Protein Structures: Experiments and Modeling

Patrice Koehl
Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Ab initio prediction

Proteins: Quality of Structures/Models
Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Ab initio prediction

Proteins: Quality of Structures/Models
Proteins: Finding the Primary Structure

Methods for finding the sequence of a protein:

- Translating gene sequence
  - For proteins from prokaryotes, direct translation
  - For proteins from eukaryotes, we need the sequence of mRNA or cDNA

- Edman degradation
  limited to “small” proteins, up to 50 amino acids for automated sequencer

- Mass spectrometry
Proteins: Finding the Primary Structure

**EDMAN DEGRADATION**

- Phenyl isothiocyanate or PITC
- Trifluoroacetic acid, TFA

(http://en.wikibooks.org/wiki/Structural_Biochemistry/Proteins/Protein_sequence_determination_techniques)
Proteins: Finding the Tertiary Structure

Methods for finding the 3D structure of a protein:

- Circular Dichroism
  (low resolution; provides information on secondary structure)

- X-ray crystallography
  (high resolution; finds structure of a protein in a crystal)

- NMR spectroscopy
  (high resolution; finds structure of a protein in solution)
Proteins: Circular Dichroism

Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed polarized light versus right-handed polarized light which arise due to structural asymmetry.

Different secondary structures in proteins have different CD spectra as they have different asymmetry.

CD therefore can detect secondary structures in protein
**Proteins: X-ray Crystallography**

**Bragg’s Law:**

\[ 2d \sin(\theta) = n\lambda \]

*From the “pattern of diffraction”, i.e. the maximum of intensities observed, we can find the angles of diffraction and for each angle we get the corresponding d using Bragg’s law.*
General principle of X-ray crystallography applied to proteins:

1) We need a crystal

2) From the diffraction pattern, we get the crystal organization

3) From the diffraction intensities, we get the electron densities

4) Once the electron density map, we fit a structure that matches with this density

5) From the atomic model, we can compute a theoretical diffraction map; if it matches with the experimental one, we are done; otherwise refine
Getting the Diffraction Pattern

Rosalyn Franklin: Diffraction pattern for DNA
One hidden problem: diffraction patterns provide intensities; for Fourier transform, need intensity and phase. A significant step in X-ray crystallography is the solve the “phase problem”.
Fitting the structure: Influence of the resolution

2.6 Å resolution  
1.2 Å resolution
# Resolution of X-ray structures

<table>
<thead>
<tr>
<th>Resolution (Å)</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;4.0</td>
<td>Individual coordinates meaningless</td>
</tr>
<tr>
<td>3.0 - 4.0</td>
<td>Fold possibly correct, but errors are very likely.</td>
</tr>
<tr>
<td>2.5 - 3.0</td>
<td>Fold likely correct except that some surface loops might be mismodeled.</td>
</tr>
<tr>
<td>2.0 - 2.5</td>
<td>Many small errors can normally be detected. Fold normally correct and number of errors in surface loops is small. Water molecules and small ligands become visible.</td>
</tr>
<tr>
<td>1.5 - 2.0</td>
<td>Many small errors can normally be detected. Folds are extremely rarely incorrect, even in surface loops.</td>
</tr>
<tr>
<td>0.5 - 1.5</td>
<td>In general, structures have almost no errors at this resolution. Geometry studies are made from these structures.</td>
</tr>
</tbody>
</table>

(www.en.wikipedia.org/wiki/Resolution_(electron_density)
Large molecular assemblies: X-ray crystallography and Cryo-EM

X-ray structure

(180 copies of the same protein)

(Norwalk virus: http://www.bcm.edu/molvir/norovirus)
Large molecular assemblies: X-ray crystallography and Cryo-EM

Cryo-EM:

- Microscopy technique; as such, do not need crystal (closer to physiological conditions)

- Not high-resolution enough to provide atomic details; used in combination with modeling
Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Ab initio prediction

Proteins: Quality of Structures/Models
Why we need Homology Modeling?

- Aim to solve the structure of all proteins: this is too much work experimentally!

- Solve enough structures so that the remaining structures can be inferred from those experimental structures

- The number of experimental structures needed depend on our abilities to generate a model.
Proteins with known structures

Unknown proteins

Why we need Homology Modeling?
Why does Homology Modeling Work?

- High sequence identity
- High structure similarity

Homology Modeling: How it works
Homology Modeling: How it works

- Find template
- Align target sequence with template
- Generate model:
  - add loops
  - add sidechains
- Refine model
Homology Modeling: Input

The query, also called target sequence

The template structure and sequence
(need high quality structure)

The sequence alignment between query and template sequence
(Probably the most important input!)
Homology Modeling: Which program to use?

Wallner B, Elofsson A.
All are not equal: a benchmark of different homology modeling programs.
Homology Modeling: Which program to use?

