International Tables for Crystallography (2006). Vol. F, Chapter 23.1, pp. 575–578.

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

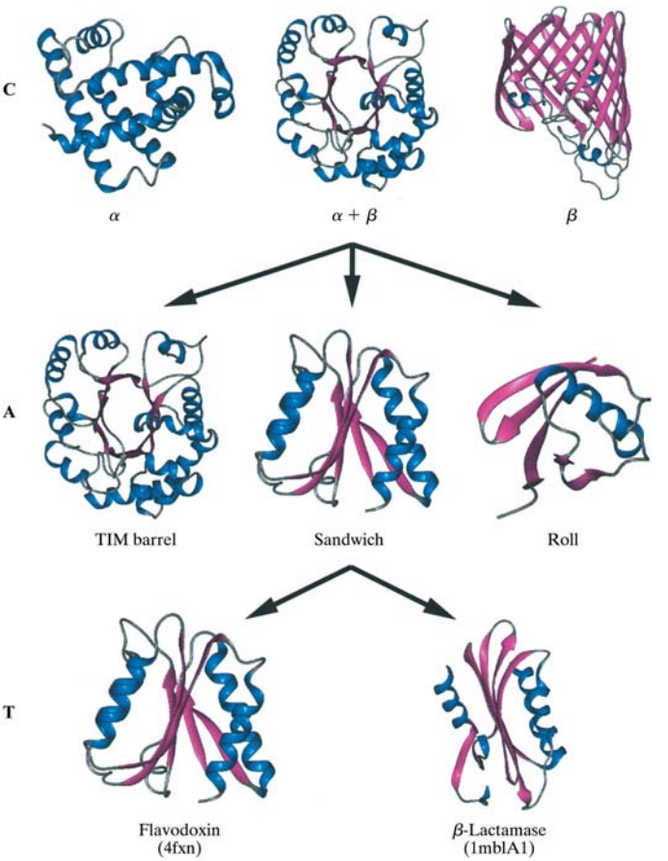

Fig. 23.1.1.1. Schematic representation of the (C)lass, (A)rchitecture and (T)opology/fold levels in the CATH database.

