PROTEIN STRUCTURE PREDICTION #1

GARLAND R. MARSHALL, PH.D.
PROFESSOR OF BIOCHEMISTRY AND MOLECULAR BIOPHYSICS AND OF BIOMEDICAL ENGINEERING
WASHINGTON UNIVERSITY, ST. LOUIS
USA
THREADING - assume sequence folds as a known 3D structure (use library of known folds)

Fragment Assembly - use library of small fragment (6-12 residues), overlap and seek most probable fold combined with minimization.

Ab Initio - Contact Matrix Approach
Given the sequence, and a set of folds observed in PDB, see if any of the sequences could adopt one the known folds. Takes advantage of knowledge of existing structures, and principles by which they are stabilized (favourable interactions).

Score the folds.

Unlike sequence-only comparison, these methods take advantage of the extra information made available by 3D structure information. In effect, they turn the protein folding problem on its head: rather than predicting how a sequence will fold, they predict how well a fold will fit a sequence.
Widely used databases of threading programs

- **CATH**
  *Orengo et al. 1997*
  - Hierarchical classification of protein domain structures

- **FSSP**
  *Holm & Sander 1996*
  - Hierarchical classification of protein chains structures

- **SCOP**
  *Murzin et al. 1995*
  - Manually classify protein domains into fold, super-family, families.
SCOP-Structural Classification of Proteins
http://scop.berkeley.edu/

- A database that describes **structural** and **evolutionary** relationships between proteins of known structure.
- Many levels exist in the hierarchy; the principal levels are **family**, **superfamily** and **fold**
- Created mainly by manual inspection.

**Family:** *Clear evolutionarily relationship*

- Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater.

- In some cases similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%. 
Fold: *Major structural similarity*

- A common fold - the same major secondary structures in the same arrangement and with the same topological connections.

- Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation.

- Proteins placed together in the same fold category may not have a common evolutionary origin

Superfamily: *Probable common evolutionary origin*

- Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexokinase together form a superfamily.
### SCOP: Structural Classification of Proteins 1.61 release

17406 PDB Entries (1 September 2002), 44327 Domains, 28 Literature References
(excluding nucleic acids and theoretical models)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>151</td>
<td>257</td>
<td>409</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>111</td>
<td>213</td>
<td>362</td>
</tr>
<tr>
<td>Alpha and beta proteins (a/b)</td>
<td>117</td>
<td>190</td>
<td>467</td>
</tr>
<tr>
<td>Alpha and beta proteins (a+b)</td>
<td>212</td>
<td>308</td>
<td>488</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>39</td>
<td>39</td>
<td>52</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>12</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>Small proteins</td>
<td>59</td>
<td>84</td>
<td>128</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>701</strong></td>
<td><strong>1110</strong></td>
<td><strong>1940</strong></td>
</tr>
</tbody>
</table>

### SCOP: Structural Classification of Proteins 1.65 release

20619 PDB Entries (1 August 2003), 55475 Domains, 1 Literature Reference
(excluding nucleic acids and theoretical models)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>179</td>
<td>299</td>
<td>480</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>126</td>
<td>248</td>
<td>462</td>
</tr>
<tr>
<td>Alpha and beta proteins (a/b)</td>
<td>121</td>
<td>199</td>
<td>542</td>
</tr>
<tr>
<td>Alpha and beta proteins (a+b)</td>
<td>234</td>
<td>349</td>
<td>567</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>38</td>
<td>38</td>
<td>53</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>36</td>
<td>66</td>
<td>73</td>
</tr>
<tr>
<td>Small proteins</td>
<td>66</td>
<td>95</td>
<td>150</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>800</strong></td>
<td><strong>1294</strong></td>
<td><strong>2327</strong></td>
</tr>
</tbody>
</table>
Structural Classification of Proteins

Welcome to SCOP: Structural Classification of Proteins 1.65 release (December 2003).

Authors: Alexey G. Murzin, Loredana Lo Conte, Antonina Andreeva, Dave Howorth, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mbc-lmb.cam.ac.uk

Access methods

- Enter SCOP at the top of the hierarchy
- Keyword search of SCOP entries
- SCOP parseable files (MRC site)
- All SCOP releases and reclassified entry history new (MRC site)
- SCOP domain sequences and pdb-style coordinate files (ASTRAL)
- Hidden Markov Model library for SCOP superfamilies (SUPERFAMILY)
- Online resources of potential interest to SCOP users

SCOP mirrors around the world may speed your access.

