Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

Secondary Structure: Computational Problems

Secondary structure characterization
Secondary structure assignment
Secondary structure prediction

Protein structure classification

Structural classes of proteins

α

β

α/β

Protein Structure Classification

SCOP - Structural Classification of Proteins
FSSP - Fold classification based on Structure-Structure alignment of Proteins
CATH - Class, architecture, topology and homologous superfamily

SCOP: Structural Classification of Proteins

Current release: 1.75
38221 PDB Entries (June 2009). 110800 Domains.

http://scop.mrc-lmb.cam.ac.uk/scop/

The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy; the principal levels are family, superfamily and fold.

Family: Clear evolutionary relationship
Superfamily: Probable common evolutionary origin
Fold: Major structural similarity

SCOP: Structural Classification of Proteins

Family: Clear evolutionary relationship

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, 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%.
SCOP: Structural Classification of Proteins

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 Statistics

<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>284</td>
<td>307</td>
<td>871</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>174</td>
<td>354</td>
<td>742</td>
</tr>
<tr>
<td>Alpha/beta proteins (a/b)</td>
<td>147</td>
<td>244</td>
<td>803</td>
</tr>
<tr>
<td>Alpha+beta proteins (a+b)</td>
<td>376</td>
<td>552</td>
<td>1055</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>66</td>
<td>66</td>
<td>89</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>58</td>
<td>110</td>
<td>123</td>
</tr>
<tr>
<td>Small proteins</td>
<td>90</td>
<td>129</td>
<td>219</td>
</tr>
<tr>
<td>Total</td>
<td>1195</td>
<td>1962</td>
<td>3902</td>
</tr>
</tbody>
</table>

Protein Structure Hierarchy

- Primary - the sequence of amino acid residues
- Secondary - ordered regions of primary sequence (helices, beta-sheets, turns)
- Tertiary - the three-dimensional fold of a protein subunit
- Quaternary - the arrangement of subunits in oligomers

Anfinsen's Dogma

Three-dimensional structure of a protein is determined solely by its amino-acid sequence.

Native conformation of the protein is the global-minimum free energy conformation.

Levinthal paradox

3 conformations per residue is a very conservative estimate
Complexity of protein structure
(Levinthal paradox)

100 residue protein
3 conformations per residue

number of distinct conformations:
\[3^{100} \approx 10^{48}\]
sampling time \(\approx 10^{30}\) years

Complexity

P (Polynomial)
complexity class of decision problems for which execution time of a
computation is no more than a polynomial function of the problem size

NP (Nondeterministic Polynomial)
complexity class of decision problems for which answers can be checked
by an algorithm whose run time is polynomial in the size of the input

Protein Folding Problem

Given: sequence
Find: structure

The problem is NP-complete

Protein Folding Problem

Problem for us, not for proteins.
They just fold...

(Ken Dill)

Dynamics of Database Growth

Protein Structure Determination

X-ray crystallography
NMR spectroscopy
Neutron diffraction
Electron microscopy
Atomic force microscopy
X-ray crystallography

Bragg’s Law

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

Phase determination: MIR and MAD
(Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction)

Fourier Transforms

X-ray crystallography

Electron density map created from multi-wavelength data (Arg)

Experimental electron density map and model fitting
(apoE four helix bundle)
X-ray crystallography

<table>
<thead>
<tr>
<th>Structural feature</th>
<th>5 Å</th>
<th>3 Å</th>
<th>2.5 Å</th>
<th>2.0 Å</th>
<th>1.5 Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain tracing</td>
<td>—</td>
<td>—</td>
<td>Fair</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Secondary structure</td>
<td>Helix</td>
<td>Fair</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Sidechain conformations</td>
<td>—</td>
<td>—</td>
<td>Fair</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Orientation of peptide planes</td>
<td>—</td>
<td>—</td>
<td>Fair</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Protein hydrogen atoms visible</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Good</td>
</tr>
</tbody>
</table>