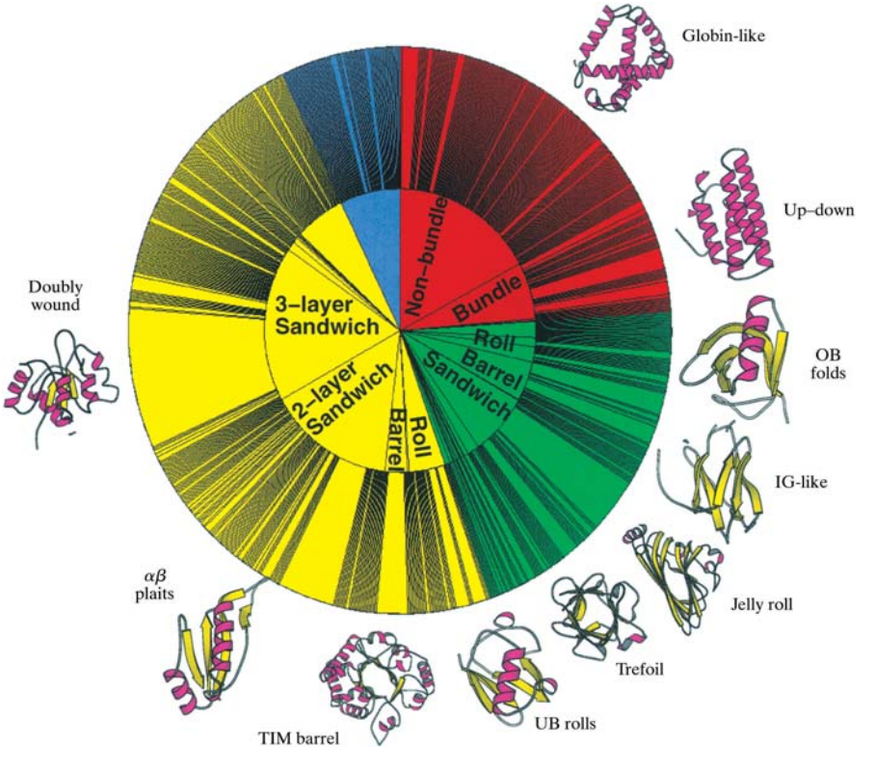

Fig. 23.1.1.2. 'CATHerine wheel' plot showing the distribution of non-homologous structures [*i.e.* a single representative from each homologous superfamily (H level) in CATH] amongst the different classes (C), architectures (A) and fold families (T) in the CATH database. Protein classes are shown coloured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

23.1.2. Locating domains in 3D structures

(L. HOLM AND C. SANDER)

23.1.2.1. *Introduction*

Modular design is beneficial in many areas of life, including computer programming, manufacturing, and even in protein folding.

Protein-structure analysis has long operated with the notion of domains, *i.e.*, dividing large structures into quasi-independent substructures or modules (Wetlaufer, 1973; Bork, 1992). In various contexts, these substructures are thought to fold autonomously, to carry specific molecular functions such as binding or catalysis, to move relative to each other as semi-rigid bodies and to speed the evolution of new functions by recombination (Fig. 23.1.2.1).

The problem of subdividing protein molecules into structural and functional units has received the attention of numerous researchers over the last 25 years. Early algorithms focused on protein folding or unfolding pathways and aimed at identifying substructures that would be physically stable on their own. Nowadays, with bulging macromolecular databases, the focus has shifted to devise automatic methods for identifying domains that can form the basis for a consistent protein-structure classification (Murzin *et al.*, 1995; Orengo *et al.*, 1997; Holm & Sander, 1999).

This review presents the concepts underlying computational methods for locating domains in 3D structures. Those interested in implementations are referred to the web services of the European Bioinformatics Institute* and related sites.

23.1.2.2. *Compactness*

A variety of ingenious techniques have been invented for locating structural domains in 3D structures. These include inspection of distance maps, clustering, neighbourhood correlation, plane cutting, interface area minimization, specific volume minimization, searching for mechanical hinge points, maximization of compactness and maximization of buried surface area (Rossmann & Liljas, 1974; Rashin, 1976; Crippen, 1978; Nemethy & Scheraga, 1979; Rose, 1979; Schulz & Schirmer, 1979; Go, 1981; Lesk & Rose, 1981; Sander, 1981; Wodak & Janin, 1981; Zehfus & Rose, 1986; Kikuchi *et al.*, 1988; Moult & Unger, 1991; Holm & Sander, 1994*b*; Zehfus, 1994; Islam *et al.*, 1995; Siddiqui & Barton, 1995; Swindells, 1995; Holm & Sander, 1996; Sowdhamini *et al.*, 1996; Zehfus, 1997; Holm & Sander, 1998; Jones *et al.*, 1998; Wernisch *et al.*, 1999).

Common to most approaches are the assumptions that folding units are compact and that the interactions between them are weak. These notions can be made quantitative, for example, by counting interatomic contacts and by locating domain borders by identifying groups of residues such that the number of contacts between groups is minimized. The hierarchic organization of putative folding units can be inferred starting from the complete structure and recursively cutting it (*in silico*) into smaller and smaller substructures. Alternatively, one may start from the residue or secondary-structure-element level and successively associate the most strongly interacting groups. The procedure involves two optimization problems.

The first optimization problem is algorithmic and concerns finding the optimal subdivisions. This problem is complicated by the possibility of the chain passing several times between domains (discontinuous domains). Without the constraint of sequential continuity, there is a combinatorial number of possibilities for dividing a set of residues into subsets (Zehfus, 1994). This hurdle has been overcome by fast heuristics (Holm & Sander, 1994*b*; Zehfus, 1997; Wernisch *et al.*, 1999).

