The SCOP Database

Alex Watters

March 7, 2000

1 Structural Clustering

With the dramatically increasing amount of DNA/protein sequence information that is becoming available, methods to sort and make sense of this information have become extremely important. Prediction of function and placement in evolutionary trees of new proteins are of great interest. Blast/Psi-Blast utilize purely sequence information in an attempt to cluster proteins by evolutionary relatedness, with the assumption that evolutionary relatedness often signifies functional relatedness (to varying degrees). In the last 5 – 10 years databases utilizing structural information of proteins have been developed to aid in the clustering. One of the more prominent databases is SCOP (structural classification of proteins). The database is constructed on proteins whose structure has been determined (found in the Protein Data Bank, `www.rcsb.org/pdb/`).

2 The SCOP Database (`scop.mrc-lmb.cam.ac.uk/scop`)

SCOP places all proteins into a taxonomic hierarchy that is split into five basic levels.

1. Protein/Domain: This level comprises the most basic elements of comparison and each category in this level is composed of a single protein or protein domain. (A *domain* in a protein is defined as a compact component whose structure appears to be independent of the rest of the protein.) This level is similar to the species level in organism taxonomy.
2. Family: Proteins or domains are clustered together into families based on two pieces of information. Proteins/domains with 30% or greater sequence similarity are included in one family. In addition, a protein may be included into families where it shares at least 15% sequence identity with members of that family, as long as its function and structure are considered to be similar to other family members.
3. Superfamily: Families are then clustered into superfamilies based on structural and functional similarities. This level of clustering is not possible with sequence information alone and so these relationships are often not found by Blast or Psi-Blast.
4. Fold: Superfamilies are grouped based on the arrangement of their major secondary structural components and overall topology of these elements. Proteins clustered together at this level, but not at lower levels are likely to have significant structural differences on their periphery (e.g. varying loop lengths and loop structures). By this level most evolutionary and functional similarities have fallen apart to the extent that structural similarities are more likely due to physical constraints on a protein polymer than due to evolutionary relatedness.
5. Class: The Folds are finally defined by the their gross composition of basic secondary structural elements (i.e. all alpha helix, all beta sheet, alpha/beta, alpha + beta, and other).

3 Structural Similarity

The major question that arises from this is what do they mean by structural and functional similarity. Proteins are structurally clustered by optimal RMSD (root mean square deviation) alignments. Initially the structure of a protein is reduced to the coordinates of a central atom in each amino acid (the alpha carbon). An alignment algorithm is then used to find the optimal alignment between the two structures, by attempting to overlay one structure upon the other, thereby assigning each amino acid in one protein to a corresponding amino acid in the second protein. The optimal alignment is achieved when the RMSD of the distances between each amino acid pair is at a minimum. However, in SCOP this is only used as a crude clustering or approximation and all final clustering decisions on structural similarity are made by human visual inspection.

There are two major uses of this database. The first is visual inspection of the proteins within the hierarchy to gain an understanding of how plastic various structures are (i.e., how does the diversity of structural differences between clustered proteins depend on their location in the hierarchy?). The second is sequence submission. Sequences can be submitted to the database through the web site. These sequences will be clustered into the existing hierarchy based on sequence similarity. This allows one to examine the structure and function of related proteins. Such information can be useful in deciding what direction experimental characterization should go.

4 References

1. Brenner, S.E., Chothia, C., Hubbard, T.J.P, and Murizin, A. (1995) Methods Enzymol., 266, 635-653.
2. Murzin, A.G, Brenner, S.E., Hubbard, T., Chothia, C. (1995) J. Mol Biol., 247, 536-540.
3. Conte, L.L, Ailey, B., Hubbard, T.J.P., Brenner, S.E., Murzin, A.G., Chothia, C. (1999) Nuc. Acid . Res. 28, 257-259.