**Protein Structure: Data Bases and Classification**

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

---

**Reference**

Bourne and Weissig
Structural Bioinformatics
Wiley, 2003

---

**More References**
### Terminology

- Primary Structure
- Secondary Structure
- Tertiary Structure
- Quaternary Structure
- Supersecondary Structure
- Domain
- Fold

### Hierarchy of Protein Structure

![Diagram of protein structure hierarchy]

### Helices

<table>
<thead>
<tr>
<th>Helix Type</th>
<th>Amino acids/turn</th>
<th>Frequency</th>
<th>H-bonding</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>3.6</td>
<td>~97%</td>
<td>i, i+4</td>
</tr>
<tr>
<td>3.10</td>
<td>3.0</td>
<td>~3%</td>
<td>i, i+3</td>
</tr>
<tr>
<td>π</td>
<td>4.4</td>
<td>rare</td>
<td>i, i+5</td>
</tr>
</tbody>
</table>
\(\alpha\)-helices have handedness:

\(\alpha\)-helices have a dipole:

\(\beta\)-sheets

(a) Antiparallel

(b) Parallel
\( \beta \)-sheets

(A) Have a right-handed twist!

(B) Can form higher level structures!

Super Secondary Structure Motifs

\( \beta \)-\( \alpha \)-\( \beta \) Loop  \( \alpha / \beta \) Barrel
What is a Domain?

Richardson (1981):

Within a single subunit [polypeptide chain], contiguous portions of the polypeptide chain frequently fold into compact, local semi-independent units called domains.

More About Domains

- Independent folding units.
- Lots of within contacts, few outside.
- Domains create their own hydrophobic core.
- Regions usually conserved during recombination.
- Different domains of the same protein can have different functions.
- Domains of the same protein may or may not interact.

Why Look for Domains?

Domains are the currency of protein function!
Domain Size

- Domains can be between 25 and 500 residues long.
- Most are less than 200 residues.
- Domains can be smaller than 50 residues, but these need to be stabilized.

Examples are the zinc finger and a scorpion toxin.

Two Very Small Domains

A Humdinger of a Domain
<table>
<thead>
<tr>
<th>What’s the Domain? (Part 1)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>What’s the Domain? (Part 2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Homology and Analogy</th>
</tr>
</thead>
</table>

- **Homology:** Similarity in characteristics resulting from shared ancestry.
- **Analogy:** The similarity of structure between two species that are not closely related, attributable to convergent evolution.

Homologous structures can be divided into orthologues (result from changes in the same gene between different organisms, such as myoglobin) and paralogues (result from gene duplication and subsequent changes within an organism and its descendants, such as hemoglobin).
Homology and Analogy
PDB File Header

The header contains information about protein and structure, date of the entry, references, crystallographic data, contents and positions of secondary structure elements, etc:

PDB File Body

The body of the PDB file contains information about the atoms in the structure:

ATOM  76  N   PRO A  12      31.129  -4.659  43.245  1.00  9.00           N
ATOM  77  CA  PRO A  12      32.426  -4.662  42.542  1.00  9.00           C
ATOM  78  C   PRO A  12      32.423  -4.009  41.182  1.00  8.02           C
ATOM  79  O   PRO A  12      33.267  -3.177  40.892  1.00  8.31           O
ATOM  80  CB  PRO A  12      32.791  -6.126  42.592  1.00 10.02           C
ATOM  81  CG  PRO A  12      32.190  -6.663  43.857  1.00 10.12           C
ATOM  82  CD  PRO A  12      30.850  -5.927  43.925  1.00  9.87           C
ATOM  90  N   ALA A  13      31.485  -4.468  40.316  1.00  8.06           N
ATOM  91  CA  ALA A  13      31.357  -3.854  39.004  1.00  7.28           C
ATOM  92  C   ALA A  13      29.947  -3.309  38.814  1.00  7.21           C
ATOM  93  O   ALA A  13      28.969  -3.932  39.200  1.00  7.56           O
ATOM  94  CB  ALA A  13      31.636  -4.879  37.897  1.00  8.54           C
What do you want to do?

- Call sequences from the whole EST to evolution, sequences known, B, etc.
- Call new own list of EST names
- Call one new EST file on the file system on another level. For instance, you can paste the last line of the top of ELAST output, we can go get the whole sequence from this
- Call new EST file of sequences in FASTA format or from ELAST output i.e., we use the fragment of sequence from the Output box in the ELAST output which we will upload.

Choose your desired thresholds:

- Minimum percentage identity
- Minimum evolution
- Minimum evidence
- Minimum S-value
- Minimum start length
- Minimum start height
- Skip any low scores?

[Images of a computer screen with lists and options]
SCOP
Structural Classification of Proteins

• Proteins are classified (manually) taking both the structural and evolutionary relationship into account.
• There are 7 classes of proteins, the main ones being all alpha, all beta, alpha/beta, and alpha+beta.
• The principle levels in the hierarchy of SCOP are fold, superfamily, and family.


SCOP Levels

• **Family**: Clear evolutionarily relationship. In general >30% pairwise residue identities between the proteins.

• **Superfamily**: Probable common evolutionary origin. Proteins have low sequence identities, but structural and functional features suggest that a common evolutionary origin is probable.

• **Fold**: Major structural similarity. Proteins have the same major secondary structures in same arrangement and with the same topological connections.
Some maybe surprising results

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5NLL</td>
<td>1AMO</td>
<td>1CHN</td>
<td>1FNB</td>
</tr>
</tbody>
</table>

Flavodoxin      Cytochrome reductase  Protein CHEY  Ferredoxin reductase

**CATH**

Protein Structure Classification

- The CATH database is a hierarchical domain classification of protein structures in the Brookhaven protein database. Only NMR structures and crystal structures solved to resolution better than 3.0 angstroms are considered.
- There are four major levels in this hierarchy: Class, Architecture, Topology (fold family) and Homologous superfamily.
- Multidomain proteins are subdivided into their domains using a consensus procedure. All the classification is performed on individual protein domains.

The CATH Hierarchy

SCOP versus CATH

The CATH Hierarchy

SCOP versus CATH

The CATH Hierarchy

SCOP versus CATH
DALI
Distance Matrix Alignment

• DALI generates alignments of structural fragments, and is able to find alignments involving chain reversals and different topologies.
• The algorithm uses distance matrices to represent each structure to be compared.
• Application of DALI to the entire PDB produces two classifications of structures: FSSP and DDD (3D).

Holm L, and Sander C (1993)
FSSP and DDD

- The families of structurally similar proteins (FSSP) is a database of structural alignments of proteins in the protein data bank (PDB). It presents the results of applying DALI to (almost) all chains of proteins in the PDB.
- The DALI domain dictionary (DDD) is a corresponding classification of recurrent domains automatically extracted from known proteins.

Other Algorithms for Domain Decomposition

- The Protein Domain Parser (PDP) uses compactness as a chief principle.
  
  http://123d.ncl.fcrf.gov/pdp.html

- DomainParser is graph theory based. The underlying principle used is that residue-residue contacts are denser within a domain than between domains.
  
  http://compbio.ornl.gov/structure/domainparser/

Oh Dear...
Parsing Sequence into Domains

- Look for internal duplication.
- Look for low complexity segments.
- Look for transmembrane segments.

Why is That Important?

- Functional insights.
- Improved database searching.
- Fold recognition.
- Structure determination.

PFAM: http://www.sanger.ac.uk/Software/Pfam/

I-Sites