The second optimization problem concerns formulating physical criteria that distinguish between autonomous and nonautonomous folding units, *i.e.*, defining termination criteria for recursive algorithms. Since compactness-related criteria do not have a clear bimodal distribution, domain-assignment algorithms (Holm & Sander, 1994*b*; Islam *et al.*, 1995; Siddiqui & Barton, 1995; Swindells, 1995; Sowdhamini *et al.*, 1996; Wernisch *et al.*, 1999) use cutoff parameters that have been fine-tuned against an external reference set of domain definitions.

23.1.2.3. *Recurrence*

Most fold classifications use a hierarchical model where evolutionary families are a subcategory of fold type and it is natural to assume that domain boundaries should be conserved in evolution. Consistency concerns lead to a reformulation of the goals of the domain-assignment problem, away from (imprecise) physical models of stable folding units and towards recognizing such units phenomenologically in the database of known structures through recurrence. The concept of recurrence has long been the cornerstone of domain assignments by experts based on visual inspection (Richardson, 1981). Recurrence means recognizing architectural units in one protein that have already been defined (named) in another.

The practical importance of domain identification is illustrated by the discov-

* EMBL–EBI (1995): http://www.ebi.ac.uk/; DALI domain dictionary (1999): http://www.ebi.ac.uk/dali/domain/; 3Dee – database of protein domain definitions (1997): http://barton.ebi.ac.uk/servers/3Dee.html.

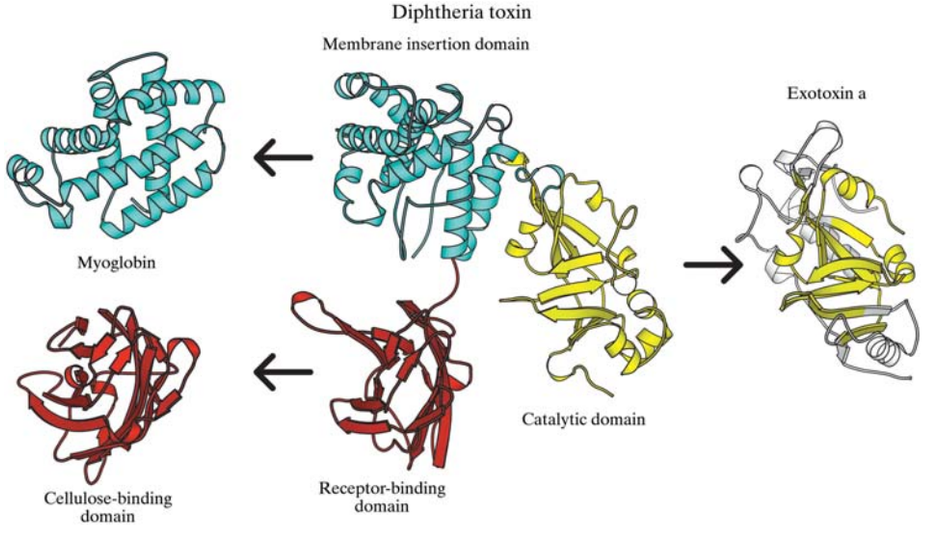

Fig. 23.1.2.1. The structure of diphtheria toxin (Bennett & Eisenberg, 1994) beautifully illustrates domains as structural, functional and evolutionary units. Structurally, note the compact globular shape of each domain and the flexible linkers between them. Functionally, note how each domain carries out a different stage of infection by the bacterium: receptor binding, membrane penetration and ADP-ribosylation of the target protein. Evolutionarily, note the occurrence of domains homologous to the catalytic domain of diphtheria toxin in exo-, entero- and pertussis toxins, and in poly-ADP-ribose polymerase (Holm & Sander, 1999). Arrows point to recurrent substructures in structural neighbours (Lionetti *et al.*, 1991; Li *et al.*, 1996; Tormo *et al.*, 1996) of each domain of diphtheria toxin. Drawn using *MOLSCRIPT* version 2 (Kraulis, 1991).

eries made by a systematic structure comparison of recurrent domains between histidine triad (HIT) proteins and galactose-6-phosphate uridylyltransferase [homodimer and internally duplicated common catalytic core, respectively (Holm & Sander, 1997)], and between beta-glucosyltransferase and glycogen phosphorylase [bare and heavily decorated common catalytic core, respectively (Holm & Sander, 1995; Artymiuk *et al.*, 1995)], even though the contours of the molecules look quite different.

