International Tables for Crystallography (2006). Vol. F, Section 23.1.1, pp. 575–576.

23. STRUCTURAL ANALYSIS AND CLASSIFICATION

23.1. Protein folds and motifs: representation, comparison and classification

By C. Orengo, J. Thornton, L. Holm and C. Sander

23.1.1. Protein-fold classification (C. Orengo and J. Thornton)

Since the first structure of myoglobin was solved in 1971, there has been an exponential growth in known protein structures with about 10 000 chains currently deposited in the Protein Data Bank (PDB; Abola *et al.*, 1987) and 200 or more solved each month. Since it is likely that the millennium will be marked by several international structural genomics projects, we can expect significant expansion of the data bank in the future. When dealing with such large numbers it is necessary to organize the data in a manageable and biologically meaningful way. To this end, several structural classifications have been developed [SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1997), DALI (Holm & Sander, 1999), 3Dee (Barton, 1997), HOMSTRAD (Mizuguchi *et al.*, 1998) and ENTREZ (Hogue *et al.*, 1996)], differing in their methodology and the degree of structural and functional annotation for the protein families identified.

Most public classification schemes have chosen to group proteins according to similarities in their domain structures, as this is generally considered to be the important evolutionary and folding unit. However, it can be difficult to identify domain boundaries either manually or using automatic algorithms, and although there are many methods available, a recent survey of these showed that even the most reliable algorithms only give the correct answer about 80% of the time (Jones *et al.*, 1998). Methods for recognizing domains are described in Section 23.1.2.

Most protocols used for clustering protein domain structures into families first identify similarities in their sequences. There are many well established methods for doing this, most based on dynamic programming algorithms, and since proteins with sequence identities of 30% or more are known to adopt very similar folds (Sander & Schneider, 1991; Flores *et al.*, 1993), it is relatively simple to cluster related proteins into evolutionary families on this basis. Very distant relatives ($<20\%$ sequence identity) are not easily identified by sequence alignment, but since structure is much more highly conserved during evolution, these relationships can be detected by comparing the 3D structures directly.

Various powerful algorithms have been developed for recognizing structurally related proteins (for reviews see Holm & Sander, 1994*a*; Brown *et al.*, 1996). These build on the rigid-body superposition methods of Rossmann & Argos (1975), which compare intermolecular distances after optimal translation and rotation of one protein structure onto the other. Other methods are based on the distance plots developed by Phillips (1970), which enable comparison of intramolecular distances between protein structures. In comparing very distantly related proteins, there are a number of problems which must be overcome. Insertions or deletions can obscure equivalent regions, though generally these appear in the loops between secondary structures. Residue substitutions can cause shifts in the orientations of the secondary structures in order to maintain optimal hydrophobic packing in the core.

A number of strategies have been developed for handling these problems. For example, some methods only consider secondary-structure elements, as these will contain fewer insertions. Artymiuk *et al.* (1989) represent secondary structures as linear vectors and use fast, efficient comparison algorithms based on graph theory. Others have adapted rigid-body methods to optimally superpose secondary structures, ignoring loops. Some methods chop the proteins being compared into fragments and then use various energy-minimization approaches (*e.g.* simulated annealing, Monte Carlo optimization) to link equivalent fragments in the two proteins. Such fragments can be identified by rigid-body superposition (Vriend & Sander, 1991) or, in the case of the DALI method (Holm & Sander, 1994*a*), by comparing contact maps for hexapeptide fragments. Several groups have modified the dynamic programming algorithms designed to cope with insertions or deletions in sequence comparison in order to compare three-dimensional (3D) information (Taylor & Orengo, 1989; Sali & Blundell, 1990; Russell & Barton, 1993). For example, the SSAP method of Taylor & Orengo (1989) uses double dynamic programming to align residue structural environments defined by vectors between Cβ atoms, whilst in STAMP (Russell & Barton, 1993), dynamic programming is used in an iterative procedure, together with rigid-body superposition.

