Protein folds and functions

Andrew CR Martin[1], Christine A Orengo[1], E Gail Hutchinson[1], Susan Jones[1], Maria Karmirantzou[1], Roman A Laskowski[2], John BO Mitchell[1], Chiara Taroni[1] and Janet M Thornton[1,2]*

Background: The recent rapid increase in the number of available three-dimensional protein structures has further highlighted the necessity to understand the relationship between biological function and structure. Using structural classification schemes such as SCOP, CATH and DALI, it is now possible to explore global relationships between protein fold and function, something which was previously impractical.

Results: Using a relational database of CATH data we have generated fold distributions for arbitrary selections of proteins automatically. These distributions have been examined in the light of protein function and bound ligand. Different enzyme classes are not clearly reflected in distributions of protein class and architecture, whereas the type of bound ligand has a much more dramatic effect.

Conclusions: The availability of structural classification data has enabled this novel overview analysis. We conclude that function at the top level of the EC number enzyme classification is not related to fold, as only a very few specific residues are actually responsible for enzyme activity. Conversely, the fold is much more closely related to ligand type.

Addresses: [1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK and [2]Crystallography Department, Birkbeck College, Malet Street, London WC1E 7HX, UK.

*Corresponding author.
E-mail: thornton@biochem.ucl.ac.uk

Key words: CATH, enzyme class, ligands, structural classification, topology

Received: **2 April 1998**
Revisions requested: **8 May 1998**
Revisions received: **27 May 1998**
Accepted: **28 May 1998**

Structure 15 July 1998, **6**:875–884
http://biomednet.com/elecref/0969212600600875

Introduction

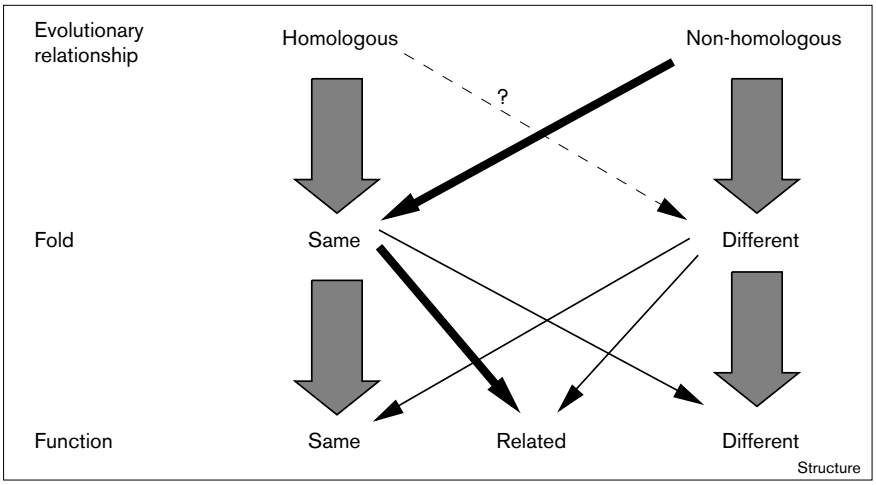
As the output from the genome projects gains pace, we are faced with a plethora of sequence data from which we wish to derive and understand biological function, both *in vitro* and *in vivo* [1]. It is timely, therefore, to consider the relationship between the three-dimensional (3D) structure of a protein and its biological function, using the relatively new structural classification schemes [2–4]. Herein we present one approach to considering the global relationships between protein fold, or topology, and function. There are several questions we would like to answer. Why does one particular protein perform a given function? Is there any significant relationship between the fold of a protein and its biological function? Can we discover any rules or guidelines which may suggest function from structure?

Whilst such questions are of major intellectual and evolutionary interest, a better understanding in this field could help practically to improve genome analysis, the search for a function for unknown open reading frames and the design of proteins with novel or modified functions.

The great majority of proteins which exhibit significant structural similarity are homologues and perform identical or similar functions. Beyond these inherited similarities, however, the different enzyme functions (as defined by their EC numbers) are performed by proteins with a wide variety of different architectures and topologies. Given this observation, it is striking that the structures of the 11 enzymes of the glycolytic pathway all belong to the $\alpha\beta$ class of structures (and use only three architectures). Functional classification for other proteins is more difficult, but we do find distinct structural class preferences for those proteins that bind some of the most common biomolecules — haems, sugars, nucleic acids and nucleotides. Nevertheless, within such a group, the individual proteins adopt a wide variety of different topologies to bind their similar ligands, which are used for different functions.

There are now more than 7000 entries in the Brookhaven Protein Databank (PDB; [5]) and these have revealed some amazing examples of fold–function relationships and evolution. Figure 1 presents a scheme to describe the possible relationships between proteins, their folds and functions. Proteins may be homologous (i.e. possess a common ancestor) or non-homologous, whilst their folds and functions may be identical, related or totally different. To date, all protein pairs with sequences which indicate a definite evolutionary relationship are observed to adopt the 'same' fold, with only minor variation (e.g. changes in domain orientations, lengths of loops or additional secondary structures). For example, globins from a wide variety of species, with widely diverged sequences, all adopt the same fold and perform an oxygen carrier/storage function.

Figure 1



Schematic representation of the relationships between fold and function. The thickness of the arrows gives an approximate indication of the relative frequency of the relationships. The dashed line with a query indicates a tentative relationship. Thus proteins with an evolutionary relationship generally have the same fold and function while non-homologous proteins have different folds and functions. Cross-overs in this scheme do occur, but are less frequent.

However, there are also a few homologous proteins which clearly have different functions, despite adopting the same structure. The classic example is that of lysozyme and α-lactalbumin. Although these enzymes possess ~35% sequence identity, the latter has lost the catalytic carboxylates from glutamate and aspartate residues necessary for sugar cleavage [6]. It has also been observed that several enzymes are expressed in the eye lens (e.g. lactate dehydrogenase) where they are thought to provide 'protein mass' rather than catalytic activity [7].