Let us restate the goal of domain identification as an economic description of all known protein structures in terms of a small set of large substructures. This is an intuitive goal and conceptually related to the principle of minimal encoding in information theory. The key ingredients of the optimization problem are the gain associated with reusing a substructure and the cost associated with using many small substructures to describe a protein. An analogy in writing is that copying blocks of text is cheap, but for coherence some thought and effort is necessary for bridging the blocks.

With a suitably defined cost function, recurrence can be used to select an optimal set of substructures from the hierarchic folding or unfolding trees generated using compactness criteria. Thus, the unsatisfactorily solved problem of defining termination criteria for compactness algorithms can be turned into an optimization problem that does not rely on any external reference and leads to an internally consistent set of domain definitions.

The key difficulty is in quantifying the notion of economy so that it leads to a selection of substructures of 'appropriate' size, *i.e.*, globular folds and not, for example, supersecondary-structure motifs. One solution, which is physical nonsense but has the desired qualitative behaviour, is a heuristic objective function used in the DALI domain dictionary (Holm & Sander, 1998). Recurrence is quantified in terms of the statistical significance of structural similarity for many pairs of substructures. The statistical significance is highest for structural similarities that involve large units and that completely cover a substructure unit. Exploiting these effects, a sum-of-pairs objective function is defined that favours recurrences of large substructures with distinct topological arrangements and packing of secondary-structure elements, and disfavours small substructures consisting of one or two secondary-structure elements despite their higher frequency of recurrence. Though other formulations of the optimization problem are possible, this empirically chosen objective function combined with a heuristic algorithm for optimization yields a useful set of substructures (domains).

23.1.2.4. *Conclusion*

While we do not foresee that automatically delineated domains will be accepted as the gold standard of the trade, modern methods, based on a combination of recurrence and compactness criteria, yield domain definitions that are consistent within protein families and often coincide with biologically functional units, recover the well known folding topologies with many members, produce clusters with good coverage of common secondary-structure elements, and provide a useful basis for large-scale structure analysis and classification.

References

23.1

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Protein Data Bank*. In *Crystallographic databases – information content, software systems, scientific applications*, edited by F. H. Allen, G. Bergerhoff, & R. Sievers, pp. 107–132.

Artymiuk, P. J., Mitchell, E. M., Rice, D. W. & Willett, P. (1989). *Searching techniques for databases of protein structures*. *J. Inf. Sci.* **15**, 287–298.

Artymiuk, P. J., Rice, D. W., Poirrette, A. R. & Willett, P. (1995). *Beta-glucosyltransferase and phosphorylase reveal their common theme*. *Nature Struct. Biol.* **2**, 117–120.

Barton, G. J. (1997). *3Dee: database of protein domain definitions*. http://barton.ebi.ac.uk/servers/3Dee.html.

Bennett, M. J. & Eisenberg, D. (1994). *Refined structure of monomeric diphtheria toxin at 2.3 Å resolution*. *Protein Sci.* **3**, 1464–1475.

Bork, P. (1992). *Mobile modules and motifs*. *Curr. Opin. Struct. Biol.* **2**, 413–421.

Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. (1996). *Understanding protein structure. Using SCOP for fold interpretation*. *Methods Enzymol.* **266**, 635–643.

Brown, N. P., Orengo, C. A. & Taylor, W. R. (1996). *A protein structure comparison methodology*. *Comput. Chem.* **20**, 359–380.