Once equivalent residues have been found, the degree of structural similarity between two proteins can be measured in a number of ways, though the most commonly used is the root-mean-square deviation (RMSD), which is effectively the average ‘distance’ between superposed residues. However, there is still no

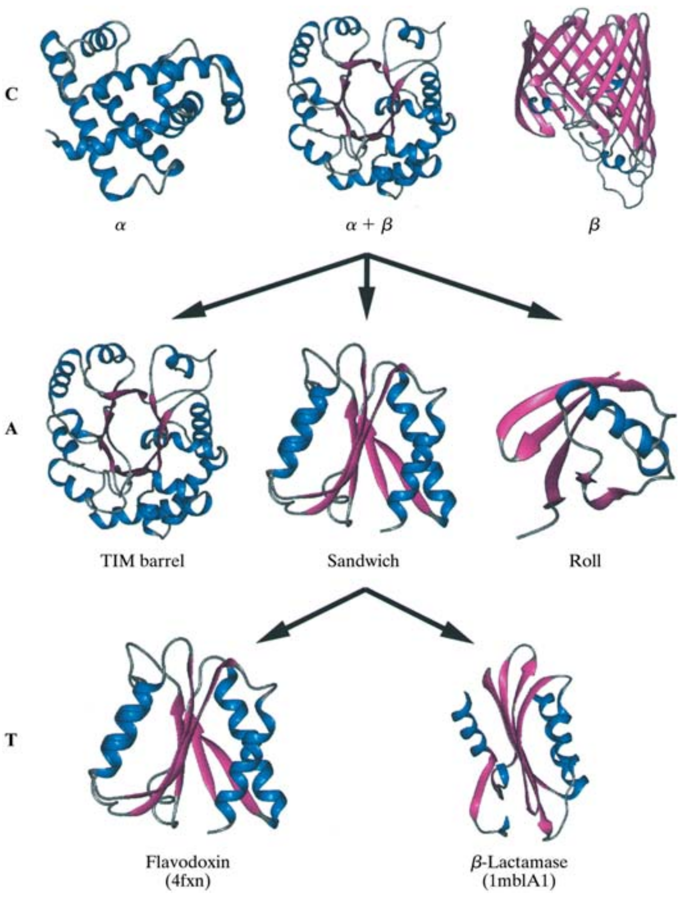

Fig. 23.1.1.1. Schematic representation of the (C)lass, (A)rchitecture and (T)opology/fold levels in the CATH database.

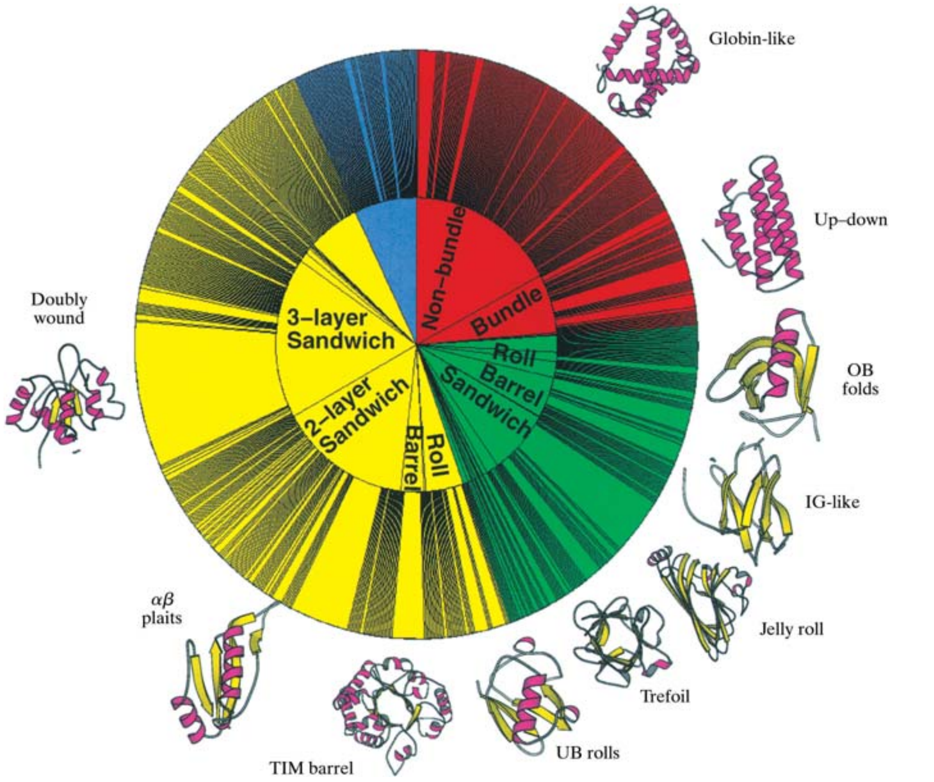

Fig. 23.1.1.2. 'CATHerine wheel' plot showing the distribution of non-homologous structures [*i.e.* a single representative from each homologous superfamily (H level) in CATH] amongst the different classes (C), architectures (A) and fold families (T) in the CATH database. Protein classes are shown coloured as red (mainly α), green (mainly β), and yellow (α–β). Within each class, the angle subtended for a given segment reflects the proportion of structures within the identified architectures (inner circle) or fold families (outer circle). *MOLSCRIPT* (Kraulis, 1991) illustrations are shown for representative examples from the superfold families.

consensus about which thresholds might imply homologous proteins or fold similarity between analagous proteins or common structural motifs. It is likely that this will become clearer as more structures are determined and the families become more highly populated, providing more information on tolerance to structural changes. These contraints will probably reflect functional requirements and/or kinetic or thermodynamic factors and will be specific to the family.

