Article No. jmbi.1999.3054 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 293, 333–342

JMB

Protein Folds, Functions and Evolution

Janet M. Thornton[1,2]*, Christine A. Orengo[1], Annabel E. Todd[1] and Frances M. G. Pearl[1]

[1]*Biochemistry and Molecular Biology Department, University College London, University of London, Gower Street, London WC1E 6BT, UK*

[2]*Crystallography Department, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK*

**Corresponding author*

The evolution of proteins and their functions is reviewed from a structural perspective in the light of the current database. Protein domain families segregate unequally between the three major classes, the 32 different architectures and almost 700 folds observed to date. We find that the number of new topologies is still increasing, although 25 new structures are now determined for each new topology. The corresponding analysis and classification of function is only just beginning, fuelled by the genome data. The structural data revealed unexpected conservations and divergence of function both within and between families. The next five years will see the compilation of a definitive dictionary of protein families and their related functions, based on structural data which reveals relationships hidden at the sequence level. Such information will provide the foundation to build a better understanding of the molecular basis of biological complexity and hopefully to facilitate rational molecular design.

Keywords: protein structure; protein function; enzymes; protein folds; protein evolution

Introduction

Within one year, it is likely that the sequence for the human genome will be largely determined. Many bacterial genomes are already completed. The challenge of the next century will be to understand and characterise how these proteins perform all the biological functions needed for life. Understanding these processes at the molecular level, requires knowledge of 3D structures, which reveal the details of binding, catalysis and signalling.

The number of known three-dimensional protein structures has increased rapidly over the last five years, with almost 10,000 entries currently in the Protein Data Bank (PDB; Bernstein *et al.*, 1978). The current rate of deposition is about 5 a day. Individually, these structures reveal much about the specific molecular mechanisms which underlie a particular biological function. However, taken together this body of data allows the biologists to explore the evolution of structure and function and the emergence of complex biochemical pathways. Since structure is much better conserved than sequence, these data facilitate the recognition of evolutionary relationships which are hidden at the sequence level. Such studies support the hypothesis that the number of protein families in biology is finite, with a minimal estimate of only 1000 different protein folds (Chothia, 1993).

In order to begin to understand and map this universe of protein structures, it is necessary to collect, collate, annotate and classify these structures in a rational scheme. This molecular taxonomy parallels the work of classical taxonomists, who sought to cluster plants and animals into species, but molecular structure now provides a more sensitive probe. Over the last ten years, several groups have evolved schemes to help to classify structures, usually based on a combination of sequence and structure comparison methods (e.g. SSAP, Taylor & Orengo, 1989; DALI, Holm & Sander, 1993a). There have been many attempts to provide lists of structural neighbours (VAST, Hogue *et al.*, 1996; DALI, Holm & Sander, 1996; Overington *et al.*, 1993; DIAL, Sowdhamini *et al.*, 1996), but several groups have taken this further in an attempt to generate more complete classification schemes, which provide phenetic descriptions of structure as well as phylogenetic relationships (e.g. SCOP,

E-mail address of the corresponding author: thornton@biochem.ucl.ac.uk

0022-2836/99/420333–10 $30.00/0

Murzin *et al*., 1995; CATH Orengo *et al*, 1997; HOMSTRAD Mizuguchi *et al*., 1998)

To date most of our knowledge about the world of proteins derives from the sequence data. However, sequence comparisons fail to identify many of the relationships which emerge once the structure of a protein is known. Brenner *et al*. (1998) showed that the extensively used BLAST algorithm finds only 10% of known relationships in the PDB. The sophisticated iterative PSI-BLAST approach (Altschul *et al*., 1997), which builds profiles for a given family, is much more sensitive, but still misses many such relationships. In the light of these results, it is clear that the definitive dictionary of protein families will require structural data to map all the relationships.

Here, we present our view of the classification of protein structure in the light of the current data in the PDB. We also consider how function relates to structure, and the evolution of new structures and functions.