In contrast, there are several examples of proteins which perform the same function, but are clearly not evolutionarily related. Here the classic examples are the trypsin and subtilisin proteinases, which not only perform the same function despite having totally different structures, but have evolved the same Asp–His–Ser catalytic-triad mechanism. This is a genuine example of functional convergence [8–10]. A slightly different example is provided by the serine proteinase inhibitors. Although these proteins adopt a wide variety of folds, they all possess a canonical 'loop' structure, which mimics the substrate and binds to the active site of the proteinase [11]. There are also proteins with dual function, perhaps the most extreme example being the trypsin proteinase in Sindbis virus, which, having performed its catalytic function, becomes the coat protein of the virus [12]. These examples serve to emphasise that the relationship between structure and function is not straightforward, and a global analysis will only reveal underlying trends to which there will always be exceptions. With this caveat in mind, we present one approach to a global analysis of the relationship between fold and function.

Our analysis represents a novel approach to studying structure/function relationships at a global level. We have demonstrated that there is little correlation between the protein class or architecture and the enzyme function, presumably because enzymic activity is defined by only a few amino acids. In contrast, there seems to be a much better correlation between class and architecture and the ligand type. Given the low degree of correlation within the enzyme groups, it is interesting to note that within the glycolytic pathway, the distribution is much more restricted and this may have some evolutionary significance.

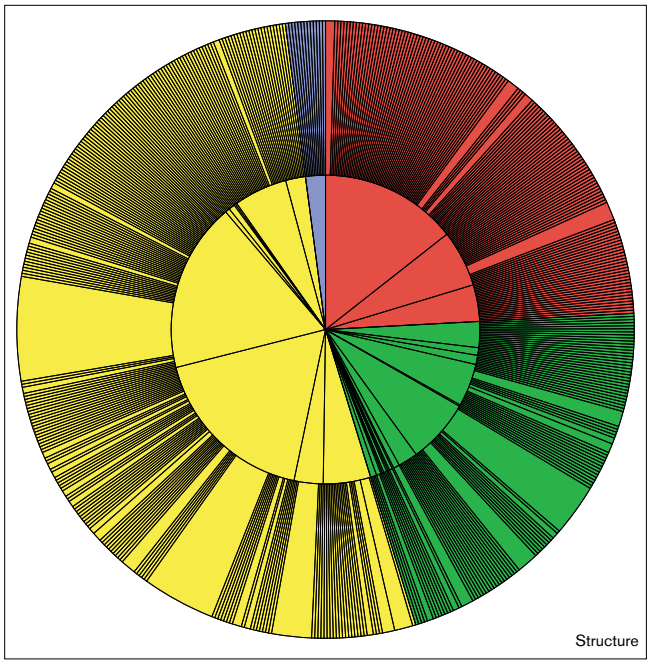
Results

The results of the semi-automated CATH domain classification can be summarised by the 'CATH wheel', shown in Figure 2. The figure shows the classification for all proteins in the July 1997 release of the PDB, which comprises more than 10,000 domains. From the plot, it is immediately apparent that approximately 25% of proteins in the PDB are mainly α, 25% are mainly β and almost half are αβ proteins. There are very few domains with low secondary structure content. The relative size of each sector in a circle reflects the number of homologous families in the PDB in that class, architecture or topology.

Extracellular location

The extracellular environment is known to be hostile to proteins, and disulphide bridges help to ensure stability; within the cell, proteins are exposed to reducing conditions. Whereas the presence of a disulphide bridge (which can be derived directly from the coordinates) is a reliable indicator of extracellular location, the absence of a disulphide bridge does not necessarily indicate intracellular location. Nevertheless, we first investigated the distribution of folds for proteins with disulphide bridges, using this criterion as an indicator of extracellular location (see Figure 3). Statistics for these examples, and for those examined later, are presented in Table 1.

Figure 2



Structural classification for all domains in the PDB. This representation (the CATH wheel) comprises a set of concentric pie charts. The colours define the class [C]: red, mainly α; green, mainly β; yellow, mixed αβ; and blue, low secondary structure. The inner circle represents the architecture [C.A] and the outer circle represents the topology [C.A.T]. The angle defined by any segment is proportional to the number of homologous families [C.A.T.H] it contains.

Compared with the overall distribution in Figure 2, there is a distinct shift towards the mainly β proteins outside the cell — dominated by the β-sandwich structures (CATH number [2.60]) found in antibodies and many extracellular receptors. This shift is at the expense of the αβ class of structures. It is possible that this fold distribution reflects an easier formation of disulphide bridges in β structures. The prevalence of disulphide bridges in β structures has been observed previously and it is known that there is a disproportionately low number of cases of disulphide bridges linking α helices in the PDB [13]. There are, however, no steric reasons why disulphide bridges should not form between α helices — indeed, phospholipase contains four helices and seven disulphide bonds [14]. It is therefore equally likely that the preference may reflect distant evolutionary events or intrinsic stability factors, suggesting that the mainly β proteins are on average more stable than other fold classes in the extracellular environment. If this is the case, the prevalence of disulphide bonds within β structures reflects their extracellular location rather than any intrinsic preference to form disulphide bonds between β strands. These distributions help to explain why it is possible to make a reasonable prediction of cellular location from amino-acid composition [15] because, on average, residues with high β propensity will be more common in extracellular proteins.

Enzyme structure and function

The enzymes are the easiest protein functional class to analyse in the PDB because they are numerous (5819 chains in the July 1997 PDB also assigned in CATH) and are logically classified in functional terms by their EC numbers [16]. In addition, the enzyme database is available electronically [17].

The primary EC number defines the class of the enzyme: 1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases; and 6, synthetases. The meaning of subsequent numbers depends on the primary class and provides information on the substrate acceptor and cofactors. Here we consider only the primary classes for single-domain

Figure 3

Structural classification for **(a)** intracellular and **(b)** extracellular protein domains as indicated by the absence or presence of disulphide bonds. For an explanation of these plots see Figure 2.

