**Master Course**

DNA/Protein Structure-function Analysis and Prediction

**Lecture 8**

**Protein Structure Prediction (II): Fold Prediction**

---

**Importance of Protein Folding**

"Understanding protein structure, function and dynamics ranks among the most challenging and fascinating problems faced by science today. Since the function of a protein is related to its three-dimensional structure, manipulation of the latter by means of mutation in the protein sequence generates functional diversity. The keys that will help us understand this mechanism and consequently protein sequence evolution lie in the yet unknown laws that govern protein folding. The knowledge of these laws would also prove useful for engineering protein molecules to optimize their activities as well as to alter their pharmacokinetic properties in the case of therapeutically important molecules." Patrice Kehdi, Stanford University

---

**Sequence-Structure-Function**

**How to get a structure: Experimental**

- Crystallography by X-ray diffraction
  - most reliable technique to date
  - depending on proteins that do want to crystallize
- Crystallography by electron diffraction
  - cryo-electron microscopy and image analysis
  - periodic ordering of proteins in two-dimensions as well as along one-dimensional helices
  - appropriate for example for membrane proteins
- Nuclear Magnetic Resonance
  - although magnets become stronger, only smaller structures can be solved
  - no need to make crystals
  - yields distance information (NOEs)
  - relies on distance geometry algorithms to convert distance information to 3D-model
- Mass Spectrometry
  - classic use is protein sequence determination
  - now used for elucidating structural features such as disulfide-bond, post translational modifications, protein-protein interaction, antigen epitopes, etc.,

---

**Protein folding**

Two very different principles are referred to when researchers talk about the “protein folding problem”:

1. The physical process of getting from the unfolded to the folded conformation: the folding pathway (biophysics)
2. Associating a three-dimensional protein structure to its sequence (computational biology, bioinformatics)
Protein folding
Classical example of folding pathway study: BPTI folding pathway studied by Tom Creighton and colleagues (see Creighton’s book *Proteins*) using disulphide arrangements (6 Cys residues making 3 disulfide bridges). Creighton has maintained for years that proteins make “mistakes” along the folding pathway (he based this on measuring “incorrect” disulphide bonds) which need to be “corrected” in order to attain the native fold. Discussions are ongoing but drifting away from this hypothesis.

How to predict a tertiary structure of a protein?
- *Ab initio* (using first principles) is difficult
- Homology modeling is most successful to date
  - For a query sequence:
    - Given a template sequence and structure that is deemed homologous
    - Model query sequence using the template structure (and sequence)
    - Crucially dependent on query-template alignment
  - threading

How to get a structure: *ab initio* modelling
- Scoring function: assume lowest energy structure is native one
  - Thermodynamic approach recognises a potential function of sequence and conformation that has its global minimum at the native conformation for many different proteins
  - Is this always the case? Think about disparities, etc.
  - Full-scale molecular force fields: e.g., ICMPP, AMBER, Mexx
  - Simplified force fields
  - Knowledge-based potentials — “Sippe” potentials (potentials of mean force)
  - “Empirical” parameters

Bioinformatics tools
Search optimisation algorithm:
- Scoring function
  - Often the most important part
- Search function

Monitoring folding pathways

Figure 1
Three-dimensional representation of the native folding space of polypeptides with 4, 5 and 6 cysteine residues (B, H and G, respectively). The numbers represent combinations, the number of disulfide bridges is indicated with numbers on the left of each panel. The edges indicate the fold-change exchange transition. Dots and lines (fully solid, dashed, thin, in the Centre of space or highly localized intermediate) show, of whatever reason, the major exchange transition. Different colors show the same facts but in a different way (fully solid, dotted, thin, in the Centre of space or highly localized intermediate) (A-G). The numbers correspond to the number of disulfide bridges created or abolished. A simple visualization tool for the Fold package. http://www.bedgeploo.org.html

PMCs Bioinformatics, 2005; 6: 19.
How to get a structure: *ab initio* modelling

- Search function: need to be able to move or change conformation
  - Molecular Dynamics (kT)<sup>3</sup>
  - Monte Carlo (Boltzmann equation)
  - Simulated annealing (very temperature)
  - Brownian motion modelling

![Techniques to enhance the reaching power of MD simulation include: use of self-core potentials, extension of the Cartesian space to 4 dimensions, local elevation of the potential energy surface, etc.](image)