Chothia, C. (1993). *One thousand families for the molecular biologist*. *Nature (London)*, **357**, 543–544.

Crippen, G. (1978). *The tree structural organization of proteins*. *J. Mol. Biol.* **126**, 315–332.

Flores, T. P., Orengo, C. A. & Thornton, J. M. (1993). *Conformational characteristics in structurally similar protein pairs*. *Protein Sci.* **7**, 31–37.

Gibrat, J. F., Madej, T., Spouge, J. L. & Bryant, S. H. (1997). *The VAST protein structure comparison method*. *Biophys. J.* **72**, MP298.

Go, M. (1981). *Correlation of DNA exonic regions with protein structural units in hemoglobin*. *Nature (London)*, **291**, 90–92.

Hogue, C. W., Ohkawa, H. & Bryant, S. H. (1996). *A dynamic look at structures: WWW-Entrez and the molecular modelling database*. *Trends Biochem. Sci.* **21**, 226–229.

Holm, L. & Sander, C. (1993). *Protein structure comparison by alignment of distance matrices*. *J. Mol. Biol.* **233**, 123–138.

Holm, L. & Sander, C. (1994*a*). *Searching protein structure databases has come of age*. *Proteins*, **19**, 165–173.

Holm, L. & Sander, C. (1994*b*). *Parser for protein folding units*. *Proteins*, **19**, 256–268.

Holm, L. & Sander, C. (1995). *Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme*. *EMBO J.* **14**, 1287–1293.

Holm, L. & Sander, C. (1996). *Mapping the protein universe*. *Science*, **273**, 595–602.

Holm, L. & Sander, C. (1997). *Enzyme HIT*. *Trends Biochem. Sci.* **22**, 116–117.

Holm, L. & Sander, C. (1998). *Dictionary of recurrent domains in protein structures*. *Proteins*, **33**, 88–96.

Holm, L. & Sander, C. (1999). *Protein folds and families: sequence and structure alignments*. *Nucleic Acids Res.* **27**, 244–247.

Islam, S. A., Luo, J. & Sternberg, M. J. (1995). *Identification and analysis of domains in proteins*. *Protein Eng.* **8**, 513–525.

Jones, S., Stewart, M., Michie, A. D., Swindells, M. B., Orengo, C. A. & Thornton, J. M. (1998). *Domain assignment for protein structures using a consensus approach: characterisation and analysis*. *Protein Sci.* **7**, 233–242.

Kikuchi, T., Nemethy, G. & Scheraga, H. A. (1988). *Prediction of the location of structural domains in globular proteins*. *J. Protein Chem.* **88**, 427–471.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures*. *J. Appl. Cryst.* **24**, 946–950.

Lesk, A. M. & Rose, G. D. (1981). *Folding units in globular proteins*. *Proc. Natl Acad. Sci. USA*, **78**, 4304–4308.

Levitt, M. & Chothia, C. (1976). *Structural patterns in globular proteins. Nature (London)*, **261**, 552–558.

Li, M., Dyda, F., Benhar, I., Pastan, I. & Davies, D. R. (1996). *Crystal structure of the catalytic domain of Pseudomonas exotoxin A complexed with a nicotinamide adenine dinucleotide analog: implications for the activation process and for ADP ribosylation*. *Proc. Natl Acad. Sci. USA*, **93**, 6902–6906.

Lionetti, C., Guanziroli, M. G., Frigerio, F., Ascenzi, P. & Bolognesi, M. (1991). *X-ray crystal structure of the ferric sperm whale myoglobin: imidazole complex at 2.0 Å resolution*. *J. Mol. Biol.* **217**, 409–412.

Mizuguchi, K., Deane, C. A., Blundell, T. L. & Overerington, J. P. (1998). *HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci.* **7**, 2469–2471.