1) **Web service: SwissModel**
   
   http://swissmodel.expasy.org/

   3 modes:
   - fully automatic
   - “Alignment mode”: you provide your own target-template alignment
   - “Project mode”: provides an environment to edit alignment

2) **Software: Modeller**
   
   http://www.salilab.org/modeller/

   Probably the best maintained software the homology modeling
Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Databases of Structures

Proteins: Homology Modeling

Proteins: Databases of Models

Proteins: Quality of Structures/Models
Do not start a homology modeling project before checking...

- **Swiss-Model** repository
  (http://swissmodel.expasy.org/repository/)
  - Companion to the Swiss-Model tools – over 3.0 million models of protein domains computed from over 2.2 million sequences.

- **ModBase**
  (http://modbase.compbio.ucsf.edu/)
  - Companion to **Modeller**

**ModBase Statistics**

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Models</td>
<td>10408644</td>
</tr>
<tr>
<td>Number of Unique Sequences modeled</td>
<td>2430327</td>
</tr>
<tr>
<td>Unique sequences attempted</td>
<td>3367150</td>
</tr>
<tr>
<td>Number of PDB chains</td>
<td>38355</td>
</tr>
</tbody>
</table>

Statistics updated at Fri Oct 1 22:06:13 PDT 2010
Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Ab initio prediction

Proteins: Quality of Structures/Models
Given a protein sequence $a_1 a_2 \ldots a_N$, secondary structure prediction aims at defining the state of each amino acid $a_i$ as being either H (helix), E (extended=strand), or O (other) (Some methods have 4 states: H, E, T for turns, and O for other).

The quality of secondary structure prediction is measured with a “3-state accuracy” score, or $Q_3$. $Q_3$ is the percent of residues that match “reality” (X-ray structure).
Secondary Structure Assignment

Determine Secondary Structure positions in known protein structures using DSSP or STRIDE:

Early methods for Secondary Structure Prediction

- **Chou and Fasman**
  

- **GOR**
  
Chou and Fasman

- Start by computing amino acids propensities to belong to a given type of secondary structure:

<table>
<thead>
<tr>
<th>Structure</th>
<th>Propensity Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helix</td>
<td>( \frac{P(i \mid Helix)}{P(i)} )</td>
</tr>
<tr>
<td>Beta</td>
<td>( \frac{P(i \mid Beta)}{P(i)} )</td>
</tr>
<tr>
<td>Turn</td>
<td>( \frac{P(i \mid Turn)}{P(i)} )</td>
</tr>
</tbody>
</table>

Propensities > 1 mean that the residue type I is likely to be found in the corresponding secondary structure type.
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>α-Helix</th>
<th>β-Sheet</th>
<th>Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>1.29</td>
<td>0.90</td>
<td>0.78</td>
</tr>
<tr>
<td>Cys</td>
<td>1.11</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>Leu</td>
<td>1.30</td>
<td>1.02</td>
<td>0.59</td>
</tr>
<tr>
<td>Met</td>
<td>1.47</td>
<td>0.97</td>
<td>0.39</td>
</tr>
<tr>
<td>Glu</td>
<td>1.44</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Gln</td>
<td>1.27</td>
<td>0.80</td>
<td>0.97</td>
</tr>
<tr>
<td>His</td>
<td>1.22</td>
<td>1.08</td>
<td>0.69</td>
</tr>
<tr>
<td>Lys</td>
<td>1.23</td>
<td>0.77</td>
<td>0.96</td>
</tr>
<tr>
<td>Val</td>
<td>0.91</td>
<td>1.49</td>
<td>0.47</td>
</tr>
<tr>
<td>Ile</td>
<td>0.97</td>
<td>1.45</td>
<td>0.51</td>
</tr>
<tr>
<td>Phe</td>
<td>1.07</td>
<td>1.32</td>
<td>0.58</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.72</td>
<td>1.25</td>
<td>1.05</td>
</tr>
<tr>
<td>Trp</td>
<td>0.99</td>
<td>1.14</td>
<td>0.75</td>
</tr>
<tr>
<td>Thr</td>
<td>0.82</td>
<td>1.21</td>
<td>1.03</td>
</tr>
<tr>
<td>Gly</td>
<td>0.56</td>
<td>0.92</td>
<td>1.64</td>
</tr>
<tr>
<td>Ser</td>
<td>0.82</td>
<td>0.95</td>
<td>1.33</td>
</tr>
<tr>
<td>Asp</td>
<td>1.04</td>
<td>0.72</td>
<td>1.41</td>
</tr>
<tr>
<td>Asn</td>
<td>0.90</td>
<td>0.76</td>
<td>1.23</td>
</tr>
<tr>
<td>Pro</td>
<td>0.52</td>
<td>0.64</td>
<td>1.91</td>
</tr>
<tr>
<td>Arg</td>
<td>0.96</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Chou and Fasman

