Homology Modelling

Thomas Holberg Blicher
Novozymes A/S

Why are Protein Structures so Interesting?

• They provide a detailed picture of interesting biological features, such as active site, substrate specificity, allosteric regulation etc.

• They aid in rational drug design and protein engineering.

• They can elucidate evolutionary relationships undetectable by sequence comparisons.

• They can be used to put mutations in the proper structural context.
Learning Objectives

• Outline the basic steps in comparative protein structure modelling.

• Explain how structure models can be used to support biological hypotheses.

• Perform simple homology modelling using web servers and evaluate the results.

PDB Growth 1971-2013

The PDB also contains nucleotide and nucleotide analogue structures.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray</td>
<td>13116</td>
<td>30860</td>
<td>86817</td>
</tr>
<tr>
<td>NMR</td>
<td>2451</td>
<td>5368</td>
<td>10324</td>
</tr>
<tr>
<td>Other</td>
<td>338</td>
<td>200</td>
<td>976</td>
</tr>
<tr>
<td>Total</td>
<td>15905</td>
<td>36428</td>
<td>98117</td>
</tr>
</tbody>
</table>
Growth of Genbank

No Structure = No Go?

• Do homology modelling.
  – Use known protein structures to make reliable models.

• Validate the model.

• Adjust expectations and use accordingly!
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 Often Can We Do It?

• Currently **103,000** structures in the PDB
  – Reduces to **22,570** structures (chains) <30 % identical (sequence).
  – These fall in ~**1,400** different structure classes (folds).

• ~**25%** of all sequences can be modelled.

• ~**50%** can be assigned to a fold class.

Protein Folds (SCOP) in PDB

No new folds!
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>
<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>P035</td>
<td>167</td>
<td>148</td>
<td>62</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>P080</td>
<td>62</td>
<td>40</td>
<td>20</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>P018</td>
<td>112</td>
<td>50</td>
<td>11</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>P100</td>
<td>105</td>
<td>75</td>
<td>10</td>
<td>11</td>
<td>115</td>
</tr>
<tr>
<td>P092</td>
<td>26</td>
<td>21</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>P111</td>
<td>45</td>
<td>44</td>
<td>23</td>
<td>21</td>
<td>2</td>
</tr>
</tbody>
</table>

Total 1639 603 172 131 436

A model is counted if it is at least 40 residues long and is assumed to have 50% of its residues within 3 Å of their true positions. The models are sorted into three classes: "Initial 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 ≤ 0.001) that failed to produce a reliable model. The calculations were performed in August 2006.

How Well Can We Do It?

- Comparable to medium resolution NMR, low resolution crystallography
- Docking of small ligands, proteins.
- Molecular replacement in crystallography.
- Supporting site-directed mutagenesis.
- Refining NMR structures.
- Finding binding/active sites by 3D motif searching.
- Annotating function by fold assignment.

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

Sequence vs. Structure

• Residues in the same column in an alignment are either:
  – Structurally equivalent/similar
  – Evolutionary equivalent/related/homologous

• Different types of similarity not necessarily equivalent.

• Use biological information to guide/adjust your alignment.
  – Functional annotation in databases
  – Active site/motifs
Alignment

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.
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
Key Parameters

- **Resolution**
- **R values**
  - Agreement between data and model.
  - Usually between 0.15 and 0.25, should not exceed 0.30.
    - $R \approx \text{Resolution} / 10$
    - $R + 0.05 > R_{\text{free}} > R$.

- **Ramachandran plot**
  - The majority of residues in most favoured regions.

- **B factors**
  - Contributions from static and dynamic disorder
    - Well determined $\sim$10-20 Å$^2$, intermediate $\sim$20-30 Å$^2$, flexible 30-50 Å$^2$, invisible $>60$ Å$^2$.

Template Quality – Ramachandran Plot

X-ray structure – good data. NMR structure – low quality data…
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.
Model Validation – ProQ

- ProQ is a neural network-based predictor
  - Structural features → quality of a protein model.

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

- Two quality measures:
  - MaxSub & LGscore

Arne Elofssons group: http://www.sbc.su.se/~bjorn/ProQ/

No Template – No Go?

