# 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>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>
Feedback Persons
Summary of Last Week

• Prediction of structural features from sequence.
  – Secondary structure
  – Disorder
  – Surface accessibility
  – Site predictions
    • Glycosylation, phosphorylation, signal peptides, TM helices…
Site predictions

- **Phosphorylation**
  - S/T or Y

- **Glycosylation**
  - N: N
  - O: S/T
<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>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>
Homology Modelling
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 $(10^{-12} \text{ sec})$ 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 Often Can We Do It?

• There are currently ~80000 structures in the PDB
  – Reduces to ~10000 structures <30% identical (sequence) with a resolution <3.0 Å.

• 25% of all sequences can be modelled.

• 50% can be assigned to a fold class.
Worldwide Structural Genomics

- "Fold space coverage"
- Complete genomes
  - Disease-causing organisms
  - Model organisms
- Membrane proteins
- Protein-ligand interactions

Hou et al., PNAS 2003, 100: 2386-2390
Homology Modeling & Structural Genomics

Table 1: Leveraging of experimental structures by comparative modeling

<table>
<thead>
<tr>
<th>Experimental Structure</th>
<th>Models or fold assignments</th>
<th>Models</th>
<th>Useful models</th>
<th>Less accurate models</th>
<th>Fold assignments only</th>
</tr>
</thead>
<tbody>
<tr>
<td>P005</td>
<td>537</td>
<td>345</td>
<td>53</td>
<td>292</td>
<td>192</td>
</tr>
<tr>
<td>P007</td>
<td>42</td>
<td>40</td>
<td>28</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>P008</td>
<td>31</td>
<td>29</td>
<td>24</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>P018</td>
<td>172</td>
<td>50</td>
<td>11</td>
<td>39</td>
<td>122</td>
</tr>
<tr>
<td>P100</td>
<td>185</td>
<td>70</td>
<td>11</td>
<td>59</td>
<td>115</td>
</tr>
<tr>
<td>P102</td>
<td>26</td>
<td>25</td>
<td>22</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>P111</td>
<td>46</td>
<td>44</td>
<td>23</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1039</strong></td>
<td><strong>603</strong></td>
<td><strong>172</strong></td>
<td><strong>431</strong></td>
<td><strong>436</strong></td>
</tr>
</tbody>
</table>

1A model is counted if it is at least 60 residues long and is assessed to have >30% of its Cα atoms within 3.5 Å of their true positions. The models are subdivided into two classes. “Useful models” are defined to be based on >30% sequence identity to the known structure, while “Less accurate models” are based on <30% sequence identity. “Fold assignments only” denotes the number of proteins with a significant PSI-BLAST relationship to a known structure (E < 0.0001) that failed to produce a reliable model. The calculations were performed in August, 2000.

How Well Can We Do It?

How Is It Done?

- Identify template(s)  
  - Initial alignment
- Improve alignment
- Backbone generation
- Loop modelling
- Side chains
- Refinement
- Validation ↬
Template Identification

- Search with sequence
  - Blast
  - Psi-Blast
  - Fold recognition methods

- Use biological information

- Functional annotation in databases

- Active site/motifs
## Alignment

![Alignment Table](image)

**Figure 25.3.** A typical residue exchange or scoring matrix used by alignment algorithms. Because the score for aligning residues A and B is normally the same as for B and A, this matrix is symmetric.
```
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

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>D</th>
<th>I</th>
<th>C</th>
<th>R</th>
<th>L</th>
<th>P</th>
<th>G</th>
<th>S</th>
<th>A</th>
<th>E</th>
<th>A</th>
<th>V</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>6</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>N</td>
<td>-3</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>5</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>5</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-3</td>
<td>2</td>
<td>-2</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>-3</td>
<td>5</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>8</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>8</td>
</tr>
</tbody>
</table>
```
<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>D</th>
<th>I</th>
<th>C</th>
<th>R</th>
<th>L</th>
<th>P</th>
<th>G</th>
<th>S</th>
<th>A</th>
<th>E</th>
<th>A</th>
<th>V</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>6</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>N</td>
<td>-3</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>8</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>8</td>
</tr>
<tr>
<td>R</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>5</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>T</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>P</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>E</td>
<td>-3</td>
<td>2</td>
<td>-2</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>-3</td>
<td>5</td>
<td>-2</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>8</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>8</td>
</tr>
</tbody>
</table>
Improving the Alignment

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 😊
The Importance of Resolution

- 4 Å
- 3 Å
- 2 Å
- 1 Å

low

high

0.5 Å
Ramachandran Plot

- Allowed backbone torsion angles in proteins

The Ramachandran Plot.

- Beta-sheet
- Left handed alpha-helix
- Right handed alpha-helix

Peptide torsion angles.

Amino acid residue
Template Quality – Ramachandran Plot

X-ray structure – good data.

NMR structure – low quality data...
Backbone Generation

• Generate the backbone coordinates from the template for the aligned regions.

• Several programs can do this, most of the groups at CASP6 use Modeller:
  http://www.salilab.org/modeller/
Loop Modelling – Connecting the Bits

• Knowledge based:
  – PDB search for fragments matching the target sequence (Levitt, Holm, Baker etc.).

• Energy based:
  – Uses an energy function to evaluate the quality of the loop and minimizes this function by Monte Carlo (sampling) or molecular dynamics (MD) techniques.

Loops – the Rosetta Method

• Find fragments (10 per amino acid) with the same sequence and secondary structure profile as the query sequence.

• Combine them using a Monte Carlo scheme to build the loop.

David Baker et al.
Side Chains

• Side chain rotamers are dependent on backbone conformation.

• Most successful method in CASP6 was SCWRL by Dunbrack et al.:
  – Graph-theory knowledge based method to solve the combinatorial problem of side chain modelling.

http://dunbrack.fccc.edu/scwrl4/index.php
Side Chains

• Prediction accuracy:
• High for buried residues
• Low for surface residues
  – Experimental reasons: side chains at the surface are more flexible.
  – Theoretical reasons: much easier to handle hydrophobic packing in the core than the electrostatic interactions including H-bonds to waters.
Side Chains

• If the seq. id is high, the networks of side chain contacts may be conserved.
• Keeping the side chain rotamers from the template may be better than predicting new ones.
Refinement

• Energy minimization

• Molecular dynamics
  – Removes big errors
    • Atom clashes
  – Introduces small errors

• Keep minimization to a minimum.
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.
Validation – ProQ Server

• ProQ is a neural network-based predictor
  – Structural features $\rightarrow$ quality of a protein model.

• ProQ is optimized to find
  – correct models…
  – …NOT (necessarily) native structures.

ProQ server: http://www.sbc.su.se/~bjornw/ProQ/ProQ.html
Structure Validation Program Suites

• **ProCheck**
  
  [http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/](http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)

• **WhatIf server**
  
  [http://swift.cmbi.ru.nl/whatif/](http://swift.cmbi.ru.nl/whatif/)
Homology Modelling Servers

• Eva-CM performs continuous and automated analysis of comparative protein structure modeling servers

• Current list of the best performing servers:

http://www.pdg.cnb.uam.es/eva/doc/intro_cm.html
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>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>
## 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>Homology modelling – Part I</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.15-11.30</td>
<td>Exercise: Homology modelling</td>
</tr>
<tr>
<td>11.30-11.40</td>
<td>Break</td>
</tr>
<tr>
<td>11.40-12.00</td>
<td>Quiz</td>
</tr>
</tbody>
</table>