News

Search the scop database [scop 1.65]

Please read the release notes for a detailed explanation and examples
Structural Classification of Proteins

Root: scop

Classes:

1. All alpha proteins [46456] (179)
2. All beta proteins [43724] (125)
3. Alpha and beta proteins (α/β) [51349] (121)
   - Mixed: parallel beta sheets (beta-allobeta units)
4. Alpha and beta proteins (α/β) [53921] (234)
   - Mixed: antiparallel beta sheets (aggregated alpha and beta regions)
5. Multi-domain protein (α-helical and beta) [56578] (39)
   - Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56583] (36)
   - Does not include proteins in the immune system
7. Small proteins [56652] (46)
   - Linearly constructed by nested zigzag, heme, and/or disulfide bridges
8. Coiled-coil proteins [57142] (6)
9. Low resolution protein structures [58117] (18)
10. Peptides [58231] (105)
11. Nondenatured proteins [58781] (39)

Experimental structures of proteins with essentially non-natural sequences.

Class: All alpha proteins

Lineage:

1. Root: scop
2. Class: All alpha proteins [46456]

Folds:

1. Globin-like [46457] (2)
   - core: 6 helices; folded leaf, partly opened
2. Long alpha-hairpin [46556] (11)
   - 2 helices; antiparallel hairpin, left-handed twist
3. Type I dockerin domain [63445] (1)
   - tandem repeat of two calcium-binding loop-helix motifs
4. LEM/SAP HeH motif [63450] (4)
   - helix-extended loop-helix; parallel helices
5. Cytochrome c [66265] (1)
   - core: 3 helices; folded leaf, opened
6. DNA/RNA-binding 3-helical bundle [46688] (12)
   - core: 3 helices; bundle, closed or partly opened, right-handed twist
7. Another 3-helical bundle [81602] (2)
   - topologically similar to the DNA/RNA-binding bundles;
8. RuvA C-terminal domain-like [46928] (6)
   - 3 helices; bundle, right-handed twist
9. S13-like RZTH domain [81297] (1)
   - core: 3 helices
10. Putative DNA-binding domain [46954] (1)
    - core: 3 helices; architecture is similar to that of the "wii
Search Results for "myoglobin" [scop 1.65]

Protein: Myoglobin from Yellowfin tuna (Thunnus albacares)

Family: Globins

Heme-binding protein

Lineage:
1. Root scop
2. Class: All-alpha proteins [46456]
3. Fold: Globin-like [46457]
   core: 6 helices, folded sheet, partly opened
4. Superfamily: Globin-like [46458]
5. Family: Globin [46463]
   Heme-binding protein
6. Protein: Myoglobin [46459]
7. Species: Yellowfin tuna (Thunnus albacares) [46478]

PDB Entry Domains:
1. *myoglobin* [15307] completed with hem
Energy Function

\[ E_{total} = E_{mutate} + E_{single} + E_{pair} + E_{gap} \]

\( E_{mutate} \) the sum of the compatibility measurements \( e_{\text{mutate}}(a_1; a_2) \) for substituting template amino acid \( a_1 \) by target amino acid \( a_2 \), PROSPECT use PAM250 matrix for \( e_{\text{mutate}} \).

\( E_{single} \) the sum of the preferences \( e_{\text{single}}(a; s; t) \) for aligning amino acid \( a \) of the target sequence onto a template position with a structural environment defined by secondary structure \( s \) and solvent accessibility \( t \)

\( E_{pair} \) the sum of pair-contact potentials \( e_{\text{pair}}(a_1, a_2) \) between amino acids \( a_1 \) and \( a_2 \) of the target sequence when they are aligned to template positions that are spatially close

\( E_{gap} \) the sum of the penalties \( e_{\text{gap}}(g) = 10.8 + 0.6^*(g-1) \) for an alignment gap of length \( g \)

D. Xu and Ying Xu, "Computational Studies of Protein Structure and Function using Threading Program PROSPECT"
3D-1D method (Inverse folding)

Instead of aligning a sequence to a sequence, we align a sequence by means of a string of descriptors that describe the 3D environment of the target structure. For each residue position in the structure, we determine:

- how buried it is (buried, partly buried or exposed)
- the fraction of surrounding environment that is polar (polar or apolar)
- the local secondary structure (-helix, -sheet or other)

There are 6 classes of environments to each position in the structure. These environments (E, P1, P2, B1, B2 and B3) depend on the number of surrounding polar residues and how buried the position is. Since there are 3 possible secondary structures for each of these, we have a total of $6 \times 3 = 18$ environment classes.

\[
\text{score}_{ij} = \ln \left( \frac{Pr(\text{residue } j \text{ in environment } i)}{Pr(\text{residue } j \text{ in any environment})} \right)
\]

The denominator is obtained from amino acid frequencies present in the PDB.
New score matrix based on structure but not evolution!!!