Moult, J. & Unger, R. (1991). *An analysis of protein folding pathways*. *Biochemistry*, **30**, 3816–3824.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *SCOP: a structural classification of the protein database for the investigation of sequences and structures. J. Mol. Biol.* **247**, 536–540.

Nemethy, G. & Scheraga, H. A. (1979). *A possible folding pathway of bovine pancreatic Rnase*. *Proc. Natl Acad. Sci. USA*, **76**, 6050–6054.

Orengo, C. A., Jones, D. T., Taylor, W. & Thornton, J. M. (1994). *Protein superfamilies and domain superfolds. Nature (London)*, **372**, 631–634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *CATH – a hierarchic classification of protein domain structures. Structure*, **5**, 1093–1108.

Orengo, C. A., Pearl, F. M. G., Bray, J. E., Todd, A. E., Martin, A. C., LoConte, L. & Thornton, J. M. (1999). *The CATH database provides insights into protein structure/function relationships. Nucleic Acids Res.* **27**, 275–279.

Phillips, D. E. (1970). *British biochemistry, past and present*, p. 11. London Biochemistry Society Symposium. Academic Press.

Rashin, A. A. (1976). *Location of domains in globular proteins*. *Nature (London)*, **291**, 85–87.

Richardson, J. S. (1981). *The anatomy and taxonomy of protein structure*. *Adv. Protein Chem.* **34**, 167–339.

Rose, G. D. (1979). *Hierarchic organization of domains in globular proteins*. *J. Mol. Biol.* **134**, 447–470.

Rossmann, M. G. & Argos, P. (1975). *A comparison of the heme binding pocket in globins and cytochrome b5*. *J. Biol. Chem.* **250**, 7525–7532.

Rossmann, M. & Liljas, A. (1974). *Recognition of structural domains in globular proteins*. *J. Mol. Biol.* **85**, 177–181.

Russell, R. B. & Barton, G. J. (1993). *Multiple protein sequence alignment from tertiary structure comparisons. Assignments of global and residue level confidences. Proteins*, **14**, 309–323.

Sali, A. & Blundell, T. B. (1990). *The definition of general topological equivalences in proteins: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming*. *J. Mol. Biol.* **212**, 403–428.

Sander, C. (1981). *Physical criteria for folding units of globular proteins*. In *Structural aspects of recognition and assembly in biological macromolecules*, Vol. I. *Proteins and protein complexes, fibrous proteins*, edited by M. Balaban, pp. 183–195. Jerusalem: Alpha Press.

Sander, C. & Schneider, R. (1991). *Database of homology-derived protein structures and structural meaning of sequence alignments. Proteins*, **9**, 56–68.

Schulz, G. E. & Schirmer, H. (1979). *Principles of protein structure*, ch. 5. New York: Springer Verlag.

Siddiqui, A. S. & Barton, G. J. (1995). *Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions*. *Protein Sci.* **4**, 872–884.

Sowdhamini, R., Rufino, S. D. & Blundell, T. L. (1996). *A database of globular protein structural domains: clustering of representa-*

23.1 (*cont.*)

tive family members into similar folds. *Structure Fold. Des.* **1**, 209–220.
Swindells, M. B. (1995). *A procedure for detecting structural domains in proteins*. *Protein Sci.* **4**, 103–112.
Taylor, W. R. & Orengo, C. A. (1989). *Protein structure alignment. J. Mol. Biol.* **208**, 1–22.
Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y. & Steitz, T. A. (1996). *Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose*. *EMBO J.* **15**, 5739–5751.
Vriend, G. & Sander, C. (1991). *Detection of common three-dimensional substructures in proteins. Proteins*, **11**, 552–558.
Wernisch, L., Hunting, M. & Wodak, J. (1999). *Identification of structural domains in proteins by a graph heuristic*. *Proteins*, **35**, 338–352.
Wetlaufer, D. B. (1973). *Nucleation, rapid folding, and globular intrachain regions in proteins*. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.
Wodak, J. & Janin, J. (1981). *Location of structural domains in proteins*. *Biochemistry*, **20**, 6544–6552.
Zehfus, M. H. (1994). *Binary discontinuous compact protein domains*. *Protein Eng.* **7**, 335–340.
Zehfus, M. H. (1997). *Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains*. *Protein Sci.* **6**, 1210–1219.
Zehfus, M. H. & Rose, G. D. (1986). *Compact units in proteins*. *Biochemistry*, **25**, 5759–5765.

