Comparative modelling

Gonzalo Lopez
Iakes Ezkurdia

Structural Computational Biology Group
Spanish National Cancer Research Centre
(CNIO)
Why Do We Need Homology Modelling?

Detailed information about the 3D structure of a protein can help understand the molecular basis for biological function.

Proteins are important therapeutic targets and 3D structure is important in rational drug design for therapeutic applications.
Why Do We Need Homology Modelling?

The rate of deposition of sequences in the databases has far outstripped the rate of deposition of structures for many years. With the many sequencing projects, this difference can only grow.

Modelling can be a way of extending the set of known 3D structures since experimental techniques for determining protein structure are relatively slow and expensive.
Structural Conservation is the Key

Early and recent comparisons of the structures of homologous proteins show that 3D structure is rarely affected by small changes in protein sequence.

For example, these two structures have only 20% sequence identity.
The First Homology Model

In 1969, bovine lactalbumin was modelled on the X-ray crystallographic structure of hen egg-white lysozyme. **Sequence identity is only 39%, but there were no loops, so the structure did not vary so much.**

The prediction was proven generally correct when the structure of lactalbumin was solved in 1989.
What Can Be Modelled?

The fact that protein 3D structures have been highly conserved during evolution allows us to build good quality theoretical models based on structural templates.

In the absence of experimental data, generating models based on known structures is the only reliable way to obtain 3D structural information.

Generally, the higher the similarity between target and template, the more accurate the target model is likely to be (less loops, fewer side chain replacements, fewer conformational changes).

If identity is high and alignments are good, comparative modelling programs can generate models with a root mean square (rms) CA error lower than 2Å. Where similarity is more distant and alignments are poorer, this can be much higher.
High- and medium-accuracy comparative models are helpful in refining functional predictions previously based on sequence alone.

Ligand binding is directly determined by the structure of the binding site.

It is also possible to correctly predict features of the target protein that do not occur in the template structure ...

Baker and Sali, Science 2001
In A, it was possible to predict the size of the ligand from the volume of the binding site cleft - the complex between the ligand docosahexaenoic fatty acid and brain lipid-binding protein was modelled on PDB structure 1ADL (62% sequence).

In B a binding site for a charged ligand was predicted for mouse mast cell protease 7 based on a cluster of charged residues on the protein.

Baker and Sali, Science 2001
The Modelling Process

Homology modelling is a 4 steps process.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
**Template Identification**

A database search program such as FASTA or BLAST is usually sufficient to detect structural templates.

Using more than one template is advantageous.

Model quality will depend on the structural and sequence similarity of the template(s).

If the sequence identity between the target sequence and the nearest template falls below 25 or 30%, homology models are less likely to be successful.
### Alignments

<table>
<thead>
<tr>
<th>Accession</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1aac</td>
<td>DKATIPSEPFAAAEVADGAIVVDIAMKYETPELHVKGVDTVTWINREAMPHNVHFVAGV</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1plc</td>
<td>IDVLLGADDGLAFVSEFSISPGKFKNAGFPHNIVFDENDS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1aac</td>
<td>L--GEAAALKGPMMKKE------QAYSLTFTEAGTYDYHCTPHF--MRGKVVVE</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1plc</td>
<td>IPGVDASKISMSEEDLLNAKGETFEVALSNKGEYSFYCSSHPQGAGMVGVKVTN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*If an accurate 3D model is to be built, it is vital that the target-template alignments are correct. Particularly at lower % identity the biggest errors stem from the alignments.*

*Most comparative modellers update alignments manually using actual and predicted secondary structure and accessibility information, and careful placement of gaps.*

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Dealing with large gaps

Not all comparative modelling programs react the same way to errors.

A – SWISS-MODEL Model
B – Backbone Trace
C – Modeller model
D – Native Structure

Target-Template alignment with large gap below.

From Wallner and Elofsson, 2005

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Backbone Building

Rigid Body Assembly
Template structures are superposed and a “framework” is calculated from the average of the co-ordinates. The core regions are derived directly from this framework. Loops and side chains are added from libraries and the equivalent conformations in the template structures.

Segment Matching
Models are built by assembling short, all-atom segments that fit “guiding” positions from template structures. The guiding positions are usually conserved in the target-template alignment.

Satisfaction of Spatial Restraints
Target structure restraints are calculated from the aligned structures. These specific restraints are supplemented by stereochemical restraints on bond lengths, bond angles, hydrogen bonds, etc. The model is then built by minimising the violations of all the restraints.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Large insertions in the backbone are difficult to model because there is no template to model from.

Although in theory there are many ways to model a loop, there are some constraints, for example end points have to match and stereochemical rules have to be followed.

Ab-initio modelling of loops is still far from easy.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Loop and Gap Strategies

**Gap filling**
The backbone of the model may contain small gaps (3 residues or less). These are relatively easy to fix since the size of the gap makes only a few configurations possible.

**Canonical loops**
Canonical loops are commonly occurring loops and they can be used by accessing a library of canonical loops.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Prolines

Proline modelling is another difficulty. Existing backbone torsion angles usually do not favour substitution by proline and this causes the backbone to bend.