Favors α-Helix
Favors β-strand
Favors turn
Chou and Fasman

**Predicting helices:**
- find nucleation site: 4 out of 6 contiguous residues with $P(\alpha) > 1$
- extension: extend helix in both directions until a set of 4 contiguous residues has an average $P(\alpha) < 1$ (breaker)
- if average $P(\alpha)$ over whole region is $>1$, it is predicted to be helical

**Predicting strands:**
- find nucleation site: 3 out of 5 contiguous residues with $P(\beta) > 1$
- extension: extend strand in both directions until a set of 4 contiguous residues has an average $P(\beta) < 1$ (breaker)
- if average $P(\beta)$ over whole region is $>1$, it is predicted to be a strand
Position-specific parameters for turn:
Each position has distinct amino acid preferences.

Examples:
- At position 2, Pro is highly preferred; Trp is disfavored
- At position 3, Asp, Asn and Gly are preferred
- At position 4, Trp, Gly and Cys preferred

<table>
<thead>
<tr>
<th>f(i)</th>
<th>f(i+1)</th>
<th>f(i+2)</th>
<th>f(i+3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala 0.060</td>
<td>0.076</td>
<td>0.035</td>
<td>0.058</td>
</tr>
<tr>
<td>Arg 0.070</td>
<td>0.106</td>
<td>0.099</td>
<td>0.085</td>
</tr>
<tr>
<td>Asp 0.147</td>
<td>0.110</td>
<td>0.179</td>
<td>0.081</td>
</tr>
<tr>
<td>Asn 0.161</td>
<td>0.083</td>
<td>0.191</td>
<td>0.091</td>
</tr>
<tr>
<td>Cys 0.149</td>
<td>0.050</td>
<td>0.117</td>
<td>0.128</td>
</tr>
<tr>
<td>Glu 0.056</td>
<td>0.060</td>
<td>0.077</td>
<td>0.064</td>
</tr>
<tr>
<td>Gln 0.074</td>
<td>0.098</td>
<td>0.037</td>
<td>0.098</td>
</tr>
<tr>
<td>Gly 0.102</td>
<td>0.085</td>
<td>0.190</td>
<td>0.152</td>
</tr>
<tr>
<td>His 0.140</td>
<td>0.047</td>
<td>0.093</td>
<td>0.054</td>
</tr>
<tr>
<td>Ile 0.043</td>
<td>0.034</td>
<td>0.013</td>
<td>0.056</td>
</tr>
<tr>
<td>Leu 0.061</td>
<td>0.025</td>
<td>0.036</td>
<td>0.070</td>
</tr>
<tr>
<td>Lys 0.055</td>
<td>0.115</td>
<td>0.072</td>
<td>0.095</td>
</tr>
<tr>
<td>Met 0.068</td>
<td>0.082</td>
<td>0.014</td>
<td>0.055</td>
</tr>
<tr>
<td>Phe 0.059</td>
<td>0.041</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>Pro 0.102</td>
<td>0.301</td>
<td>0.034</td>
<td>0.068</td>
</tr>
<tr>
<td>Ser 0.120</td>
<td>0.139</td>
<td>0.125</td>
<td>0.106</td>
</tr>
<tr>
<td>Thr 0.086</td>
<td>0.108</td>
<td>0.065</td>
<td>0.079</td>
</tr>
<tr>
<td>Trp 0.077</td>
<td>0.013</td>
<td>0.064</td>
<td>0.167</td>
</tr>
<tr>
<td>Tyr 0.082</td>
<td>0.065</td>
<td>0.114</td>
<td>0.125</td>
</tr>
<tr>
<td>Val 0.062</td>
<td>0.048</td>
<td>0.028</td>
<td>0.053</td>
</tr>
</tbody>
</table>
Chou and Fasman

*Predicting turns:*
- for each tetrapeptide starting at residue i, compute:
  - $P_{\text{Turn}}$ (average propensity over all 4 residues)
  - $F = f(i)f(i+1)f(i+2)f(i+3)$

- if $P_{\text{Turn}} > P_{\alpha}$ and $P_{\text{Turn}} > P_{\beta}$ and $P_{\text{Turn}} > 1$ and $F > 0.000075$
  tetrapeptide is considered a turn.

Chou and Fasman prediction:

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm
The GOR method

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position \( j \) in the sequence, eight residues on either side are considered.

\[
\begin{array}{c}
\text{j} \\
\text{ } \\
\end{array}
\]

A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue \( j \) is helical. The helix propensity tables have 20 x 17 entries. Build similar tables for strands and turns.

**GOR simplification:**
The predicted state of AAj is calculated as the sum of the position-dependent propensities of all residues around AAj.