23.2

Åqvist, J., Luecke, H., Quiocho, F. A. & Warshel, A. (1991). *Dipoles localized at helix termini of proteins stabilize charges*. *Proc. Natl Acad. Sci. USA*, **88**, 2026–2030.
Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *On the conformation of the hen egg-white lysozyme molecule*. *Proc. R. Soc. London Ser. B Biol. Sci.* **167**, 365–377.
Bochkarev, A., Pfuetzner, R. A., Edwards, A. M. & Frappier, L. (1997). *Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA*. *Nature (London)*, **385**, 176–181.
Burd, C. G. & Dreyfuss, G. (1994). *Conserved structures and diversity of functions of RNA-binding proteins*. *Science*, **265**, 615–621.
Cheng, X. (1995). *DNA modification by methyltransferases*. *Curr. Opin. Struct. Biol.* **5**, 4–10.
Cleland, W. W. & Kreevoy, M. M. (1994). *Low-barrier hydrogen bonds and enzymic catalysis*. *Science*, **264**, 1887–1890.
Cotton, F. A., Hazen, E. E. Jr & Legg, M. J. (1979). *Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme–thymidine 3′,5′-bisphosphate–calcium ion complex at 1.5-Å resolution*. *Proc. Natl Acad. Sci. USA*, **76**, 2551–2555.
Cusack, S., Yaremchuk, A. & Tukalo, M. (1996*a*). *The crystal structure of the ternary complex of T. thermophilus seryl-tRNA synthetase with tRNA(Ser) and a seryl-adenylate analogue reveals a conformational switch in the active site*. *EMBO J.* **15**, 2834–2842.
Cusack, S., Yaremchuk, A. & Tukalo, M. (1996*b*). *The crystal structures of T. thermophilus lysyl-tRNA synthetase complexed with E. coli tRNA(Lys) and a T. thermophilus tRNA(Lys) transcript: anticodon recognition and conformational changes upon binding of a lysyl-adenylate analogue*. *EMBO J.* **15**, 6321–6334.
Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A. L., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). *Science*, **280**, 68–77.
Freemont, P. S., Friedman, J. M., Beese, L. S., Sanderson, M. R. & Steitz, T. A. (1988). *Cocrystal structure of an editing complex of Klenow fragment with DNA*. *Proc. Natl Acad. Sci. USA*, **85**, 8924–8928.
Gerlt, J. A. & Gassman, P. G. (1993). *Understanding the rates of certain enzyme-catalyzed reactions: proton abstraction from carbon acids, acyl-transfer reaction, and displacement reactions of phosphodiesters*. *Biochemistry*, **32**, 1943–1952.
Glusker, J. P. (1991). *Structural aspects of metal liganding to functional groups in proteins*. *Adv. Protein Chem.* **42**, 1–76.
Goldgur, Y., Mosyak, L., Reshetnikova, L., Ankilova, V., Lavrik, O., Khodyreva, S. & Safro, M. (1997). *The crystal structure of phenylalanyl-tRNA synthetase from Thermus Thermophilus complexed with cognate tRNAPhe*. *Structure*, **5**, 59–68.
Harrington, R. E. & Winicov, I. (1994). *New concepts in protein–DNA recognition: sequence-directed DNA bending and flexibility. Prog. Nucleic Acid Res. Mol. Biol.* **47**, 195–270.
He, J. J. & Quiocho, F. A. (1993). *Dominant role of local dipoles in stabilizing uncompensated charges on a sulphate sequestered in a periplasmic active transport. Protein Sci.* **2**, 1643–1647.
