Protein 3D Structure Prediction

Michael Tress

CNIO
Given a new sequence that we want to know as much as possible about, one of the fundamental questions is "how does it fold?"
3D Structure Prediction

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.
Why Such a Gap?

Sequencing millions of DNA sequences is relatively easy, while experimental determination of a single protein is difficult = 1-3 years.
Fortunately, many proteins with apparently dissimilar sequences have similar folds. This means we can use databases of known 3D structures to make predictions for unknown sequences.

Since experimental techniques for determining protein structure are relatively slow and expensive, modelling is a way of extending the set of known 3D structures.
Protein Structure is Highly Conserved

Early 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.
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.
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 ...

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
Modelling Terminology

The sequence of unknown structure is called the **query** or **target** sequence.

In order to build a 3D **model** by comparative modelling two things are necessary:

1. A protein with known structure (a **template**).
2. An **alignment** between the target and template sequences (the quality of the model will depend on the quality of the alignment between the target and template).
Four Basic Comparative Modelling Steps

Template Detection
The first step is to find a sufficiently similar structural template(s) from the PDB, by sequence search or by structure-based techniques (hybrid, fold recognition etc.).

Alignment
In fact this is not a separate step - all template detection methods create alignments. But automatic alignments can be edited.

Model Building
From the simple transference of PDB coordinates to complex all atom modelling.

Evaluation
Most methods assess model quality, this is based on whether the model 3D structure is protein-like.
Template Identification

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

Domains
Many proteins are made up of several structural domains. A domain search should be carried out at the same time as the sequence search.

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.
All methods of template detection generate alignments between the query sequence and the PDB template sequence.

If an accurate 3D model is to be built, it is vital that the target-template alignments are correct.

Generally the higher the sequence similarity and the lower the number of gaps between the two sequences, the more likely the alignment is to be correct. Particularly at lower percentage identity the biggest errors stem from the alignments.
Alignments

The more sequences that are included in the alignment the more likely the alignment is to be **reliable** in an evolutionary sense.

For that reason most people generate alignments with **multiple sequence alignment** programs or profile-based methods.

Since the alignments are for the purposes of generating structures, it is important to take into account structural information such as the coincidence of real and predicted **secondary structure** and **accessibility**.

Alignments can be edited manually using actual and predicted secondary structure and accessibility information, and careful placement of **gaps**.
"Alignment 1" is chosen because of the PROs at position 7. But the 10 Angstrom gap that results is too big to close with a single peptide bond.
Rigid Body Assembly
Template structures are superposed and a “framework” is calculated from the average of the co-ordinates.

The model is derived directly from this framework in the core regions and 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 short segments are found by scanning the structural database and the guiding positions come from 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.
Loop and gap strategies

Large insertions in the backbone are difficult to model because there is no template to model from. Ab-initio modelling of loops is still far from easy.

There are some constraints, for example end points have to match and stereochemical rules have to be followed.

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.
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.
Side Chain Generation

Side chain placement usually comes after the backbone modelling step. It is problematic, since side chains might be in more or less any conformation.

In practice there are only a few populous side chain conformations, or rotamers and most side chains tend to be in one of the two most populous rotamers, thus reducing the possible conformations.

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.
Negotiating the Chicken and Egg Problem

However, there are certain rules for choosing the best rotamer that make correct rotamer prediction possible.

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

(Well, almost …)

Perfect side chain prediction is impossible.

Side chains rotamer angles have to be 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.

An ideal method would include backbone flexibility in its modelling of side chains, but no successful method does that yet.
All homology models will have errors. 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.
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.
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 CHARMM22.