Gly → Pro is often the worst mutation since glycine can have most conformations without restrictions.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Side Chain Generation

Side chain placement is potentially the most difficult step, since side chains might be in more or less any conformation.

In practice the distribution of possible side chain conformations shows that there are only a few populous side chain conformations, or rotamers.

In practise most side chains tend to be in one of the two most populous rotamers, thus reducing the possible conformations.

Side chains can be replaced by eye, using molecular graphics packages and information from rotamer databases.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Rotamer Chicken and Egg Problem

Side chain conformations are predicted by taking into account similar structures (conserved residues usually have conserved rotamers), steric clashes and packing energies.

If many side chains have to be placed onto the backbone, side chains can potentially occupy the same space. This leads to a "chicken and egg" problem.

In order to place the first residue correctly, all other residues must already be correctly positioned.

There are certain rules for choosing the best rotamer that make correct rotamer prediction possible.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model

Tyr and Phe
The Problem with Side Chain Prediction

There are two problems that make perfect side chain prediction impossible.

Side chains that are replaced are often distorted in the model in order to compensate for the rigidity of the modelled backbone.

When template structures have less than 50% identity to the target, backbone shifts will hamper side chain replacement methods.

Ideally you need a method that includes backbone flexibility in its modelling of side chains, but as yet there has been no successful method.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
All homology models will have errors. So, without an estimation of the likelihood and magnitude of errors, the model itself is fairly meaningless.

Side chains or whole loops can be misplaced. Some errors may be more crucial than others.

Though error evaluation methods can indicate where potential errors are, error correction is still not possible. However, it is possible to model the structure again to take account of the errors.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Evaluation Algorithms

Normality indices can indicate how well model characteristics resemble actual known structural characteristics.

Stereochemical features such as bond lengths, bond angles, main chain and side chain torsion angles and clashes can be evaluated.

Programs for evaluating stereochemistry include PROCHECK and WHATCHECK.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Energy minimisation and molecular dynamics can also be used to evaluate models. Models should have low energy according to a molecular mechanics force field, such as CHARMM 22.

Spatial features such as packing, the formation of a hydrophobic core and solvent accessibilities that might point to errors can also be assessed. These methods assess the environment of each residue in a model against the expected environment.

Programs implementing this statistical profile approach include VERIFY 3D and PROSA.
Most comparative model building algorithms start from the principle that the backbone of the model and structure are identical.

In practice, however, conformational changes brought about by domain movement and bending of the backbone are fairly common.

But still the most damaging errors are usually caused by incorrect alignments.

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model

2bb2
1amm
Evaluating Protein Structure Predictions

CASP, EVA & LiveBench

Sadowski M I, Jones D T, Proteins. 2007. (online)
Homology Modelling Servers

An automated comparative modelling server (ExPASy, CH)

**CPHmodels** – www.cbs.dtu.dk/services/CPHmodels/
Server using homology modelling (BioCentrum, Denmark)

**SDSC1** – cl.sdsc.edu/hm.html
Protein structure homology modelling server (San Diego, USA)

**WHATIF** – www.cmbi.kun.nl/gv/servers/WHATIF/
WHAT IF Web interface: homology modelling, drug docking, electrostatics calculations, structure validation and visualisation.

**3D-JIGSAW** – www.bmm.icnet.uk/servers/3djigsaw/
Automated system for 3D models for proteins (Cancer Research UK)

**ROBETTA** – robeta.bakerlab.org/
ROBETTA detects fragments with BLAST, FFA S03, or 3D Jury, and generates alignments with K*SYNC. Loop regions are assembled from fragments and side chains using a replace and score method.

**SPARKS2** – phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sparks2.html
SWISS MODEL - an example

The SWISS MODEL server can be used to create 3D models of protein structure starting with just a linear sequence of amino acids.

Here we will follow the output of the server step by step as it creates an approximation of the 3D model.
First Approach Mode

First go from the front page to the "First Approach Mode"
(not the most obvious entry point ...)

Enter the query sequence, eg:

>target sequence
M SK V PRN FRL L EEL EK GEK GFGPESC SY GL A D S D D I T M T K W N G T I L G PPH S N H E N
L L D L R K E M A T P A N K K L R Q P K E G E T F

Along with the sequence the SWISS MODEL server requires you to introduce a return email address.
**Step 1**

**SWISS MODEL** first searches for homologous sequences in a database of sequences of known structure (ExNRL -3d) using the sequence search program **BLASTP2**.

Results from the search are shown in order of E-value:

- Length of target sequence: 137 residues
- Searching sequences of known 3D structures
  - Found 12UCE.pdb with P(N)=5.0e-12
  - Found 11AYZ.pdb with P(N)=1.2e-08
  - Found 11AAK.pdb with P(N)=5.3e-08
  - Found 12AAK.pdb with P(N)=5.3e-08
  - Found 11UCZ.pdb with P(N)=1.1e-07
  - Found 12UCZ.pdb with P(N)=1.1e-07
  - Found 11A3S.pdb with P(N)=1.0e-05

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Step 2

The template sequences found in step 1 are extracted from the database. AlignMaster creates pairwise alignments between the templates and the target sequence.