Hibbert, F. & Emsley, J. (1990). *Hydrogen bonding and chemical reactivity. Adv. Phys. Org. Chem.* **226**, 255–379.
Hodel, A. E., Gershon, P. D. & Quiocho, F. A. (1998). *Structural basis for sequence non-specific recognition of 5′-capped mRNA by a cap-modifying enzyme*. *Mol. Cell*, **1**, 443–447.
Hodel, A. E., Gershon, P. D., Shi, X., Wang, S. M. & Quiocho, F. A. (1997). *Specific protein recognition of an mRNA cap through its alkylated base*. (Letter.) *Nature Struct. Biol.* **4**, 350–354.
Jacobson, B. L. & Quiocho, F. A. (1988). *Sulphate-binding protein dislikes protonated oxyacids: a molecular explanation*. *J. Mol. Biol.* **204**, 783–787.
Joachimiak, A., Schevitz, R. W., Kelley, R. L., Yanofsky, C. & Sigler, P. B. (1983). *Functional inferences from crystals of Escherichia coli trp repressor*. *J. Biol. Chem.* **258**, 12641–12643.
Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. (1990). *Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: a framework for understanding homeodomain–DNA interactions*. *Cell*, **63**, 579–590.
Labahn, J., Scharer, O. D., Long, A., Ezaz-Nikpay, K., Verdine, G. L. & Ellenberger, T. E. (1996). *Structural basis for the excision repair of alkylation-damaged DNA*. *Cell*, **86**, 321–329.
Ledvina, P. S., Tsai, A.-H., Wang, Z., Koehl, E. & Quiocho, F. A. (1998). *Dominant role of local dipolar interactions in phosphate binding to a receptor cleft with an electronegative charge surface potential: equilibrium, kinetic and crystallographic studies. Protein Sci.* **7**, 2550–2559.
Ledvina, P. S., Yao, N., Choudhary, A. & Quiocho, F. A. (1996). *Negative electrostatic surface potential of protein sites specific for anionic ligands*. *Proc. Natl Acad. Sci. USA*, **93**, 6786–6791.
Lindahl, T. (1982). *DNA repair enzymes*. *Annu. Rev. Biochem.* **51**, 61–87.
Luecke, H. & Quiocho, F. A. (1990). *High specificity of a phosphate transport protein determined by hydrogen bonds*. *Nature (London)*, **347**, 402–406.
Marcotrigiano, J., Gingras, A. C., Sonenberg, N. & Burley, S. K. (1997). *X-ray studies of the messenger RNA 5′ cap-binding protein (eIF4E) bound to 7-methyl-GDP*. *Nucleic Acids Symp. Ser.* pp. 8–11.
Meador, W. E., George, S. E., Means, A. R. & Quiocho, F. A. (1995). *X-ray analysis reveals conformational adaptation of the linker in functional calmodulin mutants*. (Letter.) *Nature Struct. Biol.* **2**, 943–945.
Medveczky, N. & Rosenberg, H. (1971). *Phosphate transport in Escherichia coli*. *Biochim. Biophys. Acta*, **241**, 494–506.
Nagai, K. (1996). *RNA–protein complexes*. *Curr. Opin. Struct. Biol.* **6**, 53–61.
Nagai, K., Oubridge, C., Ito, N., Jessen, T. H., Avis, J. & Evans, P. (1995). *Crystal structure of the U1A spliceosomal protein complexed with its cognate RNA hairpin*. *Nucleic Acids Symp. Ser.* 1–2.
Nicholls, A., Sharp, K. & Honig, B. (1991). *Protein folding and association; insights from the interfacial and thermodynamic properties of hydrocarbons*. *Proteins*, **11**, 281–296.
Orgel, L. E. (1966). *An introduction to transition-metal chemistry. Ligand-field theory*, 2nd ed. London: Methuen and New York: Wiley.