Several groups (Holm & Sander, 1999; Hogue *et al.*, 1996) attempt to determine the significance of a structural match by considering the distribution of scores for unrelated proteins and calculating a Z score. These approaches are very reliable for proteins possessing unusual structural characteristics but may not be as sensitive for those with highly recurring and common structural motifs. Other groups use empirical approaches (Orengo *et al.*, 1997) to establish reasonable cutoffs for identifying homologues, though these approaches obviously suffer from the currently limited size of the structure data bank.

Because of the individual strategies used to recognize relatives, the protein-structure classifications differ somewhat in their assignments. However, most classifications group proteins having highly similar sequences ($\geq 30\%$) into families. Subsequently, those families having highly similar structures and some other evidence of common ancestry [*e.g.* similar functions or some residual sequence identity (Orengo *et al.*, 1999)] are merged into homologous superfamilies. Families adopting similar folds, but where there is no other evidence to suggest divergent evolution, are usually put into the same fold group but are described as analogous proteins, since their similarity may simply reflect the physical and/or chemical constraints on protein folding.

SCOP and CATH are currently the largest of the public classifications, each with over 1000 homologous superfamilies. In SCOP (Murzin *et al.*, 1995), these families have been very carefully manually validated using biochemical information and by consideration of special structural features (*e.g.* rare β-bulges, left-handed helical connections) that may constitute evolutionary fingerprints; in CATH, homologues are validated both manually and automatically (Orengo *et al.*, 1997). Other databases [HOMSTRAD (Mizuguchi *et al.*, 1998); 3Dee (Barton, 1997)] contain similar groupings of protein structures, and there are multiple structural alignments for the family, annotated according to residue properties.

Several studies have suggested a limited number of folds available to proteins, with estimates ranging from one thousand to several thousand (Chothia, 1993; Orengo *et al.*, 1994), and this will mean an increasing number of analogous protein pairs being identified as the structural genomics initiatives continue. Recent analyses of the population of different fold families have revealed that some folds are more highly populated, perhaps because they fold more easily or are more stable. In the CATH database, ten favoured folds, described as superfolds, comprised very regular, layered architectures and were shown to contain a higher proportion of favoured motifs (*e.g.* Greek key, $\beta\alpha$ motif) than non-superfold structures. Similarly, analysis of SCOP (Brenner *et al.*, 1996) revealed some 40 or so frequently occurring domains (FODS), which included the superfolds. About one-third of all non-homologous structures ($< 25\%$ sequence identity to each other) adopt one of these folds.

Some groups avoid explicit definition of protein families. The DALI database of Holm & Sander (1999) is a neighbourhood scheme listing all related proteins for a given protein structure. Neighbours are identified using the DALI structure comparison algorithm (Holm & Sander, 1993) and range from the most highly similar, homologous proteins to those sharing only motif similarities. The ENTREZ database (Hogue *et al.*, 1996) provides a similar scheme, generated by the VAST structure comparison method of Gibrat *et al.* (1997). Both allow the user to assess significance and draw their own inferences regarding evolutionary relationships. More recently, the DALI domain database (DDD) (Holm & Sander, 1998) has provided clusters of related proteins based on calculated Z scores.

Most available databases further classify the fold groups on the basis of class. These agree with the major classes recognized by Levitt & Chothia (1976) (mainly α, mainly β, α/β, $\alpha + \beta$), although in the CATH database the α/β and $\alpha + \beta$ classes have been merged (Fig. 23.1.1.1). CATH also describes an intermediate architecture level between class and fold group (Orengo *et al.*, 1997). This refers to the arrangement of secondary-structure elements in 3D, regardless of their connectivity and so defines the shape (*e.g.* barrel, sandwich, propeller) (Fig. 23.1.1.2). There are currently 32 different architectures in CATH, with the simple barrel and sandwich shapes accounting for about 60% of the non-homologous structures.