Lecture 21: Protein Structure Prediction I

- Introduction to protein sequence and structure
- Protein secondary structure prediction

Some slides adapted from notes created by Dr. Keith Dunker

Protein Structure: Overview

§ **Dogma:** amino acid sequence determines 3-dimensional protein structure

§ We should, in theory, be able to predict structure from sequence

§ Knowledge of protein structure will give information about protein function

FUNCTION

Difficulties in Protein Structure Prediction

§ Currently, calculating a protein structure from first principles (Quantum theory) is too difficult and complex a problem

§ Different sequences can yield the same structure

§ The same sequences can yield different structures

Levels of Protein Structure

Primary (1˚) structure: amino acid sequence of protein

Secondary (2˚) structure: local structure (alpha helices or beta strands)

Tertiary (3˚) structure: 3-dimensional structure of protein

Quaternary (4˚) structure: structure of a multiple protein complex
1° Structure: Amino Acid sequence

§ The amino acid sequence of a protein specifies its 3-dimensional structure

§ Each amino acid side-chain has different chemical properties that tend towards forming different types of structures

• Amino acids with large, nonpolar side-chains (L, W, etc) will tend to pack into the central hydrophobic regions of proteins

• Amino acids with polar or charged side-chains (E, K, etc) will tend to be on the hydrophilic surface of proteins

Properties of Amino Acid Side-Chains

Nonpolar, aliphatic side chains

Aromatic side chains

Polar, uncharged side chains

Special side chains

Polar, charged side chains

Amino acid structures from: http://en.wikipedia.org/wiki/Proteinogenic_amino_acid

Post-translational Modifications

§ Post-translational modifications alter the chemical properties of the amino acid side-chain, and thus can change protein structure or interactions

§ The addition or removal of post-translational modifications are catalyzed by enzymes

§ Examples of post-translational modifications:

• Phosphorylation
• Methylation
• Glycosylation
• Acetylation
• Ubiquitination
• Farnesylation

Peptide Bonds and Rotational Conformations

§ Peptide bonds between amino acids have a partial double bond character, limiting rotational conformations of the protein chain

§ The structure of the protein is determined by the Phi (\(\psi\)) and Psi (\(\psi\)) dihedral angles for each amino acid in the protein

Picture from http://www.tulane.edu/~biochem/med/second.htm
Rotational Conformation and Steric Interference

§ Rotational conformations are also limited by steric interference
§ Many $\phi$ and $\psi$ (and $\omega$) dihedral angles lead to steric hindrance between amino acid side chains and/or the peptide backbone

![Steric interference diagram](http://www.tulane.edu/~biochem/med/second.htm)

2° Structure: Overview

§ Secondary structure represents local structure and folding of segments of the overall protein that fall into the range of favorable $\phi$ and $\psi$ angles
§ Types of Secondary Structure:
  • Alpha helix ($\phi \sim -57$; $\psi \sim -47$)
  • Beta sheet/strand:
  • Turns
  • Coil

![Secondary structure types](http://commons.wikimedia.org/wiki/File:Beta-sheets.png)

Reasons for Predicting Secondary Structure

§ Secondary structure prediction is often an important starting point for predicting the tertiary (3D) structure of a protein
§ Since the secondary structure of a protein is strongly influenced by the local amino acid sequence, its prediction should be more straightforward than tertiary (3D) structure prediction
§ Secondary structure prediction can facilitate identifying potential structural motifs or domains in the protein sequences
§ Secondary structure prediction can give insight into potential protein function
§ Secondary structure prediction can guide the design of site-directed mutations in a protein sequence

Methods for Secondary Structure Prediction

§ Chou-Fasman method
  • Based on the propensities of different amino acids to adopt different secondary structures
  • Predictions are made using a rules-based approach to identify groups of amino acids with shared secondary structure propensities
§ Garnier, Osguthorpe, Robson (GOR) method
  • Statistical method of secondary structure prediction based on information theory & Bayesian probability
§ Multiple Sequence Alignment (MSA) methods
  • Performs secondary structure prediction on a multiple sequence alignment as opposed to a single protein sequence
§ Neural network-based methods
  • Example: Profile network from Heidelberg (PHD)
Chou-Fasman (1974) Method

§ Based on observation that certain amino acids tend to be enriched (or depleted) in different secondary structure classes