The program SIM selects all those structural templates that have an identity of 25% or greater with the target sequence. The length of the aligned region also must be at least 20 residues.

The program will also detect domains. If the target sequence has more than one domain the structure can also be modelled domain by domain basing each domain on a different structural template if necessary.
Sequence identity of templates with target:

12UCE.pdb: 22.85 % identity
11AYZ.pdb: 27.8 % identity
12AAK.pdb: 17.15 % identity
11UCZ.pdb: 29.6 % identity
12UCZ.pdb: 29.6 % identity
11A3S.pdb: 22.5 % identity

Looking for template groups
Global alignment overview:

Target Sequence: |====================================================================|
12UCE.pdb        |                  -------------------------------------------------|
11AYZ.pdb        |                    ------------------------------------------------|
12AAK.pdb        |               -----------------------------------------------------|
11UCZ.pdb        |                 --------------------------                      |
12UCZ.pdb        |                 --------------------------                      |
11A3S.pdb        |                 --------------------------------------------------|

AlignMaster found 1 regions to model separately:
1: Using template(s)  11A3S.pdb 11AAK.pdb 11AYZ.pdb 11UCZ.pdb 12AAK.pdb 12UCE.pdb 12UCZ.pdb

12UCE.pdb has been rejected, too low similarity with Target sequence (22.85 % identity.)
12AAK.pdb has been rejected, too low similarity with Target sequence (17.15 % identity.)
11A3S.pdb has been rejected, too low similarity with Target sequence (22.5 % identity.)
ProMod II

Requirements

A. At least one protein of known 3D structure.
B. Good quality alignments (model reliability is related to sequence identity)

Procedures

1. Superposes related 3D structure(s).
2. Generates a multiple alignment.
3. Generates a framework for the model.
4. Rebuilds missing loops.
5. Completes and corrects backbone.
6. Corrects and rebuilds side chains.
Example ProModII Output

ProModII trace log for Batch.1
============================================================
ProModII: Loading Template: 11AYZ.pdb
ProModII: Loading Template: 11UCZ.pdb
ProModII: Loading Template: 12UCZ.pdb
ProModII: Loading Raw Sequence
ProModII: Iterative Template Fitting
ProModII: Iterative Template Fitting
ProModII: Generating Structural Alignment
ProModII: Aligning Raw Sequence
ProModII: Refining Raw Sequence Alignment
ProModII: Weighting Backbones
ProModII: Averaging Side chains
ProModII: Adding Missing Side chains
ProModII: Small Ligation (C-N < 3.0A) ignored;
ProModII: GROMOS will repair it at residue ASP 73
ProModII: Building CSP loop with anchor residues THR 60 and GLU 63
ProModII: Number of Ligations found: 1
ProModII: all loops are bad; continuing CSP with larger segment
ProModII: Building CSP loop with anchor residues PRO 59 and GLU 63
ProModII: Number of Ligations found: 10
ProModII: ACCEPTING loop 5: clash= 1 FF= 53.8 PP=-18.41
ProModII: Dumping Preliminary Model
ProModII: Dumping Sequence Alignment
ProModII: Done.
In the last step the program Gromos96 uses energy minimisation to optimise the model.

Gromos96 trace log for Batch.1
==================================================================
Now running PROCS1 on file batch-procs0.dat ... Done.
Now running PROCS2 on file batch-procs1.dat ... Done.
Now running PROGCH on file batch-procs2.dat ... Done.
Now running PROMD on file batch-progch.dat ... Done.
Now running PROMD on file batch-promd0.dat ... Done.
Detection of SS-Bonds within batch ...

1. Identify structural template(s)
2. Align templates for modelling
3. Build the model
4. Evaluate the model
Model 3D Co-ordinates

The SWISS MODEL server sends out the co-ordinates of the model based on the co-ordinates of the template chains.

It is important to remember that these co-ordinates are only predicted and are not as reliable as experimental co-ordinates.

In fact the more remote the homologue the less accurate the model co-ordinates.

An additional evaluation of the model quality is highly recommended (WHATCHEK, VERIFY3D, ...).

These evaluations can suggest small alignment modifications and/or alternative template selection for modelling specific regions of the protein (requires certain expertise).
Swissmodel Exercise

http://swissmodel.expasy.org/
Go from the front page to the "First Approach Mode"

>Sequence1
GSHMQSYQISSRLMAQEGRSTSVQTSSLIQSLFDLALRIHQQD
STAKNASLINALVSRDSSLDEFFSSVDELELSNA
PDLRFISSIONILWDDGNASFYGIAQQLNKLIRRAISGNWHLV
QTPSEGKSVHIMRSSLIEAGTGQVVGLYVGVIV
LNDNFALLLENIRSGSNSENLVVLADVPTTPLVSTLKGNEPYSLDYVV
HSAKADMRSFIVGQTFLIVEVESVPTILCVYSIQTN
QNVLTLDNESVPTILCVYSIQTNQNVRHQ
Acknowledgements

Michael L. Tress
Ana Rojas

Structural Computational Biology Group and anyone else I borrowed figures from.