Structural Biochemistry (KBB055)

Lecture on Structure Prediction: Graham Kemp

http://www.cs.chalmers.se/~kemp/teaching/KBB055/

Aims

To revise the basic principles of protein conformation.

To explain the steps involved in comparative protein modelling.

To introduce the ways in which the stereochemical quality of a protein structure can be assessed.

To introduce the concepts of fold recognition and secondary structure prediction.

Comparative modelling strategy

- identify a known structure that is predicted to be similar;
- align sequences;
- predict structurally conserved regions, and locations of insertions and deletions (sometimes called “indels”);
- build model backbone structure
  - copy predicted conserved main chain regions from template structure,
  - remodel loops with insertions or deletions;
- add side chains to the modelled main chain;
- evaluate and refine model.

Why build model structures?

Knowledge of a protein’s three-dimensional structure is vital to a full understanding of the molecular basis for its biological function.

We want to understand the function of all proteins encoded by a genome, therefore we would like to know all of their 3-D structures.

Experimental techniques for determining protein structure are relatively slow and expensive, so we look to modelling as a way of extending the set of 3-D structures.

Modelling can also be used in protein engineering when designing proteins for therapeutic applications.

Using known substructures in protein crystallography

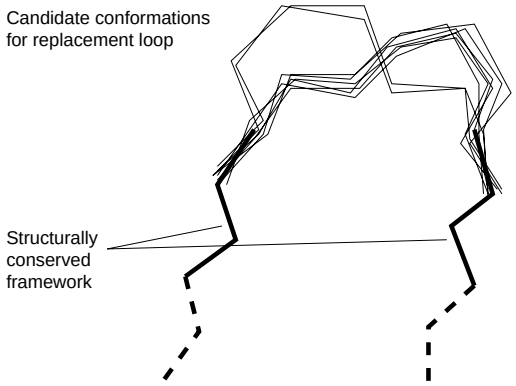
Jones, T.A. and Thirup, S. (1986)
The EMBO Journal, vol. 5, pp 819-822.

Electron density map interpretation is made easier by fitting regular α-helices and strands into the map.

This building-block approach to protein modelling can be extended to include **all** main chain fragments.

For example, a model of retinol binding protein was built using fragments from only three other proteins. A model with Cα atoms matching within an R.M.S. error of 1Å was built using only 15 fragments.

Fragment-fitting: an approach to remodelling loops



Side chain rotamers

There is an extremely large number of possible combinations of side chain conformations — infinite if we consider side-chain bonds to be continuously variable.

For practical purposes the search space can be discretised by considering a finite set of possible torsion angles for each side-chain.

The distribution of side chain conformations falls into statistically significant clusters. By using representative side chain conformations, or **rotamers**, the vast combinatorial search space can be greatly reduced.

Ponder, J.W. and Richards, F.M. (1987)
J. Mol. Biol., vol. 193, pp 775-791.

Fragment selection criteria

- steric overlap;
- packing
  - — no protruding loops;
  - — no internal cavities;
- disulphide bridges and salt bridges;
- solvent accessibility
  - — avoid burying unpaired charges;
- sequence criteria
  - — Gly and Pro residues
  - — similarity between model's sequence and the sequences of the fragments in their native structures.

PROCHECK

Roman Laskowski et al.

http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

"The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.

Unusual regions highlighted by PROCHECK are not necessarily errors as such, but may be unusual features for which there is a reasonable explanation (eg distortions due to ligand-binding in the protein's active site). Nevertheless they are regions that should be checked carefully."

Checks performed by PROCHECK (1)

Covalent geometry checks

- Main-chain bond lengths:
  Compared against the Engh & Huber small-molecule data.
  Differences greater than 0.05Å are highlighted.
- Main-chain bond angles:
  Compared against the Engh & Huber small-molecule data.
  Differences greater than 10 degrees are highlighted.