**Molecular Mechanics and Force Fields**

- AMBER, Assisted Model Building and Energy Refinement
- AMBER, Open Force Field with 3D-Viewing (O3F3)
- CHARMM, Chemistry at HARvard MolecuMolecular Mechanics
- DISCOVER, force fields of the Insight Discovery package
- ECP/LJ, a pairwise potential for proteins and peptides
- GROMOS, GB/SA Molecular Simulation package
- The Sybyl 5.5 Home page

**Potentials of mean force**

- Potentials of mean force describe the interaction between residues.
- It is possible to calculate such potentials by performing long simulations at the atomic level.
- In reality, this is not practical because of the amount of computations involved and also because our understanding of protein behavior on the atomic level is insufficient.

- However, if we assume that residues in an ensemble of proteins follow a Boltzmann distribution describing their location, mutual interaction, etc., then we can estimate the potential of mean force by analyzing the distribution of their occurrence.

\[ p_j = \exp(-\beta U_j) \]

\( \beta \) is the Boltzmann constant

**Energy potentials**

Two main types of energy functions have been explored in the context of *in silico* protein studies:

3. Semi-empirical potentials
4. Knowledge-based potentials

**Semi-empirical potentials**

- Are derived from analytical expressions, describing the different interactions encountered in proteins.
- Parameters are obtained by fitting experimental data on small molecules and/or from quantum mechanical calculations (Halgren, 1995; Moult, 1997; Lazaridis and Karplus, 2000).
- The advantage corresponds to well-defined interactions, with a clear physical basis.
- Delicate aspects of this approach include the parameterization of the functions and the inclusion of solvent and other entropic effects.
- The use semi-empirical potentials is generally very expensive in terms of computer time, as they require a full atomic protein representation and, preferentially, explicit solvent molecules.

**Knowledge-based potentials**

- Widely used in simulations of protein folding structure prediction, and protein design.
- Advantages include limited computational requirements and the ability to deal with low-resolution protein models compatible with long-scale simulations.
- Drawbacks are their dependence on specific features of the dataset from which they are derived, such as the size of the proteins it contains, and their physical meaning, which is still a subject of debate.
Knowledge-based potentials (Cnt.)

- Statistical or knowledge-based potentials are derived from datasets of known protein structures. They can be easily adapted to simplified protein models, taking the solvent implicitly into account and including some entropic contributions (Sippl, 1985; Jermign and Bahar, 1996; Moult, 1997; Lazaridis and Karplus, 2000).
- However, their physical significance is less straightforward, basically because they are mean-force potentials, usually residue-based, in which different kinds of atom–atom interactions and entropic effects are mixed.

Knowledge-based potentials (Cnt.)

- These potentials are either obtained by optimization of the parameters of a predefined analytical form by requiring them to yield a large energy gap between native and unfolded states (e.g., Crippen, 1991; Goldstein et al., 1992; Mistry and Shapiro, 1996; Tobi et al., 2000; Vendruscolo et al., 2000), or derived from observed frequencies of association of specific sequence and structure elements (e.g., Tanaka and Scheraga, 1976; Miyazawa and Jermign, 1985; Kang et al., 1993; Keohar et al., 1994; Sopp, 1995; Simons et al., 1997; Melo and Feytmann, 1997; Lu et al., 2003).
- Energy functions describing different types of interactions are obtained according to the kind of structure elements considered, the assumptions made, and the reference state used (Godzik et al., 1995; Du et al., 1998; Roosman and Gilis, 1998).

Knowledge-based potentials (Cnt.)

- Preceding slide mentions Tanaka and Scheraga, 1976; Miyazawa and Jermign, 1985; Crippen, 1991
- Despite this history, these potentials are often referred to as Sippl potentials, after Manfred Sippl who wrote a paper in 1995 that became popular (and did not cite his predecessors; mind you, he had been a postdoc in Crippen’s and Jermign’s labs...).
- As did others, Sippl played around with the distribution of pairwise residue distances observed in the protein data bank.
- Can you imagine what can be done with these potentials?

Knowledge-based potentials

Example: distance-derived potential

- Construct a database of all 20x20 or 21x21 amino acid pairs
- Derive a potential using $E = -k_B T \ln \tanh(\beta r)$
- Predict a given sequence using the pairwise potentials

Researchers Design and Build First Artificial Protein

November 23, 2003 — Using sophisticated computer algorithms running on standard desktop computers, researchers have designed and constructed a novel functional protein that is not found in nature. The achievement should enable researchers to explore larger questions about how proteins evolved and why nature “chose” certain protein folds over others.