<table>
<thead>
<tr>
<th>Position in fold</th>
<th>Environment class</th>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>...</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>V</th>
<th>W</th>
<th>Y</th>
<th>Opn</th>
<th>Ext</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E</td>
<td>12</td>
<td>-46</td>
<td>22</td>
<td>3</td>
<td>-190</td>
<td>113</td>
<td>...</td>
<td>-32</td>
<td>32</td>
<td>12</td>
<td>-91</td>
<td>-214</td>
<td>-94</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>B_2</td>
<td>-56</td>
<td>-5</td>
<td>-128</td>
<td>-135</td>
<td>105</td>
<td>-166</td>
<td>...</td>
<td>-80</td>
<td>-117</td>
<td>-78</td>
<td>60</td>
<td>102</td>
<td>112</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>E_α</td>
<td>46</td>
<td>-44</td>
<td>44</td>
<td>59</td>
<td>-220</td>
<td>68</td>
<td>...</td>
<td>-34</td>
<td>15</td>
<td>-17</td>
<td>-110</td>
<td>-135</td>
<td>-210</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>4</td>
<td>P_2_α</td>
<td>6</td>
<td>-93</td>
<td>28</td>
<td>56</td>
<td>-143</td>
<td>-50</td>
<td>...</td>
<td>50</td>
<td>-18</td>
<td>-5</td>
<td>-48</td>
<td>-114</td>
<td>-79</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>5</td>
<td>E_α</td>
<td>46</td>
<td>-44</td>
<td>44</td>
<td>59</td>
<td>-220</td>
<td>68</td>
<td>...</td>
<td>-34</td>
<td>15</td>
<td>-17</td>
<td>-110</td>
<td>-135</td>
<td>-210</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>6</td>
<td>P_2_α</td>
<td>6</td>
<td>-93</td>
<td>28</td>
<td>56</td>
<td>-143</td>
<td>-50</td>
<td>...</td>
<td>50</td>
<td>-18</td>
<td>-5</td>
<td>-48</td>
<td>-114</td>
<td>-79</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>7</td>
<td>B_2_α</td>
<td>-59</td>
<td>-10</td>
<td>-162</td>
<td>-71</td>
<td>90</td>
<td>-149</td>
<td>...</td>
<td>6</td>
<td>147</td>
<td>-150</td>
<td>68</td>
<td>50</td>
<td>85</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>8</td>
<td>E_α</td>
<td>46</td>
<td>-44</td>
<td>44</td>
<td>59</td>
<td>-220</td>
<td>68</td>
<td>...</td>
<td>-34</td>
<td>15</td>
<td>-17</td>
<td>-110</td>
<td>-135</td>
<td>-210</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>9</td>
<td>P_2_α</td>
<td>6</td>
<td>-93</td>
<td>28</td>
<td>56</td>
<td>-143</td>
<td>-50</td>
<td>...</td>
<td>50</td>
<td>-18</td>
<td>-5</td>
<td>-48</td>
<td>-114</td>
<td>-79</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>10</td>
<td>B_1_α</td>
<td>-56</td>
<td>-73</td>
<td>-197</td>
<td>-174</td>
<td>132</td>
<td>-253</td>
<td>...</td>
<td>-167</td>
<td>-273</td>
<td>-129</td>
<td>66</td>
<td>100</td>
<td>18</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Now you align your target sequence with many **know structure** proteins in a database using this matrix score, obtaining templates, as in comparative modeling...
Threading methods:

Given: a structure $P$ with positions $p_1, p_2, \ldots, p_n$, and a sequence $s_1, \ldots, s_m$.
Find: $t_1, t_2, \ldots, t_n$ (where $1 \leq t_1 < t_2 < \cdots < t_n \leq m$ and $t_i$ indicates the index of the amino acid from $s$ that occupies $p_i$) such that

$$
\sum_{i=1}^{n} \sum_{j=1}^{n} \text{score}(i, j, s_{t_i}, s_{t_j})
$$

is maximized.

