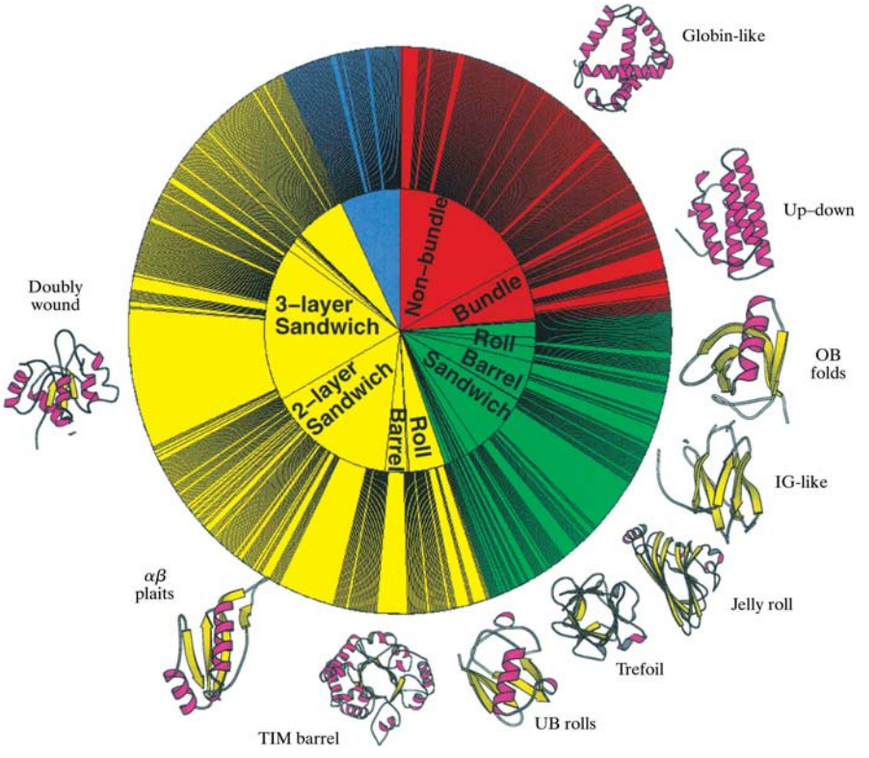

Fig. 23.1.1.2. 'CATHerine wheel' plot showing the distribution of non-homologous structures [*i.e.* a single representative from each homologous superfamily (H level) in CATH] amongst the different classes (C), architectures (A) and fold families (T) in the CATH database. Protein classes are shown coloured 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.

Copyright © 2006 International Union of Crystallography

references