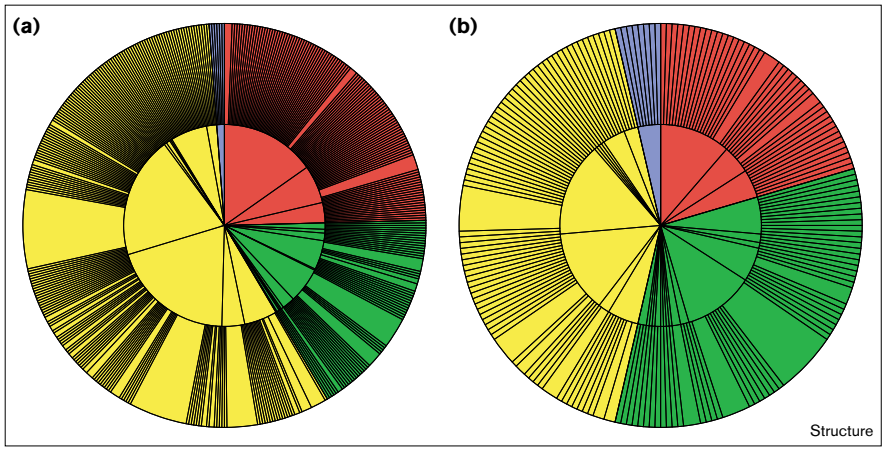


Table 1

Statistical significance of class distribution for non-homologous families.

Category	Observations				χ^2	P
	Mainly α	Mainly β	αβ and low	Total		
All proteins	151	132	342	625	–	–
Intracellular	121	82	288	491	5.98	< 0.1
Extracellular	45	73	102	220	19.21	< 0.001
EC1	6	7	10	23	1.51	< 0.5
EC2	5	7	22	34	1.88	< 0.5
EC3	11	21	46	78	4.77	< 0.1
EC4*	1	1	9	11	–	–
EC5*	2	1	9	12	–	–
EC6*	0	2	2	4	–	–
Haem-binding domains	14	2	7	23	16.99	< 0.001
Carbohydrate-binding domains	5	11	11	27	6.24	< 0.05
DNA-binding domains	9	1	7	17	8.27	< 0.02
Nucleotide-binding domains	5	2	31	38	11.44	< 0.005

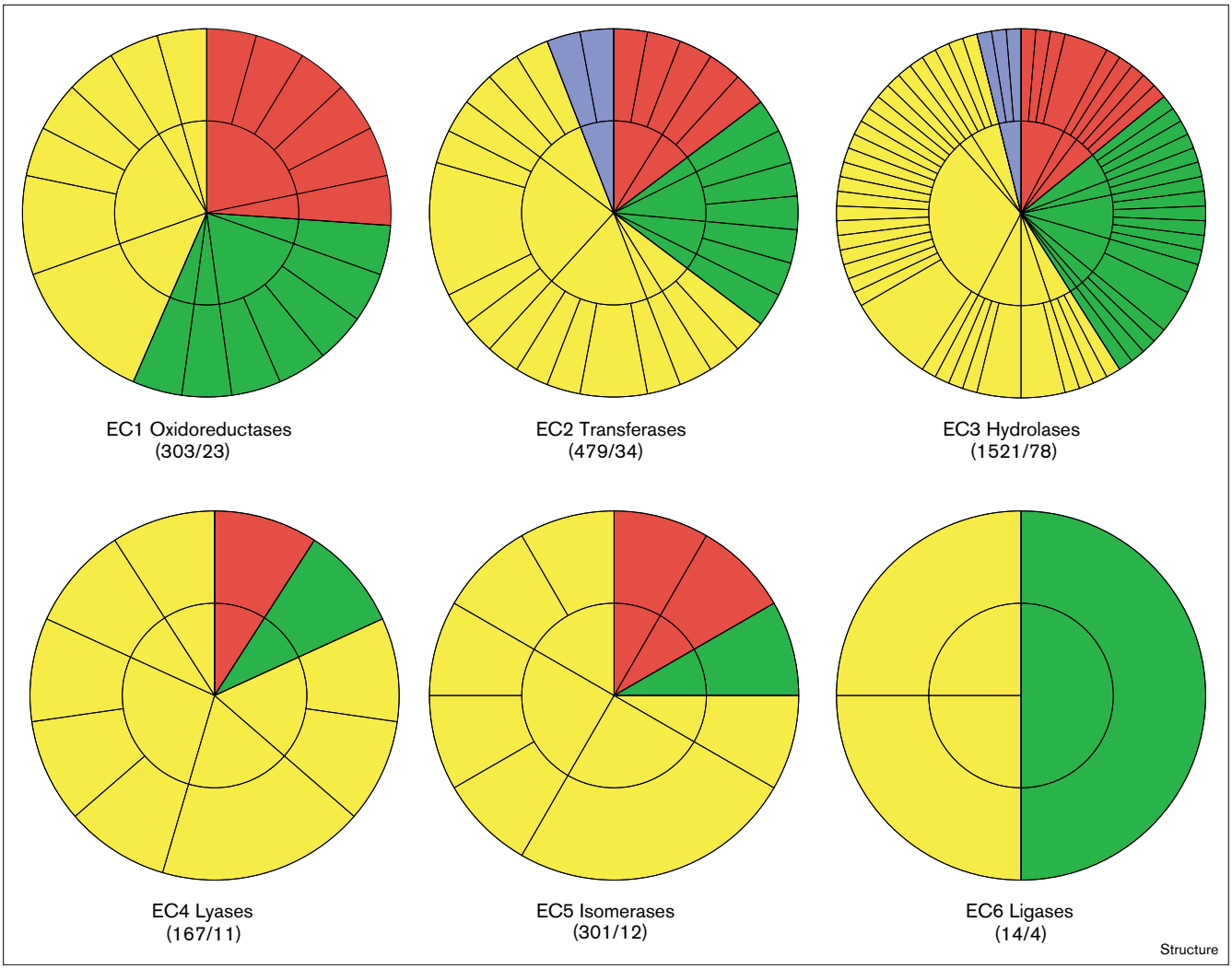
Observations of protein structural classes for non-homologous families in different categories of proteins. The χ^2 – calculated as $\Sigma((O-E)^2/E)$ and probability (P) values are calculated for each category, using expected values derived from the 'all proteins' values (i.e. $E_{cx} = O_{ax} \times (T_c/T_a)$, where x is a protein class (mainly α, mainly β, αβ plus low secondary structure), c is a category of proteins being observed and a represents all proteins. Thus, E_{cx} is the expected value for protein class x in category c, O_{ax} is the observed number of occurrences of class x in all proteins, T_c is the total number of observations in category c and T_a is the total number of observations in all proteins. As the number of observations of low secondary structure proteins is small, these have been grouped with the αβ class for the purposes of this analysis (ignoring the low secondary structure group altogether has only a small effect on the χ^2 values). The probability is a measure of the random chance of obtaining this distribution rather than the expected distribution (i.e. that observed by looking at all proteins). *For EC4–6, there are too few non-homologous family examples to obtain meaningful statistics.