Planarity checks

- Aromatic rings (Phe, Tyr, Trp and His)
  RMS distances greater than 0.04Å from best-fit plane highlighted.
- End-groups (Arg, Asn, Asp, Gln, Glu)
  RMS differences greater than 0.03Å from best-fit plane are highlighted.

Checks performed by PROCHECK (2)

Dihedral angle checks

Ramachandran plot shows phi-psi distribution.

Each residue is classified according to its region: “core”, “allowed”, “generous”, or “disallowed”.

Residues in the generous and disallowed regions are highlighted on the plot.

A log-odds score shows how normal or unusual the residue’s location is on the Ramachandran plot for the given residue type.

Chirality check

Provides a measure of the C-alpha tetrahedral distortion. Measured by the notional zeta torsion angle, defined by the atoms C-alpha, N, C and C-beta.

The expected value is 33.9 degrees.

A negative value signifies a D-amino acid.

Checks performed by PROCHECK (3)

Non-bonded interactions check

Any two non-bonded atoms are deemed to make a bad contact if they are as close as 2.6Å apart.

Possible hydrogen-bonding partners are excluded by ignoring all atom-pairs where one of the atoms is a possible H-bond donor (eg a main-chain nitrogen) and the other is a possible H-bond acceptor (eg a water molecule, or a main-chain oxygen).

Main-chain hydrogen bonds check

A check is made of main-chain hydrogen-bond energies, calculated using the Kabsch & Sander (1983) method.

Significant deviations from the ideal value of -2.0 kcal/mol are highlighted.

Checks performed by PROCHECK (4)

Disulphide bond checks

The S-S separation in each disulphide bond is compared with the ideal distance of 2.0Å

The chi-3 torsional angle, defined by the S-S bridge, is compared against the ideal values:

-85.8 degrees for a left-handed conformation
96.8 degrees for a right-handed conformation

Significant deviations from the ideal values are highlighted.

Various other stereochemical parameters are computed and compared with values from well-refined structures.

Fold recognition

The idea behind “threading”:

Imagine a wire wound into the shape of a known protein’s main chain “fold”.

Imagine next that our new sequence is represented by beads that are “threaded”, in order, onto the wire, and are pushed along the wire.

At each step, a score is calculated based on which residues are adjacent in space, which residues are buried, etc.

Repeat this process for each different known fold.

A high score indicates that the sequence is compatible with that fold.

Heuristics for manual secondary structure prediction

- Many α-helices are amphipathic. Conserved hydrophobic residues at positions i, i+3, i+4, i+7, etc. are highly indicative of an α-helix.
- Half-buried strands will tend to have hydrophobic and hydrophilic residues at alternate positions.
- In proteins containing both α-helices and strands the strands are often completely buried and tend to contain only hydrophobic residues.

For more details and references, see:
http://www.bmm.icnet.uk/people/rob/CCP11BBS/secstrucpred.html

Secondary structure prediction

If neither sequence comparison nor fold recognition identifies a structure that can be used as a template for comparative modelling, then we can consider predicting secondary structure elements and how these might be assembled into a compact structure.

However, as noted by Ponder and Richards (1987):

“a major problem lies in the secondary structure prediction itself ... the problem appears to lie in the non-negligible effect of long-range tertiary structural features upon secondary structure”

and

“the problem of docking the preformed secondary units is formidable when considered in atomic detail.”

Alternative secondary structure assignment methods

Cuff J. A. and Barton G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, PROTEINS: Structure, Function and Genetics. 34:508-519 (1999)

“Secondary structure definition methods DSSP[38], DEFINE[39] and STRIDE[40] were compared. All three agree at only 75% of positions. This is mainly due to differences between DEFINE and DSSP/STRIDE. DSSP and STRIDE agree at 95% of positions, though DSSP defines many more 4 residue helices than STRIDE.”

[38] W. Kabsch and C. Sander. A dictionary of protein secondary structure. Biopolymers, 22:2577-2637, 1983.
[39] F. M. Richards and C. E. Kundrot. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins, 3:71-84, 1988.
[40] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. Proteins, 23:566-579, 1995.