The ability to specify and design artificial proteins also opens the way for researchers to engineer artificial protein enzymes for use as medicines or industrial catalysts, said the study’s lead author, Howard Hughes Medical Institute investigator David Baker at the University of Washington.

Baker and his colleagues took advantage of methods for sampling alternative protein structures that they have been developing for some time as part of the Rosetta ab initio protein structure prediction methodology. “Indeed, the integration of protein design algorithms (to identify low energy amino acid sequences for a fixed protein structure) with protein structure-prediction algorithms (which identify low energy protein structures for a fixed amino acid sequence) was a key ingredient of our success,” Baker said.

In their design and construction effort, the scientists chose a version of a globular protein of a type called an alpha/beta conformation that was not found in nature. “We chose this conformation because there are many of this type that are currently found in nature, but there are glaring examples of possible folds that haven’t been seen yet,” he said. “We chose a fold that has not been observed in nature.”
Their computational design approach was iterative, in that they specified a starting backbone conformation and identified the lowest energy amino acid sequence for this conformation using the RosettaDesign program they had developed previously. RosettaDesign is available free to academic groups at www.ssc.unc.edu/kushnir/rosetta.htm.

They then kept the amino acid sequence fixed and used the Rosetta structure prediction methodology they had previously used successfully for ab initio protein structure prediction to identify the lowest energy backbone conformation for this sequence.

Finally, they fed the results back into the design process to generate a new sequence predicted to fold to the new backbone conformation. After repeating the sequence optimization and structure prediction steps 10 times, they arrived at a protein sequence and structure predicted to have lower energy than naturally occurring proteins in the same size range.

The result was a 93-amino acid protein structure they called Top7. “It’s called Top7, because there was a previous generation of proteins that seemed to fold right and were stable, but they didn’t appear to have the perfect packing, seen in native proteins,” said Baker.

The researchers synthesized Top7 to determine its real-life, three-dimensional structure using x-ray crystallography. As the x-rays pass through and bounce off of atoms in the crystal, they have a diffraction pattern, which can then be analyzed to determine the three-dimensional shape of the protein.

“Of the real surprises came when we actually solved the crystal structure and found it to be marvelously close to what we had been trying to make,” said Baker. “That gave us encouragement that we were on the right track.”

According to Baker, the achievement of designing a specified protein fold has important implications for the future of protein design. "Probably the most important lesson is that we can now design completely new proteins that are very stable and are very close in structure to what we were aiming for," he said. "And secondly, this design shows that our understanding of the energetics of proteins and other macromolecules cannot be too far off; otherwise, we never would have been able to design a completely new molecule with this accuracy." The next big challenge, said Baker, is to design and build proteins with specified functions, an effort that is now underway in his laboratory.

---

The artificial protein Top7 was designed from a starting configuration and sequence by iterating a threading technique and an ab initio 3D-model building protocol (Rosetta software suite).

**Ab initio**

**Sequence**

**Structure**

**Threading**

---

The resulting protein sequence and structure predicted Top7 had a lower (calculated) energy than naturally occurring proteins in the same size range!

A computer-generated image of the artificial protein, Top7.

---

**Top 7 recipe:**

*Choose globular protein of a type called an alpha-beta conformation (antiparallel 5-stranded beta-sheet with 2 alpha-helices at one side of the sheet)*

*Design starting backbone conformation and identify the lowest energy amino acid sequence (threading)*

*Keep amino acid sequence fixed and use Rosetta for ab initio protein structure prediction to identify the lowest energy backbone conformation for this sequence.*

*Then feed results back and generate a new sequence predicted to fold to the new backbone conformation (threading).*

*Iterate sequence optimization and structure prediction steps 10 times.*

---

**Convergent and Divergent Evolution**

There are entire groups of sequentially unrelated, but structurally similar (i.e. homologous), proteins. Thus, even when sequence similarity is not detectable, correct structural templates might exist in the database of solved protein structures such as in the Protein Data Bank. If such topological cousins could be easily identified, the number of proteins whose structures could be predicted would increase significantly.