enzyme chains (this removes the problem of assigning the enzyme activity to a specific domain). Figure 4 illustrates, for all examples in the PDB, the structure distributions for each enzyme class. The CATH wheels suggest that the distributions for EC1–3 are only marginally different from the expected distribution (i.e. that seen for all proteins). The statistics in Table 1 support this assertion, with a relatively high probability of obtaining these distributions by chance (significant only at the 10% level or worse). Although the CATH wheels suggest that EC4–6 are significantly different, the numbers of non-homologous examples are too small for the statistics to be meaningful (χ^2 values have not been calculated).

These data show that all classes of domains form enzymes, although the mainly α class is under represented and the αβ class is over represented, compared with the distribution for all proteins currently known. There has been some discussion of the possible relationship of class and enzyme activity and it has been suggested that helices may be required for mechanical actions in enzymes [18]. If this is the case, it appears to be the exception rather than the rule because helices, being less flexible than strands, may not be able to make adequate, subtle movements during catalysis. Another factor contributing to the under representation of the mainly α class in enzymes may be that, in helices, the mainchain polar groups are all satisfied and not available for interactions, whereas at the edge of a β sheet these groups are accessible, yet held rigidly in well-defined conformations, and may be used to bind a polar substrate or be part of the catalytic process (e.g. in the serine proteinases). The dominance of the αβ folds largely reflects the presence of the nucleotide-binding domains found in many enzymes (see below). Given these distributions, the similarity of the 11 enzyme structures of the glycolytic pathway is even more striking ([19]; Figure 5) and may reflect an evolutionary process. The 12 active-site domains from the 11 proteins make use of only three architectures and nine topologies.

In enzymes, the catalytic mechanism and function depend on the precise location and orientation of a very few amino acids. Therefore, of all proteins, enzymes are the least likely to exhibit a fundamental relationship between their gross structure, as encapsulated by the higher levels of the CATH numbers, and their specific function. Indeed, amongst the single-domain enzymes in the current database we found 37 examples of one EC number corresponding to more than one topology (including five examples of the same EC number being assigned to four different folds) and 36 examples of members of one homologous family having different EC numbers. In the most extreme case, the [3.20.20.70] family within the TIM (triose phosphate isomerase) barrel fold has 13 different EC numbers associated with it. These structures include primary EC numbers 1–5, highlighting the lack of correlation between

Figure 4



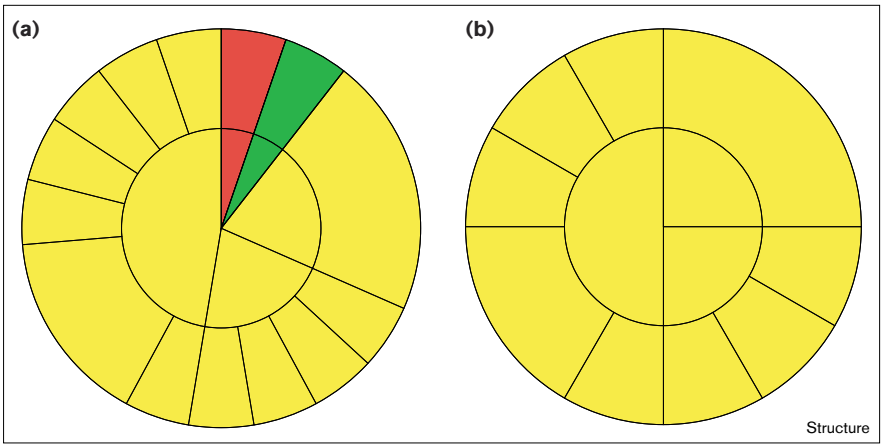
Structural classification for all single-domain enzymes in the PDB, grouped according to their primary EC number. The numbers of examples followed by the number of non-homologous families (as defined by CATH) are shown in parentheses under each CATH wheel.

EC number and topology. Therefore, in our view, the differences between the distributions seen for the different enzyme classes are unlikely to represent a fundamental correlation between fold and function. A more likely source of correlation is to be found by considering not function per se, but the type of molecule the protein binds in performing its biological activity. We have therefore grouped and analysed proteins which bind similar molecules to see if any similarity of fold is observed. The rationale is that molecules of a certain shape or polarity may only be recognised by certain folds or structural classes of protein. Below we present the results for four major types of 'ligands' which are particularly common biomolecules. In practice, we can automatically generate the data for any ligand.

Haem-binding domains

Figure 6a shows the CATH distribution for the 13 non-homologous protein families, comprising 23 non-homologous domains, which are known to bind haem. (In total there are 523 such domains, but only one representative is automatically chosen from each family.) Diverse examples are shown in Figure 7a. The class distribution is seen to be radically different (Table 1) from that shown in Figure 2. The dominance of the mainly α proteins is striking. Inspection of the individual binding sites shows that although they are very different in topology and detail, the preferred binding mode is for the haem to slot between two or more helices, with its hydrophobic faces shielded from the solvent and interacting with hydrophobic sidechains in the helices (Figure 8a). Indeed, in

