**Programme**

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Quiz results</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Protein folds, de novo structure prediction and (structural) genomics (revisited)</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.45-12.00</td>
<td>PyMOL hand-in exercise II</td>
</tr>
</tbody>
</table>

**Feedback Person(s)**
Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Quiz results</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Protein folds, de novo structure prediction and (structural) genomics (revisited)</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.45-12.00</td>
<td>PyMOL hand-in exercise II</td>
</tr>
</tbody>
</table>

Outline & Learning Objectives

• After today’s lecture you should be able to:
  – explain how genomic sequencing can be used in relation to structural modelling of proteins.
  – describe and select appropriate methods for structural modelling of proteins depending on the level of data available.
  – explain how experimental data can be integrated in the structural modelling of proteins.
Genomics

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

Hou et al., PNAS 2003, 100: 2386-2390
PDB Growth


Structural Coverage of Swiss-Prot

Protein Folds in PDB

No new folds!
Data Levels I

• Structure of (close) homologue available
  • Quality of model: High
  • Workload: Low
  homology modelling (any size)
  – Modeller (HHpred)
  – SwissModel
  – CPHmodels

Data Levels II

• Structure of remote homologue exists
  • Quality of model: Medium
  • Workload: Intermediate
  fold recognition
  homology modelling (any size)
  – HHpred
Data Levels III

- Many (diverse) sequences from same family available but no structure
  - Structure from sequence
    - EVfold (any size)
    - ROSETTA (up to medium size)

- Quality of model: Low to intermediate
- Workload: Intermediate to high

Data Levels IV

- Few (or low diversity) sequences from same family but no structure
  - Ab initio structure prediction (small)
    - ROSETTA
    - Classical physics force fields
    - Quantum mechanics-based methods

- Quality of model: Generally low
- Workload: Career choice (!)
Building with fragments

The ROSETTA method

**Rosetta**

*The “Frankenstein” method*

- Search for overlapping fragments of 3 and 9 amino acids in PDB
- Stitch together many models
- Refine and score according to energy function.
Design of a New Fold

Directional Statistics

“Sampling Realistic Protein Conformations Using Local Structural Bias”, Hamelryck et al.

*PLoS Computational Biology* 2006
Directional Statistics Distributions

Hidden Markov Model
Ab Initio Structure Prediction

Performance

<table>
<thead>
<tr>
<th>Target Protein</th>
<th>ROSETTA</th>
<th>FB5–HMM (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;6 Å</td>
<td>RMSD</td>
</tr>
<tr>
<td></td>
<td>(Percent)</td>
<td>(Å)</td>
</tr>
<tr>
<td>Protein A</td>
<td>95</td>
<td>3.3</td>
</tr>
<tr>
<td>Homeodomain</td>
<td>47</td>
<td>2.7</td>
</tr>
<tr>
<td>Protein G</td>
<td>0</td>
<td>6.3</td>
</tr>
<tr>
<td>Cro repressor</td>
<td>18</td>
<td>4.2</td>
</tr>
<tr>
<td>Protein L7/L12</td>
<td>6</td>
<td>5.3</td>
</tr>
<tr>
<td>Calbindin</td>
<td>17</td>
<td>4.7</td>
</tr>
</tbody>
</table>

(Column 1) Protein target name.
(Column 2,3) Percentage of good decoys (RMSD < 6 Å) and RMSD of the best decoy (Å) predicted by ROSETTA (out of 100 predictions).
(Column 4,5) Percentage of good decoys and RMSD of the best decoy generated by FB5–HMM (out of 100,000 compact decoys, method S in Table 1).
DOI: 10.1371/journal.pcbi.0020131.t004
EVfold

“Protein structure prediction from sequence variation”, Marks et al. Nature Biotechnology 2012

Direct Information from Sequence
Membrane Proteins

Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing
Hopf et al. 2012 Cell
Combining Modelling and Experimental Data

Determining the architectures of macromolecular assemblies

Adding Details Through Data

The Nuclear Pore Complex

a. Cytoplasmic side and Nucleoplasmic side

b. Heat map and network graph

c. Cross-section and side-view of the nuclear pore complex

d. 3D model of the nuclear pore complex
Summary

• Several methods exist to model protein structures:
  – From simple homology modelling…
  – …fragment and sequence-guided modelling…
  – …to true ab initio calculations based on classical or quantum mechanics.

• Learn to select the proper method!

Useful Links

• EVfold: http://evfold.org/evfold-web/evfold.do
• HHpred: http://toolkit.tuebingen.mpg.de/hhpred
• Modeller http://salilab.org/modeller/
• ROSETTA: https://www.rosettacommons.org
• Links: http://bioinformatics.ca/links_directory/
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00-8.20</td>
<td>Quiz results</td>
</tr>
<tr>
<td>8.20-9.00</td>
<td>Protein folds, de novo structure prediction and (structural) genomics (revisited)</td>
</tr>
<tr>
<td>9.00-9.15</td>
<td>Break</td>
</tr>
<tr>
<td>9.45-12.00</td>
<td>PyMOL hand-in exercise II</td>
</tr>
</tbody>
</table>