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 VERIFY3D and PROSA.
## Homology Modelling Servers

<table>
<thead>
<tr>
<th>Server</th>
<th>Website</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPHmodels</td>
<td><a href="http://www.cbs.dtu.dk/services/CPHmodels/">www.cbs.dtu.dk/services/CPHmodels/</a></td>
<td>Server using homology modelling (BioCentrum, Denmark)</td>
</tr>
<tr>
<td>SDSC1</td>
<td><a href="http://cl.sdsc.edu/hm.html">cl.sdsc.edu/hm.html</a></td>
<td>Protein structure homology modelling server (San Diego, USA)</td>
</tr>
<tr>
<td>3D-JIGSAW</td>
<td><a href="http://www.bmm.icnet.uk/servers/3djigsaw/">www.bmm.icnet.uk/servers/3djigsaw/</a></td>
<td>Automated system for 3D models for proteins (Cancer Research UK)</td>
</tr>
<tr>
<td>ROBETTA</td>
<td><a href="http://robetta.bakerlab.org/">robetta.bakerlab.org/</a></td>
<td>ROBETTA detects fragments with BLAST, FFAS03, or 3DJury, and generates alignments with K*SYNC. Loop regions are assembled from fragments and side chains using a replace and score method.</td>
</tr>
</tbody>
</table>
Comparison of Homology Modelling Servers by EVA
What Can Be Modelled?

Since protein 3D structures have been highly conserved during evolution we can build good quality theoretical models based on structural templates.

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.
What to expect from a model

Although homology modelling programs can produce good models of the backbone of the protein, they are less good at building loop regions and replacing side chains.

Because model building algorithms start from the principle that the backbone of the template and target structure are identical, models built by comparative modelling will always resemble the template more than the target.

Unfortunately conformational changes brought about by domain movement and bending of the backbone are fairly common.

Despite this important information can still be obtained from the models …
Without experimental data generating models based on known structures is the only reliable way to obtain 3D structural information.
Simplified Protein Structure Prediction Flow Chart

1. EXPERIMENTAL SEQUENCE → DATABASE SEARCHING → STRUCTURAL HOMOLOGUE?
2. YES → HOMOLOGY MODELLING
3. NO → SECONDARY STRUCTURE PREDICTION
   - FOLD PREDICTION HYBRID METHODS
4. Final Structure???
No Template Found by BLAST?

Pairwise sequence search methods can detect folds when sequence similarity is high, but are very poor at detecting relationships that have less than 20% identity.

One possibility is to use profile-based sequence search methods (PSIBLAST, FFAS, HMMs). These have evolved greatly, and can find templates with very low sequence similarity.

Hybrid methods and fold recognition methods can find folds that are too distantly related to be detected by sequence based methods, because they evaluate not only sequence similarity, but also structural fit.
How fold recognition can help

Small changes in sequence have little effect on backbone structure

The development of fold recognition methods came from the observation that many apparently unrelated sequences had very similar 3-dimensional structures (folds).
Profile-Based and Hybrid Techniques

PSI-BLAST
Profile method, can be as accurate as many fold recognition techniques at detecting remote homologues. It is possible to spot biologically meaningful templates from careful analysis of low-scoring hits.

Intermediate search methods
Profile-profile alignment methods use evolutionary information in both query and template sequences. As a result, they are able to detect homologies beyond the reach of other sequence-based methods.

Hidden Markov Models
Hidden Markov models regard the sequence as a series of nodes, each corresponding to a column in a multiple alignment. Each node has a residue “state” and states for insertion and deletion. The HMMs built from aligned sequences have many similarities to profiles.

META-PROFILES (Hybrid methods)
Many profile methods now also use predicted secondary structure. By adding structural information to the profiles (meta-profiles) it is often possible to find homologues that have very low sequence similarity but are still structurally similar.
Profile/Hybrid Sequence-Based Servers

PSI-BLAST ...

SAM T02 - www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html
The query is checked against a library of hidden Markov models. This is NOT a threading technique, it is sequence based, but it does use secondary structure information.

Meta-BASIC - basic.bioinfo.pl
Meta-BASIC is based on consensus alignments of profiles. It combines sequence profiles with predicted secondary structure and uses several scoring systems and alignment algorithms.

FFAS – ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl
FFAS03 is a profile-profile alignment method, which takes advantage of the evolutionary information in both query and template sequences.
Fold Recognition – prediction from sequence-structure alignments

Start with the Query sequence:

MTYGFRIPLNCERWGHKLISTVILKRPMTYGFRIPLNCERWP...

Align the sequence with all folds from the library.

