knowledge of query proteins

- Specialists can help to find a correct template and correct alignments

---

Knowledge of function
Cysteines forming disulfide bridges or binding e.g. zinc molecules
Proteolytic cleavage sites
Other metal binding residues
Antibody epitopes or escape mutants
Ligand binding
Results from CD or fluorescence experiments
Human intervention II

- **Fold It: The Protein Folding Game**
- **Rosetta Energy Potentials**

- [http://fold.it/portal/](http://fold.it/portal/)

- Uses the HUMAN brain’s pattern recognition resources for finding the lowest energy fold
Meta-servers

• Democratic modeling
  – The highest scoring hit is often wrong
  – Many prediction methods have the correct fold among the top 10-20 hits
  – If many different prediction methods all have the same fold among the top hits, this fold is probably correct

<table>
<thead>
<tr>
<th>Server 1</th>
<th>Server 2</th>
<th>Server 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template 1 -&gt; Model 1</td>
<td>Template 1 -&gt; <strong>Model 1</strong></td>
<td>Template 2 -&gt; Model 1</td>
</tr>
<tr>
<td>Template 2 -&gt; Model 2</td>
<td>Template 2 -&gt; Model 2</td>
<td>Template 2 -&gt; <strong>Model 2</strong></td>
</tr>
<tr>
<td>Template 3 -&gt; <strong>Model 3</strong></td>
<td>Template 3 -&gt; Model 3</td>
<td>Template 3 -&gt; Model 3</td>
</tr>
</tbody>
</table>
Example of a meta-server

- **3DJury** [http://meta.bioinfo.pl/submit_wizard.pl](http://meta.bioinfo.pl/submit_wizard.pl)
  - Inspired by *Ab initio* modeling methods
    - Average of frequently obtained low energy structures is often closer to the native structure than the lowest energy structure
  - Find most abundant high scoring model in a list of prediction from several predictors
    1. Use output from a set of servers
    2. Superimpose all pairs of structures
    3. Similarity score based on # of Cα pairs within 3.5Å
  - Similar methods developed by A. Elofsson (Pcons [http://pcons.net/](http://pcons.net/)) and D. Fischer (3D shotgun)
3DJury

• Because it is a meta-server it can be slow
• If queue is too long some servers are skipped
• Alternative conformations for a sequence are easily obtained
Take home messages

• Hybrid methods using both threading methods and profile-profile alignments are the best

• Use only *Ab initio* methods if necessary and know that the quality is really low!

• Try to use as much knowledge as possible for alignment and template selections in difficult cases

• Use meta-servers when you can

• TRY FOLDIT!
## Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Last week’s quiz results + Summary</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Fold recognition</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.20</td>
<td>Exercise: Modelling remote homologues</td>
</tr>
<tr>
<td>11.20-11.40</td>
<td>Summary &amp; discussion</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>