Proteins
Classification, Databases, and Visualization

Terminology

• Primary Structure
• Secondary Structure
• Tertiary Structure
• Quaternary Structure
• Supersecondary Structure
• Folds
• Domains

Bourne and Weissig
Structural Bioinformatics
Wiley, 2003
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.
Homology: Similarity in characteristics resulting from shared ancestry.

Homologous structures: Structures in different species that are similar because of common ancestry.

Analogy: The similarity of structure between two species that are not closely related; attributable to convergent evolution.

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 α, all β, α/β, and α+β.
- The principle levels in the hierarchy are fold, superfamily, family.

Levels

- Family: Clear evolutionary 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.

Hubbard, Murzin, Brenner and Chothia (1997)
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)

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).

References: Holm and Sander (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.

References: Holm and Sander

• The FSSP Database of Structurally Aligned Protein Fold Families, Nucleic Acids Research 22 (17), pp 3600-3609, 1994.
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/

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/