A new class of structure prediction methods, termed *inverse folding or threading*, has been specifically formulated to search for such structural similarities. However, topological cousins may differ substantially in their structural details, even when their overall topology is identical. For example, the root mean square deviation, RMSD, of their backbone atoms may differ by 3-4 Å in the core and sequence identity can be as low as 10%. Thus, it is a non-trivial problem to recognize such topological cousins as being related.
Convergent and Divergent Evolution

This question touches on an important problem: are these proteins related by evolution (i.e., homologous) or not? Perhaps current sequence-based similarity searches are simply not sensitive enough to detect very distant homologies. For many such protein groups, there are hints of distant evolutionary relationships, such as functional similarity or limited sequence similarity in the important regions of the protein. For some other protein fold groups, there are no obvious relations between their function or any other observations that suggest homology—for example the globin-like fold of bacterial toxin colicin. Such protein groups may indicate that the universe of protein structures is limited, and proteins end up having similar folds because they must choose from a limited set of possibilities.

Convergent or Divergent Evolution

Convergent

Our goal is to predict a structure likely to be adopted by the given sequence, while avoiding pitfalls of ab initio folding simulations such as long simulation times and exploring conformations that are unlikely to be seen in folded proteins. To allow for scanning of large structural databases within a reasonable length of time, algorithms use an extremely simplified description of a protein structure.

Convergent or Divergent Evolution

Divergent

Different tools would be appropriate to recognize proteins from extended homologous families vs. non-homologous but structurally converging protein groups. The first choice would indicate the enhancement of tools of standard sequence analysis. For instance, multiple alignments could be used to create “profiles” where invariant positions within the family of related proteins are weighted more heavily than more variant positions.

Threading

Template sequence

Query sequence

Template structure

Structure-based function prediction

Threading

- Scoring function for measuring to what extent query sequence fits into template structure
- For scoring we have to map an amino acid (query sequence) onto a local environment (template structure)
- We can use the following structural features for scoring:
  - Secondary structure
  - Is environment inside or outside? – Residue accessible surface area
  - Polarity of environment
- The best (highest scoring) “thread” through the structure gives a so-called structural alignment, this looks exactly the same as a sequence alignment but is based on structure.
Threading – inverse folding
Map sequence to structural environments

Threading
Searching for compatibility between the structure and the sequence (in principle disregarding possible evolutionary relationships) – inverse folding

*3D profiles of Bowie et al. (1991) are formally equivalent to the “frozen approximation” of the topology fingerprint method of Godzik et al. In each case, a position dependent mutation matrix is created and used in the dynamic programming alignment. For 3D profiles, it is based on the classification of environments of each position. In the topology fingerprint method, the energy of each possible mutation is calculated by summing up interactions at each position.

*Some potential energy parameters used in sequence-structure recognition methods contain a strong sequence-sequence similarity component, because the same amino acid features are important to both. For instance, hydrophobicity is a main component in both mutation matrices and some interaction parameter sets.

Find the best compromise over all environments

What is the optimal thread for each local environment?

Fold recognition by threading

Threading
Searching for compatibility between the structure and the sequence (in principle disregarding possible evolutionary relationships) – inverse folding.

*Some similarities between methods also occur when potential energy parameters contain a strong “sequence memory” by including contributions from amino acid composition or size.

*There are also methods that explicitly combine elements of both approaches, such as enhancing sequence similarity by residue burial status, secondary structure, or a generalized “interaction environment”. Algorithms that follow these ideas are still being developed.

Bowie et al. (1991) 3D-1D structure to sequence matching

*Define 17 different structural environments for each residue position in the structure (based on secondary structure, hydrophobicity, solvent exposure)

*Bowie et al. (1991) 3D-1D structure to sequence matching

*Make a 20x17 amino acid to structural template matrix

*Align structure against sequence using the structure–sequence matrix (using Dynamic Programming)
The Inverse Folding Paradigm

In an inverse folding approach, one threads a probe sequence through different template structures and attempts to find the most compatible structure. Since large structural databases must be screened, such threading algorithms are optimized for speed. Normally, a simplified representation of the protein with a simplified energy function is used to evaluate the fitness of the probe sequence in each structure. In the last few years, different fitness functions and algorithms have been developed, and protein threading has become one of the most active fields in theoretical molecular biology. In all cases, the paradigm of homology modeling is followed with its three basic steps of identifying the structural template, creating the alignment and building the model. As a result, the threading approach to structure prediction has limitations similar to classical homology modeling.

The Inverse Folding Paradigm (Cont.)