GOR can be used at: [http://abs.cit.nih.gov/gor/](http://abs.cit.nih.gov/gor/) (current version is GOR IV)
Both Chou and Fasman and GOR have been assessed and their accuracy is estimated to be \( Q3 = 60-65\% \).
Secondary Structure Prediction

- Available servers:

  - JPRED:  http://www.compbio.dundee.ac.uk/~www-jpred/
  - PHD: http://cubic.bioc.columbia.edu/predictprotein/
  - PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/
  - NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
  - Chou and Fassman: http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

- Interesting paper:

Protein Structure Prediction

One popular model for protein folding assumes a sequence of events:

- Hydrophobic collapse
- Local interactions stabilize secondary structures
- Secondary structures interact to form motifs
- Motifs aggregate to form tertiary structure
Protein Structure Prediction

A physics-based approach:

- find conformation of protein corresponding to a thermodynamics minimum (free energy minimum)

- cannot minimize internal energy alone!
  Needs to include solvent

- simulate folding…a very long process!

Folding time are in the ms to second time range
Folding simulations at best run 1 ns in one day…
The Folding @ Home initiative

(Vijay Pande, Stanford University)

Our goal: to understand protein folding, protein aggregation, and related diseases

What are proteins and why do they "fold"? Proteins are biology's workhorses -- its "nanomachines." Before proteins can carry out their biochemical function, they remarkably assemble themselves, or "fold." The process of protein folding, while critical and fundamental to virtually all of biology, remains a mystery. Moreover, perhaps not surprisingly, when proteins do not fold correctly (i.e. "misfold"), there can be serious effects, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, and Parkinson's disease.

Results from Folding@Home

http://folding.stanford.edu/
The Folding @ Home initiative

What does Folding@Home do? Folding@Home is a distributed computing project which studies protein folding, misfolding, aggregation, and related diseases. We use novel computational methods and large scale distributed computing, to simulate timescales thousands to millions of times longer than previously achieved. This has allowed us to simulate folding for the first time, and to now direct our approach to examine folding related disease.

How can you help? You can help our project by downloading and running our client software. Our algorithms are designed such that for every computer that joins the project, we get a commensurate increase in simulation speed. One can also help by donating funds to the project, via Stanford University.

What have we done so far? We have had several successes. You can read about them on our Science page, Results section, or go directly to our press and papers page.

Since October 1, 2000, over 1,000,000 CPUs throughout the world have participated in Folding@Home. Each additional CPU gives us an added boost in performance, allowing us to tackle more difficult problems or solve existing research faster or more accurately.
Folding @ Home: Results

Predicted folding time (nanoseconds)

Experimental measurement (nanoseconds)

Experiments:

villin:
Raleigh, et al, SUNY, Stony Brook

BBAW:
Gruebele, et al, UIUC

beta hairpin:
Eaton, et al, NIH

alpha helix:
Eaton, et al, NIH

PPA:
Gruebele, et al, UIUC

http://pande.stanford.edu/
Protein Structure Prediction

DECOYS:
Generate a large number of possible shapes

DISCRIMINATION:
Select the correct, native-like fold

Need good decoy structures
Need a good energy function
Structural Bioinformatics: Proteins

Proteins: Sources of Structure Information

Proteins: Homology Modeling

Proteins: Ab initio prediction

Proteins: Quality of Structures/Models
What is a “good” protein structure?

- **The structure must have good stereochemistry**
  - Bond lengths and bond angles close to ideal values
  - Backbone dihedral angles in “allowed” regions of the Ramachandran plot
  - Side-chains close to “rotamer” states

- **The structure should be energetically favorable**
  - No “clashes” (such as two atoms too close to each other)
  - Hydrophobic residues should be mostly buried
  - Polar residues should be on the surface. No charges buried.

I advise running Procheck to check the quality of a protein structure:

- **As a standalone program:**
  [http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/](http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)

- **Or through the PDBsum web portal:**
  [http://www.ebi.ac.uk/pdbsum/](http://www.ebi.ac.uk/pdbsum/)
What have we learnt?

- X-ray crystallography is the method of choice for the elucidation of the high-resolution structure of a protein.

- Alternative methods include NMR and Cryo-EM.

- The mapping between protein sequence and protein structure is not one-to-one: for one sequence, there is (usually) one structure, but for one structure, there can be many sequences.
What have we learnt?

- Homology modeling is a computational technique that predicts the structure of a target protein based on a template structure and an alignment between the target and template sequences.

- The quality of the alignment defines the quality of the homology models.

- Two good programs for homology modeling: Swiss-Model (web-based) and Modeller (standalone software)

- The swiss model repository and ModBase are useful databases of protein structure models generated by homology modeling.

- It is usually good practice to check the quality of a protein structure or model. ProCheck is a useful tool for this: it checks both the stereochemistry and energetics of the protein.