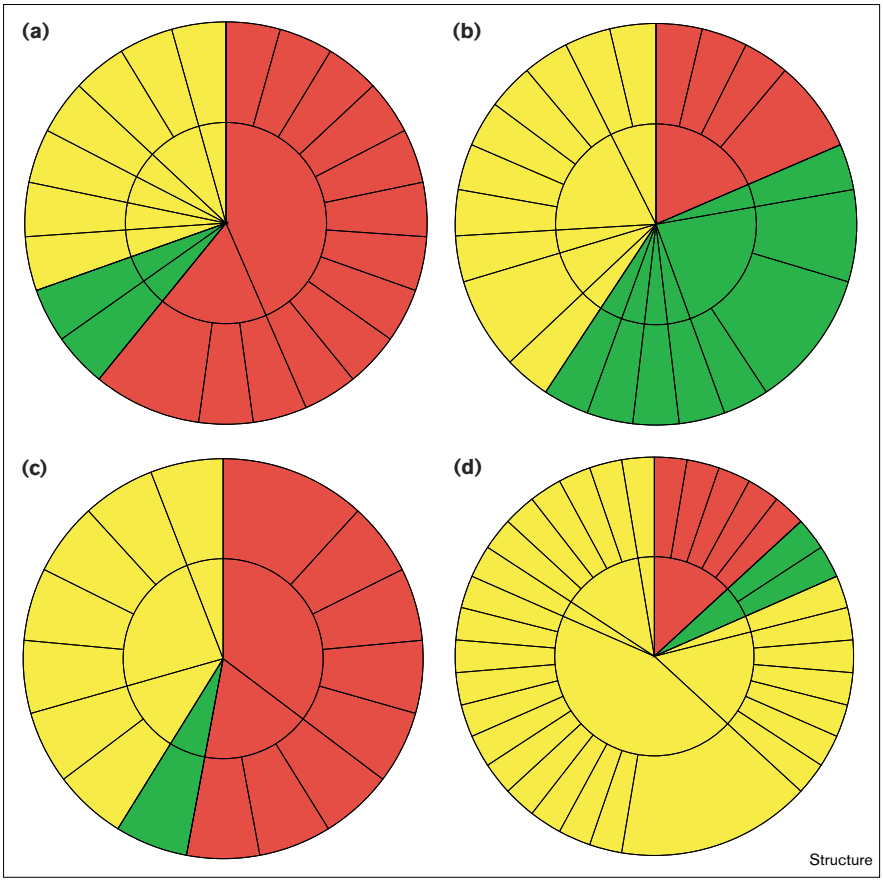
Figure 5



Structural classification for the structures of the 11 enzymes in the glycolytic pathway. The CATH wheel representations for **(a)** whole proteins and **(b)** active-site domains are shown.

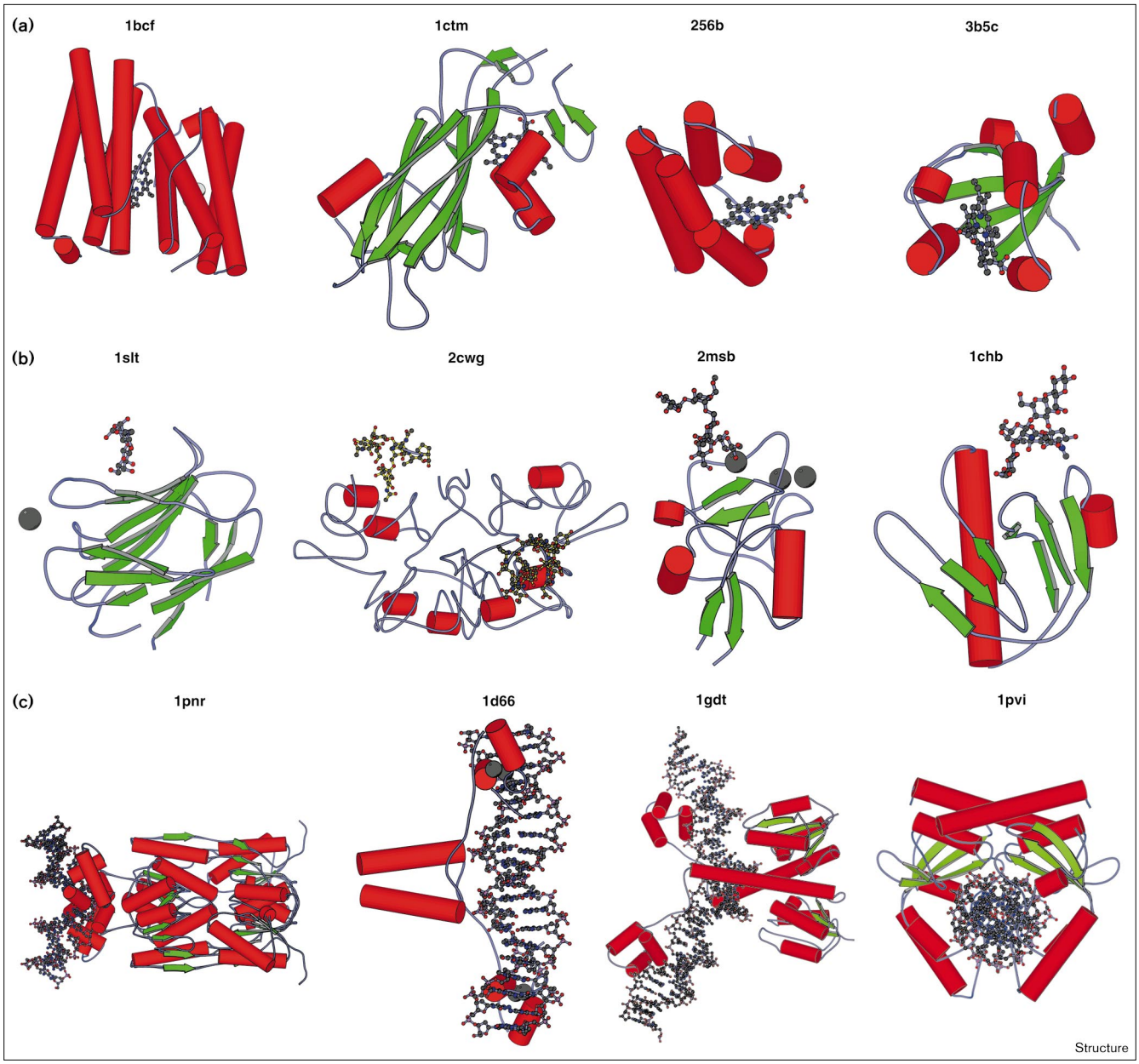
the 14 mainly α homologous families of domains in this group, the binding sites are constituted in the same way, whereas in the two mainly β domain families, the loops act in a similar role to the helices. It is apparently difficult to bury the large planar hydrophobic haem group using β sheets alone. Nevertheless, we realise that it is possible for differing motifs to be used to bind identical ligands and one recent example, cytochrome cd_1, not yet classified in CATH, shows a haem group sandwiched within an eight-bladed β-propeller structure [20].

