

SCOP: Structural Classification of Proteins

Alexey Murzin et al.

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

Proteins are classified to reflect both structural and evolutionary relatedness.

Many levels exist in the hierarchy, but the principal levels are family, superfamily and fold, described below.

The exact position of boundaries between these levels are to some degree subjective.

The evolutionary classification is generally conservative: where any doubt about relatedness exists, new divisions were made at the family and superfamily levels. Thus, some researchers may prefer to focus on the higher levels of the classification tree, where proteins with structural similarity are clustered.

Graham Kemp, Chalmers University of Technology

SCOP 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.

Graham Kemp, Chalmers University of Technology

SCOP Fold: Major structural similarity

Proteins are defined as having a common fold if they have 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.

In some cases, these differing peripheral regions may comprise half the structure.

Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

Graham Kemp, Chalmers University of Technology

SCOP 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.

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%.

Graham Kemp, Chalmers University of Technology

CATH — Protein Structure Classification

Christine Orengo et al.

<http://cathwww.biochem.ucl.ac.uk/latest/>

CATH: Class

Class is determined according to the secondary structure composition and packing within the structure.

It can be assigned automatically for over 90% of the known structures using the method of Michie et al. (1996). For the remainder, manual inspection is used and where necessary information from the literature taken into account.

Three major classes are recognised; mainly-alpha, mainly-beta and alpha-beta. This last class (alpha-beta) includes both alternating alpha/beta structures and alpha+beta structures, as originally defined by Levitt and Chothia (1976). A fourth class is also identified which contains protein domains which have low secondary structure content.

Graham Kemp, Chalmers University of Technology

CATH: Topology (Fold family)

Structures are grouped into fold families at this level depending on both the overall shape and connectivity of the secondary structures. This is done using the structure comparison algorithm SSAP (Taylor & Orengo (1989)). Parameters for clustering domains into the same fold family have been determined by empirical trials throughout the databank (Orengo et al. (1992), Orengo et al. (1993)). Structures which have a SSAP score of 70 and where at least 60% of the larger protein matches the smaller protein are assigned to the same T level or fold family.

Some fold families are very highly populated (Orengo et al. (1994)) particularly within the mainly-beta 2-layer sandwich architectures and the alpha-beta 3-layer sandwich architectures. In order to appreciate the structural relationships within these families more easily, they are currently subdivided using a higher cutoff on the SSAP score (75 for some mainly-beta and alpha-beta families, 80 for some mainly-alpha families, together with a higher overlap requirement (70%)).

Graham Kemp, Chalmers University of Technology

CATH: Architecture

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures.

It is currently assigned manually using a simple description of the secondary structure arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g. the beta-propeller or alpha four helix bundle). Procedures are being developed for automating this step.

Graham Kemp, Chalmers University of Technology

CATH: Homologous Superfamily

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified first by sequence comparisons and subsequently by structure comparison using SSAP. Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:

- Sequence identity $\geq 35\%$, 60% of larger structure equivalent to smaller
- SSAP score ≥ 80.0 and sequence identity $\geq 20\%$
60% of larger structure equivalent to smaller
- SSAP score ≥ 80.0 , 60% of larger structure equivalent to smaller, and domains which have related functions

Graham Kemp, Chalmers University of Technology

PROCHECK

Roman Laskowski et al.

<http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>

"The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.

Unusual regions highlighted by PROCHECK are not necessarily errors as such, but may be unusual features for which there is a reasonable explanation (eg distortions due to ligand-binding in the protein's active site). Nevertheless they are regions that should be checked carefully."

Graham Kemp, Chalmers University of Technology

Checks performed by PROCHECK (2)

Dihedral angle checks

Ramachandran plot shows phi-psi distribution.

Each residue is classified according to its region: "core", "allowed", "generous", or "disallowed".

Residues in the generous and disallowed regions are highlighted on the plot.

A log-odds score shows how normal or unusual the residue's location is on the Ramachandran plot for the given residue type.

Chirality check

Provides a measure of the C-alpha tetrahedral distortion. Measured by the notional zeta torsion angle, defined by the atoms C-alpha, N, C and C-beta.

The expected value is 33.9 degrees.

A negative value signifies a D-amino acid.

Graham Kemp, Chalmers University of Technology

Checks performed by PROCHECK (1)

Covalent geometry checks

- Main-chain bond lengths:
Compared against the Engh & Huber small-molecule data.
Differences greater than 0.05Å are highlighted.
- Main-chain bond angles:
Compared against the Engh & Huber small-molecule data.
Differences greater than 10 degrees are highlighted.

Planarity checks

- Aromatic rings (Phe, Tyr, Trp and His)
RMS distances greater than 0.04Å from best-fit plane highlighted.
- End-groups (Arg, Asn, Asp, Gln, Glu)
RMS differences greater than 0.03Å from best-fit plane are highlighted.

Graham Kemp, Chalmers University of Technology

Checks performed by PROCHECK (3)

Non-bonded interactions check

Any two non-bonded atoms are deemed to make a bad contact if they are as close as 2.6Å apart.

Possible hydrogen-bonding partners are excluded by ignoring all atom-pairs where one of the atoms is a possible H-bond donor (eg a main-chain nitrogen) and the other is a possible H-bond acceptor (eg a water molecule, or a main-chain oxygen).

Main-chain hydrogen bonds check

A check is made of main-chain hydrogen-bond energies, calculated using the Kabsch & Sander (1983) method.

Significant deviations from the ideal value of -2.0 kcal/mol are highlighted.

Graham Kemp, Chalmers University of Technology

Checks performed by PROCHECK (4)

Disulphide bond checks

The S-S separation in each disulphide bond is compared with the ideal distance of 2.0Å

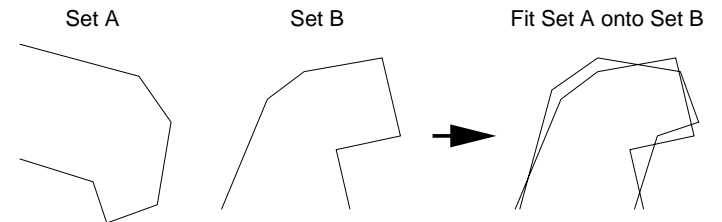
The chi-3 torsional angle, defined by the S-S bridge, is compared against the ideal values:

- 85.8 degrees for a left-handed conformation
- 96.8 degrees for a right-handed conformation

Significant deviations from the ideal values are highlighted.

Various other stereochemical parameters are computed and compared with values from well-refined structures.

Superposing structures



- 3-D transformation to map Set A onto Set B
- Root Mean Square (RMS) distance

Structural clues from multiple sequence alignments

- Residues at highly conserved positions often have important functional or structural roles.
- Insertions and deletions can be accommodated most easily in surface loops.
- Conserved patterns of hydrophobic residues can suggest secondary structure.
- The root mean square deviation between pairs of homologous proteins generally increases as the percent residue identity decreases.

“HSSP-curve”

- Shows the length-dependent threshold for significant sequence identity.
- Proposed by Sander and Schneider (1991) and revised by Rost (1999).
- Above the curve, identifying true positives is easy.
- Just below the curve, the number of false positives rises rapidly; distinguishing between true and false positives in the “twilight zone” is difficult.

(HSSP stands for “Homology-derived Secondary Structure of Proteins”)