Most importantly, an example of the correct structure must exist in the structural database that is being screened. If not, the method will fail. The quality of the model is limited by the extent of actual structural similarity between the template and the probe structure. At present, one cannot readjust the template structure to more correctly accommodate the probe sequence. In practice, for the best threading algorithms, the accuracy of the template recognition is well above 50%, and the quality of the predicted alignments, while somewhat better than sequence-based alignments, is still far from those obtained on the basis of the best structural alignments. In the last several years, over 15 threading algorithms have been proposed in the literature. An example is GenoFold, which has been described in a number of publications and has been utilized by a number of groups to make structural predictions, where it has performed quite favorably when compared to other approaches.

Top score structure 20 a.a. fragments in the high specificity regions --- Sequence: 3icb (residues 31–50)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Starting position</th>
<th>Score</th>
<th>Cα rms.d.</th>
<th>Secondary structure (DSSP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ucb</td>
<td>31</td>
<td>5.75</td>
<td>0.80</td>
<td>HHHHHTTTTHHHHH</td>
</tr>
<tr>
<td>Kick</td>
<td>30</td>
<td>6.10</td>
<td>0.60</td>
<td>COSTKTCKTAS</td>
</tr>
<tr>
<td>Score</td>
<td>258</td>
<td>4.81</td>
<td>4.40</td>
<td>HOOT TT TTHTTHTHTHHH</td>
</tr>
<tr>
<td>Test A</td>
<td>71</td>
<td>2.88</td>
<td>0.80</td>
<td>TKKKKKKKKTTTTE</td>
</tr>
<tr>
<td>Test B</td>
<td>38</td>
<td>0.79</td>
<td>0.37</td>
<td>EKKKATTTTVE</td>
</tr>
<tr>
<td>Test C</td>
<td>85</td>
<td>1.73</td>
<td>3.11</td>
<td>KKKKKKKKKKKKKKKKKKKKK</td>
</tr>
<tr>
<td>Test D</td>
<td>75</td>
<td>0.98</td>
<td>0.90</td>
<td>KKKKKKKKKKKKKKKKKKKKK</td>
</tr>
<tr>
<td>Splice</td>
<td>28</td>
<td>0.87</td>
<td>0.87</td>
<td>KKKKKKKKKKKKKKKKKKKKK</td>
</tr>
<tr>
<td>Usp</td>
<td>199</td>
<td>0.54</td>
<td>2.50</td>
<td>KKKKKKKKKKKKKKKKKKKKK</td>
</tr>
<tr>
<td>Rgap</td>
<td>1.49</td>
<td>0.84</td>
<td>0.84</td>
<td>KKKKKKKKKKKKKKKKKKKKK</td>
</tr>
</tbody>
</table>

> The native structure is on top

Top-scoring structural 20 a.a. fragments in regions where the native state does not have lowest score but the Cα rms.d.s are low --- Sequence: 3icb (residues 36–55)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Starting position</th>
<th>Score</th>
<th>Cα rms.d.</th>
<th>Secondary structure (DSSP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iuda</td>
<td>75</td>
<td>5.94</td>
<td>3.16</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>Iuda’</td>
<td>72</td>
<td>6.59</td>
<td>3.04</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>Top A</td>
<td>102</td>
<td>6.33</td>
<td>3.14</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>Top B</td>
<td>106</td>
<td>5.80</td>
<td>3.00</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>Uuda</td>
<td>67</td>
<td>5.87</td>
<td>3.76</td>
<td>TTTTTTTTTTTTTTTTTTTTTTTT</td>
</tr>
<tr>
<td>Uuda’</td>
<td>106</td>
<td>5.42</td>
<td>4.68</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>Top A</td>
<td>94</td>
<td>5.96</td>
<td>3.58</td>
<td>TTTTTTTTTTTTTTTTTTTTTTTT</td>
</tr>
<tr>
<td>Top B</td>
<td>104</td>
<td>5.65</td>
<td>2.47</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
<tr>
<td>Uuda</td>
<td>36</td>
<td>7.06</td>
<td>8.18</td>
<td>TTTTTTTTTTTTTTTTTTTTTTTT</td>
</tr>
<tr>
<td>Rgap</td>
<td>1.58</td>
<td>7.06</td>
<td>8.18</td>
<td>HHHHHTTTTHHHHHHHHHHHHHHHH</td>
</tr>
</tbody>
</table>

> The native structure is not on top