Figure 6



Structural classification (CATH wheel) for **(a)** haem-binding domains, **(b)** carbohydrate-binding domains, **(c)** DNA-binding domains, and **(d)** nucleotide-binding domains.

Figure 7

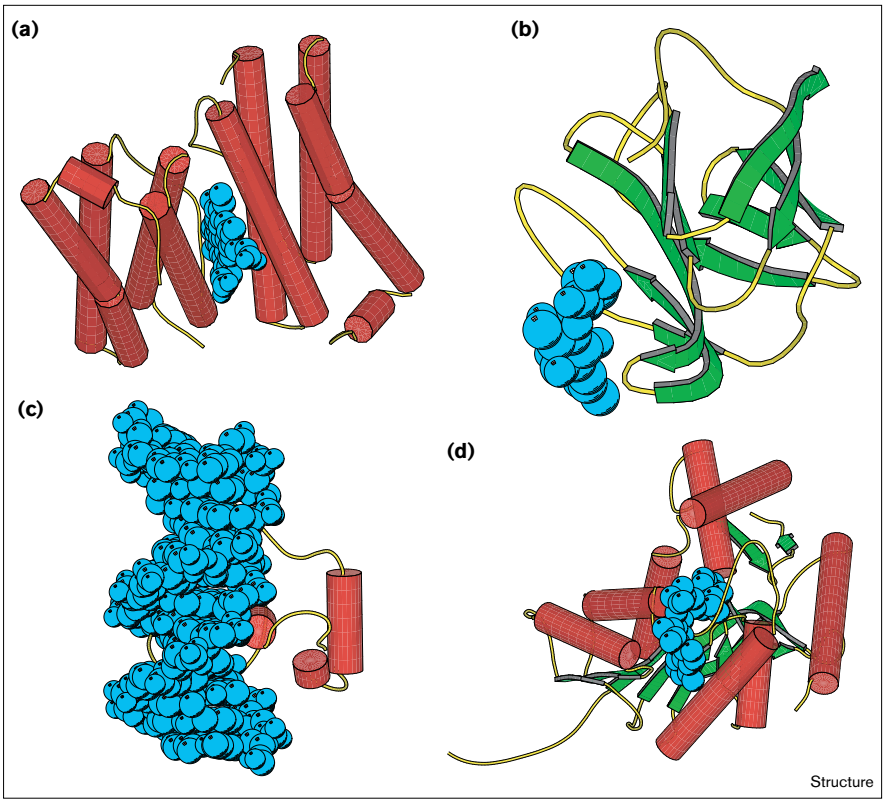


Collages of diverse examples of ligand-binding proteins including **(a)** haem-binding proteins: bacterioferritin (cytochrome *b1*, 1bcf), cytochrome *f* (1ctm), cytochrome *b562* (256b), cytochrome *b5* (3b5c); **(b)** carbohydrate-binding proteins: galectin *1* (1slt, chain A), wheat germ agglutinin (2cwg), mannose-binding protein (2msb, chain B), cholera toxin (1chb, chain H) and **(c)** DNA-binding proteins: purine repressor(1pnr), transcriptional activator GAL4 (1d66), gamma-delta resolvase (1gdt), PVUII endonuclease (1pvi). Pictures were generated using the program MOLSCRIPT V2 [30]. For references to the individual structures, see their entries in the PDB [5].

Carbohydrate-binding domains

The class distribution of carbohydrate-binding domains differs significantly from that expected from all proteins (Table 1). This group of 22 proteins (51 domains) covers principally the sugar-processing enzymes and the lectins. Figure 6b shows a large group of mainly β proteins, which dominate the lectin family illustrated in Figure 7b. As can be seen in the figure, the details of the binding sites are very different, but in lectins many sugars are bound in a shallow depression, cradled in a β sheet or loop structure (see Figure 8b). In the sugar-processing enzymes, the binding sites are very different from those of the lectins, as they are contained in rather deep clefts, which render the sugar inaccessible to solvent and primed for catalytic

Figure 8



Examples of representative structures which bind their ligand in the mode most often observed in other proteins which bind the same ligand. **(a)** Haem-binding protein, 1bcf – a classic haem-binding pocket, lined with α helices [31]; **(b)** carbohydrate-binding protein, 1slt (chain A) – a lectin binding a sugar in a shallow depression on the surface of the mainly β structure [32]; **(c)** DNA-binding protein, 1hcr – a classic helix interaction binding in the major groove of the DNA [33] and **(d)** nucleotide-binding protein, 9ldt (chain A) with NAD [34]. The figures were generated using the program MOLSCRIPT V2 [30]. In all cases, α helices are shown in red, β strands in green, and the rest of the protein in yellow; the ligand is shown in cyan.

attack. These enzymes have many different architectures and topologies [21], partly reflecting the many different sorts of carbohydrates found *in vivo*.

DNA-binding domains

The fold distribution for 18 DNA-binding proteins (33 domains; [22]) seen in Figure 6c and Table 1, shows many mainly α and αβ proteins, but few mainly β structures. Again there is a multitude of different structures and folds, examples of which are shown in Figure 7c, but DNA recognition is dominated by the helix motif binding in the major groove, such that the base sequence can be recognised (Figure 8c). This motif is found in the mainly α and αβ classes of proteins. The origin of this distribution must reflect the exquisite fit of a helix into the major groove. Some structures do exhibit β-hairpin binding in the major groove or involve complex loop structures, but the helix interaction is clearly the most common.

