Protein Structure: Data Bases and Classification

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

A Foine Reference

Bourne and Weissig
Structural Bioinformatics
Wiley, 2003
Terminology

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

Hierarchy of Protein Structure
**Helices**

- \( \alpha \)
- 3.10
- 97%
- Rare

Amino acids/turn:
- 3.6
- 3.0
- 4.4

Frequency:
- ~97%
- ~3%
- Rare

H-bonding:
- \( i, i+4 \)
- \( i, i+3 \)
- \( i, i+5 \)

---

**\( \alpha \)-helices**

- Amino terminus
- Carboxyl terminus
- 5.4 Å (3.6 residues)
- (a) (b) (c) (d)
**α-helices**

α-helices have handedness:  

α-helices have a dipole:

**β-sheets**

(a) Antiparallel  
Top view  
Side view  

(b) Parallel  
Top view  
Side view
β-sheets

(A) Have a right-handed twist!

(B) Can form higher level structures!
Super Secondary Structure Motifs

β-α-β Loop  α/β Barrel

Protein Structure and Function

Cytochrome c  Lysozyme  Ribonuclease
Structural Proteins

Membrane Proteins
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

What’s the Domain? (Part 1)
What’s the Domain? (Part 2)

Homology and Analogy

- 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 (a result from changes in the same gene between different organisms, such as myoglobin) and paralogues (a result from gene duplication and subsequent changes within an organism and its descendents, such as hemoglobin).
The RCSB Protein Data Bank

The Calcium Pump

Explore the Structure

Structure Explorer - 4MDH

Summary Information

Title: Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5-A resolution.

Component: Cytoplasmic Malate Dehydrogenase (EC 1.1.1.37)

Authors: J. J. Birkett, L. J. Bannister

Exp. Method: X-ray Diffraction

Classification: Oxidoreductase(NADP+/NADH)-Malate dehydrogenase

ICD Number: 1.1.1.37 (Malate dehydrogenase)

Sources: S. cerevisiae

Primary Citation: Birkett, J. J., Rhodes, G., Banister, L. J. Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5-A resolution. Biochemistry 28, 6005 (1989)

Deposition Date: 12-Apr-1989

Release Date: 19-Apr-1989

Resolution (Å): 1.50

Space Group: P2_1 2_1 2

Unit Cell:

\[ a = 139.30 \quad b = 86.60 \quad c = 58.80 \]

\[ \alpha = 90.00 \quad \beta = 90.00 \quad \gamma = 90.00 \]
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 Header**

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

**PDB File Body**

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

Unique Folds in the PDB
New Folds Become Rare

SCOP Structural Classification of Proteins

- Proteins are classified (manually!) taking both 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 are fold, superfamily, and family.

Hubbard, Murzin, Brenner and Chothia (1997)
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

<table>
<thead>
<tr>
<th></th>
<th>5NLL</th>
<th>1AMO</th>
<th>1CHN</th>
<th>1FNB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flavodoxin</td>
<td>Cytochrome reductase</td>
<td>Protein CHEY</td>
<td>Ferredoxin reductase</td>
<td></td>
</tr>
</tbody>
</table>

CATH
Protein Structure Classification

- The CATH database is a hierarchical domain classification of protein structures in the Brookhaven protein databank. 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.

Orengo, Michie, Jones, Jones, Swindells, Thornton (1997)
The CATH Hierarchy

SCOP versus CATH

<table>
<thead>
<tr>
<th>Correspondence between SCOP and CATH hierarchies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class</strong></td>
</tr>
<tr>
<td><strong>Architecture</strong></td>
</tr>
<tr>
<td><strong>Topology</strong></td>
</tr>
<tr>
<td><strong>Sequence family</strong></td>
</tr>
</tbody>
</table>

- TIM barrel
- Sandwich
- Roll
- Flavodoxin (4fxn)
• 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).
Unaligned:

1bbo  1 KYICFEDGNIKMKEXLELGGSTTVTPFYQCTCNSEKTNGLNRESPREMKK  57
3drpA 103 FTKROGCTYRCVKVSQYTHMLSNPCVYTSRHSVSVVYEPFCKKEPTKRDNMAFVFLIF  165

Aligned:

1bbo  1 ......KYICFEDGNIKMKEXLELGGSTTVTPFYQCTCNSEKTNGLNRESPREMKK  57
3drpA 103 f:kegehTVLAVAVCSQYTHMLSNPCVYTHSVSNTSVYEPFCKKEPTKRDNMAFVFLIF... 165
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.

References: Holm and Sander


Other Algorithms for Domain Decomposition

- The Protein Domain Parser (PDP) uses compactness as a chief principle.  
  http://123d.ncifcrf.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.

PRODOM:

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