§ The Chou, Fasman and co-workers calculated the propensity of each amino acid to adopt an alpha helix, beta-strand, or coil conformation (later turn as well)

§ Propensity values (Pᵣ or Pₜ) were calculated using a database of experimentally determined protein structures (first database had 15 proteins)

§ Propensity for an amino acid of type i to form an α-helix (Pᵣᵢ) is calculated using the following formula:

\[ Pᵣᵢ = \frac{\text{fraction of residues of type } i \text{ in } \alpha \text{-helix}}{\text{fraction of all residues in } \alpha \text{-helix}} \]

Chou-Fasman Propensities (1978)

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>( Pᵣᵢ )</th>
<th>( Pₜᵢ )</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1.45</td>
<td>1.70</td>
</tr>
<tr>
<td>A</td>
<td>1.42</td>
<td>1.60</td>
</tr>
<tr>
<td>L</td>
<td>1.21</td>
<td>1.47</td>
</tr>
<tr>
<td>K</td>
<td>1.16</td>
<td>1.38</td>
</tr>
<tr>
<td>F</td>
<td>1.13</td>
<td>1.37</td>
</tr>
<tr>
<td>Q</td>
<td>1.11</td>
<td>1.30</td>
</tr>
<tr>
<td>W</td>
<td>1.08</td>
<td>1.19</td>
</tr>
<tr>
<td>I</td>
<td>1.08</td>
<td>1.10</td>
</tr>
<tr>
<td>V</td>
<td>1.06</td>
<td>1.05</td>
</tr>
<tr>
<td>D</td>
<td>1.01</td>
<td>0.93</td>
</tr>
<tr>
<td>H</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td>R</td>
<td>0.98</td>
<td>0.87</td>
</tr>
<tr>
<td>T</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>S</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td>C</td>
<td>0.70</td>
<td>0.75</td>
</tr>
<tr>
<td>Y</td>
<td>0.69</td>
<td>0.74</td>
</tr>
<tr>
<td>N</td>
<td>0.67</td>
<td>0.55</td>
</tr>
<tr>
<td>P</td>
<td>0.57</td>
<td>0.54</td>
</tr>
<tr>
<td>G</td>
<td>0.57</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Chou-Fasman Method

§ Example calculation for α-helix propensity (\( Pₛ \)) for alanine (A):

§ Data from protein structure database (29 proteins):
- Number of alanine residues in the database = 434
- Number of alanine that occur in alpha helix = 234
- Number of all residues in the database = 4741
- Number of all residues that occur in alpha helix = 1798

\[ P₄^A = \frac{234}{434} \times \frac{1798}{4741} = 0.539 \]

Adapted from notes by Mallika Veeramalai and David Gilbert
http://www.brc.dcs.gla.ac.uk/~drg/courses/bioinformatics_mscIT/slides/slides7/sld014.htm
Chou-Fasman Algorithm

1. Alpha Helix Prediction:
   A. Nucleate a helix by scanning for groups of 6 residues with at least 4 helix formers (Hₙ⁻ and Hₙ) and no more than 1 helix breaker (Bₙ and Bₙ). (regions with both α-helix and β-strand assignment)
   • Two Iᵣ residues count as one helix former for nucleating a helix
   B. Propagate predicted helix in both directions until reach a four residue window with average propensity (Pₐ) < 1.0
   C. The average propensity (Pₐ) for a predicted helix must be Pₐ > 1.03

2. Beta Strand Prediction:
   A. Nucleate a β-strand by scanning for groups of 5 residues with at least 3 strand formers (Hₜ and Hₜ) and no more than 1 strand breaker (Bₜ and Bₜ).
   B. Propagate predicted β-strand in both directions until reach a four residue window with average propensity (Pₜ) < 1.0
   C. The average propensity (Pₜ) for a predicted β-strand must be Pₜ > 1.05

Chou-Fasman Algorithm (simplified)

1. Alpha Helix Prediction:
   A. Nucleate a helix by scanning for groups of 6 residues with at least 4 helix formers (Hₙ⁻ and Hₙ) and no more than 1 helix breaker (Bₙ and Bₙ).
   • Two Iᵣ residues count as one helix former for nucleating a helix
   B. Propagate predicted helix in both directions until reach a four residue window with average propensity (Pₐ) < 1.0
   C. The average propensity (Pₐ) for a predicted helix must be Pₐ > 1.03 and Pₐ > Pₜ