Nucleotide-binding domains

The nucleotide-binding domains are found in many different proteins with various functions (71 domains from 31 proteins). The fold distribution shown in Figure 6d and Table 1 is striking, being dominated by the αβ class of structures (81.6%). Since the first observation of a nucleotide-binding domain, the Rossmann fold [23] in lactate dehydrogenase, many different structures have been determined which bind one of a variety of nucleotides, although the Rossmann fold (CATH number [3.40.50]) remains dominant. Almost all of these alternative folds, however, are found to be αβ proteins adopting a three-layer αβα architecture with various topologies. The nucleotide is usually located in similar binding sites extended along the C-terminal end of the parallel β sheet (an example is shown in Figure 8d).

Discussion

Through the use of both CATH wheels and statistical tests, we have shown that fold is strongly correlated with the nature of the ligand (at least for four major biological ligands, which are all significant at the 5% level or better — in some cases far better than the 0.1% level), but not with enzyme function (significant at only the 10% level or worse). We have been very conservative in our analysis, including only one representative from each homologous family in the PDB. We found that the vast majority of proteins with the same topology are homologues and have similar functions. Beyond these inherited similarities, this work only approaches the correlation between structure and function from a limited perspective — that of using the enzyme classification or comparing proteins which bind similar molecules. For the enzymes, the EC classification

shows little correlation, at this gross level, with structure. Enzymes with the same EC number may exhibit different folds and vice versa. It may be that more significant relationships occur within pathways, where the substrate is successively transferred from enzyme to enzyme along the pathway, requiring similar binding sites at each stage. For several of the common biological ligands, we have shown that there is a distinct bias towards certain protein classes defined by the stereochemical requirements for binding the ligand. Beyond this, however, the exact geometry of the binding site can be constituted very differently, with different topologies, to provide the complementarity in shape, hydrogen bonding, and hydrophobic and electrostatic profiles between the protein and its ligand. With the advent of the structural classification schemes and the emerging functional classifications, it will be possible to extend this work to explore more easily the evolution of folds, functions and pathways. This will give us an insight into how these relatively simple molecules have evolved to cooperate and create the complex biochemical pathways and cascades that are essential for life.

All CATH wheels in this paper can be viewed at http://www.biochem.ucl.ac.uk/bsm/cathwheels/. These wheels are hyperlinked so that all information on the class, architecture, fold and homologous family is available, as well as details and references to the individual structures. In addition, there is a server to generate CATH wheels automatically, by using a list of PDB or domain codes, as well as the ability to cut and paste hit lists from our HETGROUP database [24] or the PDB ligand search tool, Relibase [25].

Biological implications

The recent rapid increase in the number of available protein sequences and three-dimensional structures has precipitated the need for reliable structural classification schemes. We report here one approach used to consider the global relationships between protein fold and function. The analysis presented here is an early application of the novel protein structure classification schemes to the understanding of structure/function relationships. The results separate the notion of protein function into enzyme activity on the one hand, and ligand type on the other; the structure/function relationships are very different in these two cases. The fold distribution for a selection of proteins was examined in terms of protein function and bound ligand. Although little correlation was observed between the protein class or architecture and enzyme function, a strong correlation was seen between class and architecture and ligand type. In contrast, the observation that structures of the enzymes of the glycolytic pathway are more closely related may have evolutionary implications and we intend to explore other pathways in the same way. These results have implications for the prediction of function from structure in performing genome analysis.

Materials and methods

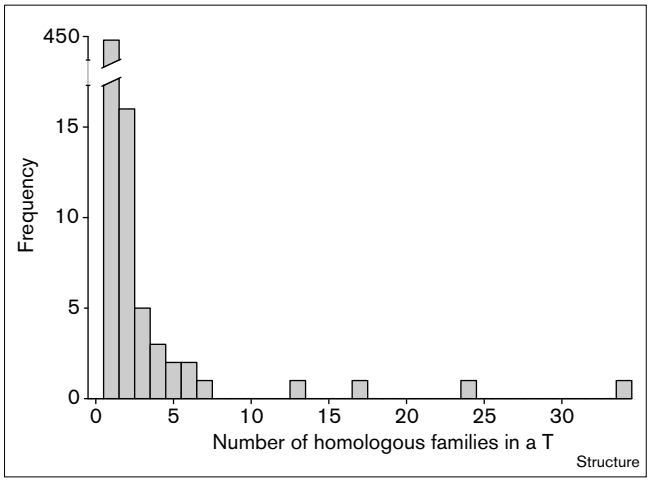
Brief overview of the CATH classification scheme

Recently several groups, including our own, have developed structural classification schemes [2–4]. In the CATH classification scheme [3], protein domains are grouped by four criteria: class (α, β, $\alpha\beta$, or low secondary structure); architecture (a level describing the gross arrangement of secondary structures in 3D space, but ignoring connectivity); topology (or fold), which groups together domains with the same topology as judged by the SSAP (Sequence and Structure Alignment Program – a method which uses double dynamic programming to align two protein domains) algorithm [26]; and homologous family, in which proteins with sequence, structure and/or functional evidence for a common ancestor are grouped. This classification can be browsed on the internet at http://www.biochem.ucl.ac.uk/bsm/cath/. Each protein is assigned a CATH number, which defines its class, architecture, topology and homologous family (e.g. triose phosphate isomerase is classified as [3.20.20.80], denoting that it is in the $\alpha\beta$ class [3] with a barrel architecture [3.20], a TIM fold [3.20.20], and belongs to the triose phosphate isomerase homologous family [3.20.20.80]). These numbers are similar in concept to the EC numbers for enzyme classification and facilitate computational analysis and data mining. Most importantly for this analysis, we can automatically generate an annotated CATH wheel for any subset of proteins in the PDB (e.g. those with a given function).

Topology, homology and function

Analysis of all the structures in the PDB reveals that most proteins with the same topology belong to the same homologous family (i.e. they are evolutionarily related). This can be quantified by plotting the number of topologies with a given number of homologous families (Figure 9). The 'singlet' folds (i.e. those with one H family) comprise 93% of all topologies to date. It is our perception that most homologues have a similar or related function, although considerable further analysis is required to prove this. If this is the case, it follows that if two proteins adopt the same topology, they are likely to be related and have related functions. The exceptions to this rule are the 'superfolds' [27] – a small number of protein folds which recur frequently and have probably arisen more than once during evolution. These fold clusters appear as large segments in the outer circle of the CATH wheel in Figure 2. For example, 13 different homologous families adopt an $(\alpha\beta)_8$ TIM barrel structure and whilst almost all are enzymes, they have very different activities. Even in this set of apparently unrelated structures, it is well known that

Figure 9



The number of instances of topologies [C.A.T] in the PDB containing a given number of homologous families [C.A.T.H]. It is clear that the majority of topologies contain a single homologous family.

the binding sites are colocated at the C-terminal end of the parallel strands which form the barrel [28].