Evaluate the sequence-structure fit
Fold Recognition Algorithms: General Principle

It was thought when fold recognition methods were developed that they could detect analogues, proteins that were structurally similar but that had no evolutionary relationship.

In fact most of these predictions were later shown to be homologous (have an evolutionary relationship) by advanced sequence comparison methods, such as PSI-BLAST.

They still have a place though, in part because the newer fold recognition methods are more sensitive than PSIBLAST, in part because research also shows that no one method can always hope to correctly identify a fold.
Library of core folds (structural templates)

- all known structures
- representative subset (seq. similarity filters)
- structural cores with loops removed

Score functions to evaluate the sequence-structure alignments:

- Coincidence of the observed and expected \textit{structural environments} of each residue
- Pair potentials
- Solvation energy
- Coincidence of real and predicted secondary structure and accessibility
- Evolutionary information (from aligned structures and sequences)
Bowie et al. (1991) created a fold recognition approach that described each position of a fold template as being in **one of eighteen environments**.

These environments were defined by measuring the **area of side chain** that was buried, the fraction of the side chain area that was **exposed to polar atoms**, and the **local secondary structure**.

Other researchers have developed similar methods, where the structural environments described include exposed atomic areas and type of **residue-residue contacts**.
How Structural Environments are Used

Scoring matrices are pre-generated for the probabilities of finding each of the twenty amino acids in each of the environment classes. Probabilities are drawn from databases of known structures.

Using these probabilities a 3D profile is created for each fold in the fold library. This 3D matrix defines the probability of finding a certain amino acid in a certain position in each fold.

When the target sequence is aligned with the fold a score is calculated from the pre-generated 3D profile for each of the positions in the alignment. The fit of a fold is the sum of the probabilities of each residue being found in each environment.
Solvation potential is a term used to describe the preference of an amino acid for a specific level of residue burial. It is derived by comparing the frequency of occurrence of each amino acid at a specific degree of residue burial to the frequency of occurrence of all other amino acid types with this degree of burial.

The degree of burial of a residue is defined as the ratio between its solvent accessible surface area and its overall surface area.
Pair or contact potentials - the tendency of residues to be in contact

Counts become propensities (frequency at each distance separation) or energies (Boltzmann principle, -KT ln)

Make count of interacting pairs of each residue type at different distance separations
3DPSSM Secondary Structure and Solvation Potentials

In the server 3DPSSM each residue in each fold is assigned a solvation potential. The degree of burial of each residue is defined as the ratio between its solvent accessible surface area and its overall surface area.

Solvation potential is divided into 21 bins, ranging from 0% (buried) to 100% (exposed).

Secondary structure type is also assigned to each fold based on the annotation in the STRIDE database.
Fold Recognition Servers

3D-PSSM - www.sbg.bio.ic.ac.uk/~3dpssm/
Based on sequence profiles, solvation potentials and secondary structure.

SPARKS2 - http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sparks2.html

mGenTHREADER - www.psipred.net/
Combines profiles and sequence-structure alignments. A neural network-based jury system calculates the final score based on solvation and pair potentials.
Consensus Fold Recognition

It has long been recognised that human experts are better at fold prediction than the methods these same experts had developed. Human experts usually use several different fold recognition methods and predict folds after evaluating all the results (not just the top hits) from a range of methods.

An algorithm that mimics the experts

In first consensus server, Pcons, the target sequence was sent to six publicly available fold recognition web servers.

Models were built from all the predictions. The models were superimposed and evaluated for their similarity.

The quality of the model was predicted from the rescaled score and from its similarity to other predicted models.
Consensus Fold Recognition Servers

**3D Jury** - http://bioinfo.pl/meta/
3D Jury is a consensus predictor that utilizes the results of fold recognition servers, such as FFAS, 3D-PSSM, FUGUE and mGenTHREADER, and uses a jury system to select structures.

Another meta-server using fold recognition, profile-based servers and secondary structure prediction to aid consensus prediction.

**Pcons** - www.sbc.su.se/~arne/pcons/
Pcons was the first consensus server for fold recognition. It selects the best prediction from several servers. PMOD can also generate models using the alignment, template and MODELLER.