Protein Structure Prediction

Protein Folding vs Structure Prediction

- Protein folding is concerned with the process of the protein taking its three dimensional shape.
- Protein structure prediction is solely concerned with the 3D structure of the protein.

Levinthal’s Paradox

Consider this greatly simplified view of protein folding for a protein of 100 amino acids:

- If each amino acid can adopt only 3 possible conformations, the total number of conformations is $3^{100} = 5 \times 10^{47}$.
- Assuming it would take $10^{(-13)}$ seconds to change each conformation, the time required to test all conformations would be $5 \times 10^{34}$ seconds, or $10^{27}$ years (the age of the universe is thought to be about $10^{10}$ years).
- Proteins usually fold in fractions of a second.

The Central Dogma

“The three dimensional structure of a protein is determined by its sequence and its environment without the obligatory role of extrinsic factors”.

Anfinsen (1973) – Renaturation of ribonuclease.

Flavors of Structure Prediction:

- Homology modeling.
- Fold recognition (also called ‘inverse folding’, ‘threading’, or ‘sequence-structure threading’).
- Ab initio (also called ‘de novo’ or ‘new folds methods’).

Predicting protein interaction (for example docking) also has to do with structure prediction, but is not considered in this lecture.
Secondary structure classification

Eight states from DSSP:

- H: α-helix
- G: 3_10-helix
- I: β-sheet
- E: β-strand
- B: bridge
- T: β-turn
- S: bend
- C: coil

CASP standard:
H = (H, B, I), E = (E, B), C = (C, T, S).

Secondary Structure Prediction

Given the sequence of amino acids of a protein, what is its secondary structure?

Primary structure:
GHWIATQRLIREAYEDYRHFSREGFIFIP

Secondary structure:
CREERECHHRHRHRHRHCCCHHCCCCC

Notation: H: Helix  E: Strand  C: Coil

A little bit of history…

The early methods for secondary structure prediction suffered from a lack of data, and usually were performed on single sequences.

- 1974: Chou and Fasman.
  - Propensities of formation based upon frequency of occurrence, rule based.
- 1974: Lim.
  - Theory based on chemical side-chain properties, very complex rules.
- 1978: Garnier, Osguthorpe, Robson.
  - Sliding window, consensus approach.

The prediction accuracy for all of these methods were roughly 50-55%.
Measures for Prediction Accuracy

The standard measure for prediction accuracy is (still) the Q3 measure

\[
Q3 = 100 \frac{\sum M_{ii}}{N_{obs}}
\]

where \(N_{obs}\) is the number of amino acids in the protein and \(M_{ii}\) is the number of matches for state \(i\) (C, E, H).

In recent years, the segment overlap measure (SOV) has been used more extensively. It aims for measuring how well secondary structure elements have been predicted rather than individual residues.


Automated Methods

The availability of large families of homologous sequences together with advances in computing techniques has pushed the prediction accuracy well above 70%. Most methods are available as web servers. They include:

- PHD
  http://www.embl-heidelberg.de/predictprotein/predictprotein.html
- PSI-PRED
  http://bioinf.cs.ucl.ac.uk/psipred/
- JPRED
  http://www.compbio.dundee.ac.uk/~www-jpred/

Other References


I-SITES
Main Effects in Protein Structure/Stability

Net protein stability:

- Hydrophobic effect.
- Atomic packing.
- Conformational entropy (rotamers).
- Electrostatic effects (ion pairs and hydrogen bonds).
- Disulfide bridges.

Ab Initio Methods

- Ab initio (latin): “from the beginning”.
- Assumption: the structure that a protein folds into inside cells is the structure with the lowest free energy.
- Finding native-like conformations require:
  - A scoring function (potential).
  - A search strategy.

Rosetta

- The scoring function is a model generated using various contributions. It has a sequence dependent part (including for example a term for hydrophobic burial), and a sequence independent part (including for example a term for strand packing).
- The search is carried out using simulated annealing. The move set is defined by a fragment library for each three and nine residue segment of the chain. The fragments are extracted from observed structures in the PDB.
Rosetta Scoring Function

\[ P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure}) \]

Sequence dependent:
- Hydrophobic burial
- Residue pair interaction

Sequence independent:
- Helix-helix packing
- Strand-strand packing
- sheet configurations
- vdWL interactions

Hydrophobic Burial

Residue Pair Interaction

Angles between Elements of Secondary Structure

Strand Packing – Helps!

Sheer Angles – Help not!
Picking the best structures

Rosetta’s scoring function is not perfect. Therefore, thousands of so-called decoys are generated. Among those, the best structures are picked using the following strategies:

- Cluster according to 3D structure.
- Filter poor sheet conformations.
- Add sidechains and remove poorly packed structures.
- Manual inspection, and more!