Instead of modeling energies from first physical principles, simplify the problem by positioning only amino acids, and compute empirical energies from the observed associations of amino acids

$$
E(\text{interaction}) = -KT \ln[\text{frequency of interaction}]
$$

where $K$ is constant, $T$ is temperature (constant), frequency of interaction measured in database of known structures.

More frequent $\rightarrow$ more favorable.
This problem is NP-complete for pairwise interactions.

- restriction in pairwise distances
- priority to well defined secondary structure

\[
\Delta E_k^{ab} = RT \ln(1 + m_{ab} \sigma) - RT \ln \left( 1 + m_{ab} \sigma \frac{f_k^{ab}(S)}{f_k(S)} \right)
\]

Journal of Molecular Biology 287, (1999), 797-815

where \( m_{ab} \) is the number of pairs \( ab \) observed with sequence separation \( k \), \( \sigma \) is the weight given to each observation, \( f_k(S) \) is the frequency of occurrence of all residue pairs at topological level \( k \) and separation distance \( s \), \( f_k^{ab}(S) \) is the equivalent frequency of occurrence of residue pair \( ab \), and \( RT \) is taken to be 0.582 kcal/mol

An obvious disadvantage:

). Doesn’t compute the actual environment created by mounting the sequence on the structure. Assumes that the environment is relatively constant, and that only amino acid details change
A schematic of sequence-structure alignment: The goal is to find an alignment between the template and the target sequence so that $E_{\text{total}}$ in Eq. is minimized.
→ Secondary Structure prediction methods:

http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
**Predict Secondary Structure (PSIPRED)**

PSIPRED is a simple and reliable secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST).

Version 2.0 of PSIPRED includes a new algorithm which averages the output from up to 4 separate neural networks in the prediction process.

Using a very stringent cross validation method to evaluate the method's performance, PSIPRED 2.0 is capable of achieving an average Q3 score of nearly 78%.

Predictions produced by PSIPRED were also submitted to the CASP4 server and assessed during the CASP4 meeting, which took place in December 2000 at Asilomar. PSIPRED 2.0 achieved an average Q3 score of 80.6% across all 40 submitted target domains with no obvious sequence similarity to structures present in PDB, which placed PSIPRED in first place out of 20 evaluated methods (an earlier version of PSIPRED was also ranked first in CASP3 held in 1998).

\[
Q_3 = 100 \cdot \frac{1}{N_{res}} \cdot \sum_{i=1}^{3} M_{ij}
\]

\[
M_{ij} = \text{number of residues observed in state } i \text{ and predicted in state } j, \text{ with } i \text{ and } j \in \{H, E, L\}
\]


**Fig. 1.** PSIPRED graphical output from prediction of methylglyoxal synthase (CASP3 target ‘T0081’) produced by PSIPREDView—a Java visualization tool that produces two-dimensional graphical representations of PSIPRED predictions.
Rotamer side chain placement

Loop sampling
Ab initio protein-structure prediction

- Search fold space to generate plausible candidates, or thread sequence through known folds
- Evaluate candidate folds for protein-like properties (density/compactness, residue-residue contacts (statistical scoring function), energetic evaluations)
- Refine best candidate folds by molecular modeling
Ab initio protein-structure prediction

- Search fold space to generate plausible candidates, or thread sequence through known folds
- Evaluate candidate folds for protein-like properties (density/compactness, residue-residue contacts (statistical scoring function), energetic evaluations)
- Refine best candidate folds by molecular modeling
Contact matrix folding

- If Cα of residue $i$ is within 8 Å of Cα of residue $j$, they are in contact and a 1 is placed at position $ij$ of contact matrix.

- Secondary structural elements can be predicted and characteristic patterns entered into initial matrix.

- Frequencies of contacts between amino acid types are used to optimize matrix.

- Matrices can be restricted to protein-density space.

- The correct 3D Cα trace can be regenerated from a correct contact matrix.

- Fold prediction done in limited 2D contact matrix space vs. 3D coordinate space.
Residue–residue contact matrix for predicted 3D structure of 3c2c (blue and green lines). The constant part $A_c$ is shown in red, the “noncontact” matrix $A_n$, is shown in green, and predicted variable contacts $A_x$, are shown in blue.