- Use De novo/ab initio/free modelling methods:
  - 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).
Fold recognition models in CASP6

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

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/
Problems with empirical potentials

Fragments with correct local structure

Nature’s potential

Empirical potential

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
Summary

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

• Always validate your final model!

Modelling Servers

• Comparative (homology) modelling:
  – CPHmodels (simple)
    • [http://www.cbs.dtu.dk/services/CPHmodels/](http://www.cbs.dtu.dk/services/CPHmodels/)
  – SwissModel (intermediate)
  – HHpred (complex)
    • [http://toolkit.tuebingen.mpg.de/hhpred](http://toolkit.tuebingen.mpg.de/hhpred)

• Ab initio
  – Robetta (also comparative; intermediate)
    • [http://robetta.bakerlab.org](http://robetta.bakerlab.org)
A Modelling Example

TMX3

A Novel Thioredoxin-Like Protein of the Endoplasmic Reticulum
Background

- Disulfide bond formation
  - Mostly by protein disulfide isomerase (PDI) in the ER
  - Also by ERp57 (a PDI homologue), calreticulin and calnexin
  - Many other PDI relatives with unknown function

Disulfide Metabolism

Mixed disulfide intermediate
Discovery of TMX3

- Search for Trx-like proteins
- Found several conserved proteins
- TMX3
  - Conserved in all higher eukaryotes
  - Signal peptide
  - Trx domain
  - Transmembrane helix
  - Glycosylation sites
  - ER retention signal

Overview of TMX3

[Diagram showing the structure of TMX3, including ER signal sequence, thioredoxin-like domain, and predicted transmembrane region.]
Experimental Data – I

- Expression in most tissues – highest in muscle tissues.
- Inside ER
- Transmembrane protein
- Glycosylated

Experimental Data – II

- CD spectroscopy
  - Secondary structure content
- Fluorescence spectroscopy
  - Functional implications
Experimental Data – III

- Redox titration
- Redox potential: -157 mV
- Similar to PDI
- Similar function to PDI?

Building a Model

- Alignments showed similarity to
  - PDI domains
    - 40% sequence ID to a domain
  - ERp57 domains
  - Calsequestrin
    - An acidic Trx-domain protein involved in calcium storage in muscle cells.
    - 25% sequence ID to Trx2+3
Modelling Alignment

- Model built with MODELLER

Model vs. Template
ProQ

ProQ - Results
Prediction using secondary structure

Predicted LGscore : 4.109
Predicted MaxSub : 0.487

Different ranges of quality:
LGscore>1.5 fairly good model
LGscore>2.5 very good model
LGscore>4 extremely good model

MaxSub>0.1 fairly good model
MaxSub>0.5 very good model
MaxSub>0.8 extremely good model

http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi

Surface Potential

• CSQ is an acidic protein

• TMX3 is not...

• ...and is probably not regulated by calcium ions.
The Catalytic Domain

A Hydrophobic Pocket?
Missing Parts

• Transmembrane helix

![Predicted transmembrane region](image)

• Dimerisation?

![PDI](image)

---

Levamisole

• Unc-74 and C. elegans
• Cysteine-loop ligand-gated ion channels

![Coleopteran](image)

![Levamisole](image)
Conclusions

• Suggestions from model & experiments
  – TMX3 is a thioredoxin-like molecule of the ER
  – TMX3 is responsible for formation of cysteine-loop-gated ion channels involved in voluntary muscle contraction.

• Future experiments
  – Structure(!)
  – Test suggestions from model
    • Hydrophobic pocket/target?
  – Identify interaction partner(s)

Acknowledgements

• Johannes Haugstetter (ETH, Zürich)
• Lars Ellgaard (University of Copenhagen)

• Further reading:
  Haugstetter J, Blicher T, Ellgaard.L.
  Identification and characterization of a novel thioredoxin-related transmembrane protein of the endoplasmic reticulum.

  Haugstetter J, Maurer MA, Blicher T, Pagac M, Wider G, Ellgaard.L.
  Structure-function analysis of the endoplasmic reticulum oxidoreductase TMX3 reveals interdomain stabilization of the N-terminal redox-active domain.