2. Beta Strand Prediction:
   A. Nucleate a β-strand by scanning for groups of 5 residues with at least 3 strand formers (Hₜ and Hₜ) and no more than 1 strand breaker (Bₜ and Bₜ).
   B. Propagate predicted β-strand in both directions until reach a four residue window with average propensity (Pₜ) < 1.0
   C. The average propensity (Pₜ) for a predicted β-strand must be Pₜ > 1.05 and Pₜ > Pₐ

Chou-Fasman Algorithm (simplified)

3. Resolving conflicting predictions:
   (regions with both α-helix and β-strand assignment)
   • If average Pₐ > average Pₜ, then the region is alpha helix
   • If average Pₜ > average Pₐ, then the region is beta strand

§ Notes about Chou-Fasman algorithm:
   • Later versions of the algorithm included predictions for turns
   • The original algorithm contained additional rules about the location of certain residues (e.g., proline) in α-helices and β-strands
   • More recent versions of the algorithm have used sequential tetrapeptide average propensities to predict secondary structure
   • The propensity values have also been variously recalculated with larger protein data sets (original data sets based on 15 and 29 proteins)

GOR Method

§ Key difference: Chou-Fasman uses individual amino acid propensities, while GOR incorporates information about neighboring amino acids to make prediction

§ A 20 x 17 matrix of directional information values for each secondary structure class was calculated from a database of known structures

§ These matrices are used to predict the secondary structure of the central (9th) residue in a 17 residue window:

   MLNPKSYENAQLGRCTTHYA
   
   2' structure prediction for residue I is based on 17 residue window

§ The secondary structure class with highest information score over 17 residue window (e.g., I is predicted to be α-helix)
Multiple Sequence Alignment Method

§ Inclusion of information from a homologous protein sequences in the form of a multiple sequence alignment can improve accuracy of secondary structure prediction

§ First, identify homologous protein sequences to the input/target protein sequence, and perform multiple sequence alignment of these sequences

§ Second, perform secondary structure prediction using the aligned sequences as input
  • Secondary structure prediction can be done using classical (GOR, Chou-Fasman) or more modern (e.g., neural networks) secondary structure prediction methods

§ Multiple sequence alignment provides additional information about amino acid substitution patterns that can be indicative of a certain category of secondary structure (i.e., α-helix)

Neural network secondary structure prediction methods

§ Neural networks are a sophisticated computational method to perform pattern recognition, classification, or prediction tasks (among others)

§ Neural networks are modeled as a network of artificial ‘neurons’

§ Key concept: neural network software can be trained to accomplish a computational task by ‘learning’ from an example training data set

§ In secondary structure prediction, neural network methods are trained using sequences with known secondary structure, and then asked to predict the secondary structure of proteins of unknown structure

§ Example: Profile network from Heidelberg (PHD) uses multiple sequence alignment with neural network methods to predict secondary structure

Fragment Database Mining — Variant of sequence alignment method of secondary structure prediction

§ Uses BLAST to identify homologous protein sequence fragments in a protein structure database (PDB)

§ Secondary structure prediction is made using the known structures of matching sequence fragments

§ A weighting approach is used to predict secondary structure for a region with multiple matching sequence fragments of known structure
  • Weighting is based on a variety of parameters, including sequence similarity between the matching fragments and the input sequence

Accuracy of Secondary Structure Prediction

§ Prediction accuracy
  • Accuracy is usually measured by $Q_3$ (or $Q_{index}$) value
  • For a single conformation state, i:
    $$ Q_i = \frac{\text{number of residues correctly predicted in state } i}{\text{number of residues observed in state } i} \times 100\% $$
  • where $i$ is either helix, strand, or coil. For all three states:
    $$ Q = \frac{\sum_{i} \text{number of residues correctly predicted in state } i}{\text{number of all residues}} \times 100\% $$

§ Accuracy of prediction methods
  • A random prediction has a $Q_3$ value of ~ 33-38%
  • Chou-Fasman method typically has a $Q_3 \sim 56-60$
  • GOR method (depending upon version) has a $Q_3 \sim 60-65$
  • MSA, neural network methods have $Q_3 \sim 70$