Here, we consider proteins which are not obviously related from sequence and structure analysis, and seek to explore the fundamental relationship between fold and function. There are many different ways to define and classify function (see [29] for discussion). Such classification schemes are difficult, sometimes ambiguous, and have yet to be applied systematically to all the proteins in the PDB. In this paper we only attempt to consider those aspects of function which are available electronically and can be determined semi-automatically – location, enzyme function and ligand-binding properties.

Implementation

We have previously linked the PDB to the Enzyme database in our 3D Enzyme Database [24], but for the current work, mapping individual PDB chains to EC numbers was far from straightforward and involved scanning SWISSPROT, the Enzyme Database and the data in the PDB entries using a number of programs written in Perl (ACRM).

A relational database was created using the freely available PostgreSQL database (http://www.PostgreSQL.org), linking the EC classification to the domain classification in CATH. This makes access and queries fast and simple. Again, programs written in Perl were used to populate the database from the raw CATH data, with additional information from the PDB files and EC number information.

Acknowledgements

The original software used to generate the CATH wheels was written by Alex Michie at University College London.

References

1. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. (1997). A genomic perspective on protein families. *Science* **278**, 631-637.
2. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
3. 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.
4. Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**, 231-234.
5. Bernstein, F.C., *et al*., & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
6. Acharya, K.R., Ren, J.S., Stuart, D.I., Phillips, D.C. & Fenna, R.E. (1991). Crystal-structure of human α-lactalbumin at 1.7 Å resolution. *J. Mol. Biol.* **221**, 571-581.
7. Wistow, G.J., Mulders, J.W.M. & Dejong, W. (1987). The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. *Nature* **326**, 622-624.
8. Barth, A., Wahab, M., Brandt, W. & Frost, K. (1993). Classification of serine proteinases derived from steric comparisons of their active sites. *Drug Des. Discov.* **10**, 297-317.
9. Wallace, A.C., Laskowski, R.A. & Thornton, J.M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in serine proteinases and lipases. *Protein Sci.* **5**, 1001-1013.
10. Wallace, A.C., Borkakoti, N. & Thornton, J.M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308-2323.
11. Hubbard, S.J., Campbell, S.F. & Thornton, J.M. (1991). Molecular recognition and conformational-analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* **220**, 507-530.
12. Tong, L., Wengler, G. & Rossmann, M.G. (1993). The refined structure of Sindbis virus core protein in comparison with other chymotrypsin-like serine proteinase structures. *J. Mol. Biol.* **230**, 228-247.
13. Thornton, J.M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261-287.
14. Dijkstra, B.W., Kalk, K.H., Hol, W.G. & Drenth, J. (1981). Structure of bovine pancreatic phospholipase A2 at 1.7 Å resolution. *J. Mol. Biol.* **147**, 97-123.
15. Nakashima, H. & Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. *J. Mol. Biol.* **238**, 54-61.
16. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). (1992). *Enzyme Nomenclature.* Academic Press, New York, NY.
17. Bairoch, A. (1996). The ENZYME databank in 1995. *Nucleic Acids Res.* **24**, 221-222.
18. Williams, R.J.P. (1993). Are enzymes mechanical devices? *Trends Biochem. Sci.* **18**, 115-117.
19. Fothergill-Gilmore, L.A. (1986). Domains of glycolytic enzymes. In *Multi-domain Proteins – Structure and Evolution*. (Coggins, D.G., ed), pp. 85-174, Elsevier, Amsterdam.
20. Baker, S.C., Saunders, N.F.W., Willis, A.C., Ferguson, S.J., Hajdu, J. & Fülöp, V. (1997). Cytochrome cd_1 structure: unusual haem environments in a nitrile reductase and analysis of factors contributing to β-propeller folds. *J. Mol. Biol.* **269**, 440-455.
21. Henrissat, B. & Davies, G. (1997). Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol.* **7**, 637-644.
22. Berman, H.M., *et al*., & Schneider, B. (1992). The nucleic-acid database – a comprehensive relational database of 3-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751-759.
23. Rossmann, M.G., Moras, D. & Olsen, K.W. (1974). Chemical and biological evolution of a nucleotide binding protein. *Nature* **250**, 194.
24. Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L. & Thornton, J.M. (1997). PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**, 488-490.
25. Hendlich, M., Rippmann, F., Barnickel, G., Hemm, K. & Aberer, K. (1996). RELIBase – an object-oriented comprehensive receptor-ligand database. *Abstr. Papers Am. Chem. Soc.* **211**, 49.
26. Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 208-229.
27. Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631-634.
28. Branden, C. & Tooze, J. (1991). *Introduction to Protein Structure.* Garland Publishing, Inc., London, UK.
29. Riley, M. (1993). Functions of the gene products of *Escherichia coli. Microbiol. Rev.* **57**, 862-952.
30. Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946-950.
31. Frolow, F., Kalb, A.J. & Yariv, J. (1994). Structure of a unique, two-fold symmetrical haem-binding site. *Nat. Struct. Biol.* **1**, 453-460.
32. Liao, D.-I., Kapadia, G., Ahmed, H., Vasta, G.R. & Herzberg, O. (1994). Structure of S-lectin, a developmentally regulated vertebrate β-galactoside-binding protein. *Proc. Natl Acad. Sci. USA* **91**, 1428-1432.
33. Feng, J.-A., Johnson, R.C. & Dickerson, R.E. (1994). HIN recombinase bound to DNA: the origin of specificity in major- and minor-groove interactions. *Science* **263**, 348-355.
34. Dunn, C.R., *et al*., & Holbrook, J.J. (1991). Design and synthesis of new enzymes based on the lactate dehydrogenase framework. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **332**, 177.