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.

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.

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.

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.

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.

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 hexakinase together form a superfamily.

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.

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 absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

CATH — Protein Structure Classification

Christine Orengo et al.

http://www.cathdb.info/

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.

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%)).

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-propellor or alpha four helix bundle). Procedures are being developed for automating this step.

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 >= 35%, 60% of larger structure equivalent to smaller
- SSAP score >= 80.0 and sequence identity >= 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

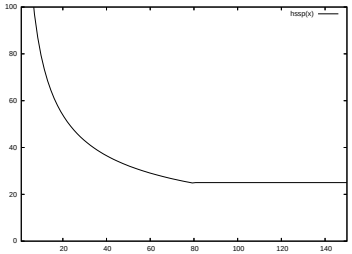
Protein stability

- good stereochemistry; no steric clashes;
- buried charged atoms must be paired;
- enough hydrophobic surface must be buried, and the interior must be sufficiently densely packed, to provide thermodynamic stability.

Modular proteins

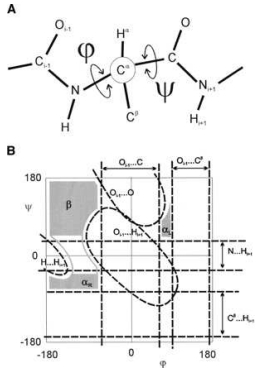
- multi-domain proteins, often with many copies of related domains;
- domains recur in many proteins in different structural contexts.

HSSP-curve



[Sander C. and Schneider, R., Proteins: Structure, Function and Genetics, 1991, 9:55-68]

Ramachandran steric map



[Ho, K.H., Thomas, A. and Brasseur, R., Protein Science, 2003, 12:2508-2522]

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, identifing 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”)

Comparative modelling and fold recognition

Comparative modelling (homology modelling):

Given:

- sequence of target protein with unknown structure
- known structure of a related protein

Predict:

- three-dimensional structure of target protein

Fold recognition:

Given:

- sequence of target protein with unknown structure
- library of known folds

Predict:

- known fold that is most compatible with the target protein's sequence

Secondary structure prediction

If neither sequence comparison nor fold recognition identifies a structure that can be used as a template for comparative modelling, then we can consider predicting secondary structure elements and how these might be assembled into a compact structure.

However, as noted by Ponder and Richards (1987):

> "a major problem lies in the secondary structure prediction itself ... the problem appears to lie in the non-negligible effect of long-range tertiary structural features upon secondary structure"

and

> "the problem of docking the preformed secondary units is formidable when considered in atomic detail."

Fold recognition

The idea behind "threading":

Imagine a wire wound into the shape of a known protein's main chain "fold".

Imagine next that our new sequence is represented by beads that are "threaded", in order, onto the wire, and are pushed along the wire.

At each step, a score is calculated based on which residues are adjacent in space, which residues are buried, etc.

Repeat this process for each different known fold.

A high score indicates that the sequence is compatible with that fold.

Heuristics for manual secondary structure prediction

- Many α-helices are amphipathic. Conserved hydrophobic residues at positions i, i+3, i+4, i+7, etc. are highly indicative of an α-helix.
- Half-buried strands will tend to have hydrophobic and hydrophilic residues at alternate positions.
- In proteins containing both α-helices and strands the strands are often completely buried and tend to contain only hydrophobic residues.

For more details and references, see:
http://www.russell.embl.de/gtsp/secstrucpred.html