Protein structure classification

Phylogenetic and phenetic classification

There are two distinct approaches to protein structure classification. The phylogenetic approach seeks to provide an evolutionary tree, relating different protein families according to their evolutionary history (i.e. grouping together proteins with a common ancestor). Theoretically, there is a single ''correct'' result for this classification (excluding complex phenomena such as the exchange of genetic material between species), which will gradually emerge and stabilise as more data become available. The phenetic approach provides higher level descriptors of protein structure (such as class, architecture and fold type) without reference to their evolutionary past, to describe and group the proteins according to their structural characteristics. Such descriptors by nature are not absolute, but very useful. In practice, it is observed that proteins that are grouped in the same homologous protein family by sequence comparisons, all adopt the same 3D fold. Therefore the most common approach to organising protein structure data is a hierarchical classification scheme, first grouping proteins according to their phylogenetic relationships, and then clustering them further by fold, architecture and class structural descriptions. Such classifications will be most successful if the distributions they seek to classify are discrete, rather than forming a continuum. For protein structure it has been found that the distributions are sufficiently discrete to allow a useful classification (for a review, see Orengo *et al*., 1997), although in highly populated areas of fold space (e.g. the three-layer αβα sandwiches) the segregation of families into distinct fold groups can be difficult.

Domains

From sequence analysis it is observed that larger proteins often comprise recognisable smaller sequence domains, which recur in other proteins in various combinations. These domains can be thought of as ''units of evolution''. The structural data similarly reveal that larger proteins often comprise distinct structural units, which are compact and often local in sequence. There is evidence that some of these structural domains can fold independently. These domains can be covalently linked to generate multi-modular proteins, which may have extensive or minimal contact between domains. One domain may also be inserted into another, splitting the original structure to form a discontinuous domain, comprising two or more non-local segments of chain. Therefore it has long been recognised that an organisation of protein structure can only be sensibly approached by classification at the domain level.

Thus all classification schemes attempt to subdivide proteins into domains. This is a difficult procedure, since even the structural data are often indeterminate, i.e. two different algorithms will give different domain assignments (Swindells, 1995; Siddiqui & Barton, 1995; Islam *et al*., 1995; Holm & Sander, 1993b; Jones *et al*., 1998). More recent approaches attempt to incorporate both structural and sequence data across the archive data, to identify ''recurrent domains'', with a view to defining the most likely ''evolutionary unit'' (Holm & Sander, 1998). The definitions will hopefully become increasingly refined as more data become available, although the ''decoration'' of core domain folds by adding additional secondary structures is frequently observed and complicates the assignments.

The CATH classification of protein domains

A very brief description of our approach to protein structure classification is presented below, with some comparisons to other approaches included to highlight similarities and differences. Details can be found in a study by Orengo *et al* (1997) and at http://www.biochem.ucl.ac.uk/bsm/cath/.

In CATH, domain boundaries are assigned by inheritance (for proteins with significant sequence similarity to a protein already in the database) or by consensus if three different algorithms identify the same structural domain boundaries or by hand if the algorithms provide different results. The aim is to identify ''structural'' domains, defined as semi-independent structural units. In SCOP the assignments are made by hand and proteins are only split into domains if one of the domains occurs independently elsewhere in the database. Thus the serine proteinase fold is divided into two β-barrel domains in CATH, but occurs as a single ''domain'' in SCOP. In total in CATH we currently have 6367 entries and 14,518 domains. Since both

the SCOP and CATH classifications require some manual intervention, there is usually some time delay between a structure being available in PDB and being included in the classification. In contrast, the DALI and VAST systems are fully automated and therefore more up-to-date.

Protein structures are first grouped according to sequence and structure similarity measures, the latter being calculated using the SSAP algorithm by Taylor & Orengo (1989). This calculates a length-independent score between 0-100. Empirical cut-offs provide a range of definitive positive and negative matches (e.g. a SSAP score >80 generally occurs between a pair of homologous proteins), but there remains a twilight zone where additional data (e.g. function) is used to certify the assignments. Proteins are merged into the same homologous family if they fulfil one of the following conditions: (i) evidence of significant sequence similarity; (ii) evidence of significant structural similarity with a weaker sequence similarity; or (iii) in the absence of any sequence similarity, evidence of significant structural similarity, combined with a functional similarity at a co-located active site.