- Numbers correspond to the predicted loops
SCORING FUNCTION FOR DE NOVO PROTEIN STRUCTURE PREDICTION

Efficient Method for Fold Sampling Based on Contact Matrix Developed


Status 1999 - For 100 residue proteins, ten predicted folds usually contained one within 4 Å CRMS.

Question – Which One??
Ab initio (New Fold methods)

- Assumption
  The structure that a protein folds is the structure with the lowest global free energy (or a structure very similar to it)

- Finding native-like protein conformations requires developing:
  - an accurate potential function that permits calculation of the free energy given a structure
  - an efficient method for searching for energy minima

Molecular dynamics is not the answer due to constraints on simulation time, but it may work on very small systems (Trp cage, etc.)
Protein Folding

Simulated folding in 1 μsec; peptide in box of water
**Relevant timescales**

- **Bond vibration**
- Isomerisation
- Water dynamics
- Helix forms
- Fastest folders
- Typical folders
- Slow folders

- $10^{-15}$ femtoseconds
- $10^{-12}$ picoseconds
- $10^{-9}$ nanoseconds
- $10^{-6}$ microseconds
- $10^{-3}$ milliseconds
- $10^{0}$ seconds

- **16 order of magnitude range**
  - Femtosecond timesteps
  - Need to simulate micro to milliseconds
peptide backbone

<table>
<thead>
<tr>
<th>Conformation</th>
<th>$\phi$</th>
<th>$\psi$</th>
<th>$\omega$</th>
<th>Res / turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$-helix</td>
<td>-57</td>
<td>-47</td>
<td>180</td>
<td>3.6</td>
</tr>
<tr>
<td>$3_{10}$ helix</td>
<td>-49</td>
<td>-26</td>
<td>180</td>
<td>3.0</td>
</tr>
<tr>
<td>$\pi$-helix</td>
<td>57</td>
<td>-70</td>
<td>180</td>
<td>4.4</td>
</tr>
</tbody>
</table>
What is the average structure of a helix in highest-resolution protein structures? Do torsional angles center around classical $\alpha$-helical values of $\Phi = -57^\circ, \Psi = -47^\circ$ or $3_{10}$-helical values of $\Phi = -67^\circ, \Psi = -30^\circ$?


Problem: Only examine high-resolution crystal structures where carbonyl oxygen is clearly visible in electron density. Assign hydrogen bonds of backbone atoms. If a residue has a combination of at least two $i\rightarrow i+3$ or $i\rightarrow i+4$ H-bonds in a row, calculate the phi and psi angles. Display frequency of occurrence of helical phi and psi angles at 1$^\circ$ resolution.

Results - Over 900 helical residues in PDB in crystal structures with 1.0 to 1.4 Å resolution clustered at $\Phi = -60^\circ, \Psi = -41^\circ$.

Similar to values $\Phi = -63.8^\circ, \Psi = -41.1^\circ$ found by Hovmöller et al., Acta Cryst D58:768-776 (2002).

The Devil's in the details!
Frequency of \( \phi, \psi \) values for helical residues in high-resolution structures in the PDB, Kuster and Marshall, unpublished

November 11, 2008

Bio5476 - Garland R. Marshall
Agreement with Theoretical Results Using Next-Generation Force Field

The absence of polarizability and the use of monopole electrostatics introduces significant errors in molecular mechanics. AMOEBA is a next generation force field parameterized for multipole electrostatics and polarizability that can reproduce geometry of water clusters as well as bulk properties of water (not possible with current modeling packages).

Helical residues 7-18 of crambin (46 residues, PDB = 1EJG, resolution = 0.54 Å) after long MD simulations in explicit water.

Results using three different force fields:

AMOEBA - helix agrees better with experimental X-ray data, minimum shifted toward 3_{10} helix;

CHARMM and OPLS-AA - helix significantly shorter than experimental, minima close to classical helical values.

Preliminary data from Sergio Urahata and Jay Ponder (CCB, WUSTL)

The Devil's in the details!

November 11, 2008

Bio5476 - Garland R. Marshall
Simplification of Problem

A. Fold on lattice - limit conformational space to be explored
B. Reduce complexity of molecule - united atom (only polar atoms have explicit hydrogens)
C. Heuristics - what rules can we deduce from known structures?