Proteins which show a significant structural similarity not necessarily linked to sequence or functional similarity are then clustered into the same fold/topology group. Of course as with sequence analysis, there is a ''twilight zone'' in structure comparisons, where it is uncertain whether two proteins have arisen through convergent or divergent evolution. Indeed it is not clear how far divergent evolution can go (Murzin, 1998), and one advantage of clustering homologous superfamilies together into fold groups is to highlight the possibility of identifying such distant relationships.

The class (mainly α, mainly β, $\alpha\beta$, low secondary structure) is assigned using an algorithm developed for this purpose (Michie *et al*., 1996). In SCOP the $\alpha\beta$ class is further subdivided into the alternating α/β and $\alpha+\beta$ classes (Levitt & Chothia, 1976). In our analysis, we found that although many proteins clearly fell into one or the other category, there were too many proteins lying on the boundary to allow a confident assignment. Since ''class'' is at the highest level of the hierarchy, we decided to combine the α/β and $\alpha+\beta$ classes into a single $\alpha\beta$ class to avoid splitting homologues into different classes.

The architecture assignment, which defines the secondary structure packing regardless of chain connectivity, is assigned by hand, but is currently being automated. This level, which is absent in SCOP, allows the folds in a given class to be grouped together if their secondary structures pack in a similar way, reflecting the stereo-chemical principles of helix/sheet packing.

Current status of protein structure data

The current data in CATH, which include all data released by PDB in 1997 (but also more recent entries in PDB which are sequence homologues of old entries), are shown in Figures 1 and 2. From 1993 the number of entries in the PDB started to rise rapidly, although the increase has been approximately linear over the last four years (Figure 1(a)). As expected, the number of domains closely follows the number of entries, with approximately 2.3 domains per entry (Figureure 1(b)). Interestingly, up to the end of 1997 the numbers of new homologous families and new topologies are also rising rapidly and have not yet reached saturation (Figure 1(c) and (d)). The CATH ratios in Figure 2(a) and (b) show that a new topology is only observed for approximately 25 new entries in the PDB and the average number of 1.6 homologous families per fold has been constant for the last few years.

The CATH wheel (Figure 3) shows the universe of protein structures derived using one representative for each homologous family. This use of representatives is important, since clearly many families have multiple entries in the sequence and structure databases, introducing considerable bias if not treated appropriately. This analysis is fundamentally different from the analyses of whole genomes (e.g. Gerstein, 1997; Tatusov *et al*., 1997), where it is clear that gene duplication has played a major role in amplifying the genome so that some genes have over 50 copies even in small genomes (Teichmann *et al*., 1998). Such duplications radically change the observed distributions. Our analysis concentrates on identifying and characterising protein domain families as the fundamental building blocks used in the evolution of complex biological structure. From this perspective we see that approximately 20% of all families are mainly α, 25% are mainly-β and approximately 50% are mixed $\alpha\beta$. Very few folds have little or no secondary structure content.

The structural data on their own do not yet give any indication of the total number of protein families or folds. If there are only 1000 folds then we can expect to see a significant levelling-off in Figure 1(c) and (d) over the next three years. The genome sequence data provide more information, but given the problems of recognising distant relatives, they are difficult to analyse accurately. Attempts to cluster all proteins into families at the sequence level give results which are entirely dependent on the method used and the cut-off applied. Until recently, most used BLAST which is known to miss 90% of the relations in the PDB and all methods leave many sequences unclassified, which have been termed ''singletons''. Are these really proteins which only occur in one species? This seems unlikely and probably represents our inability to recognise distant homologues. As more structures are determined and relationships accurately identified, we can expect a more accurate picture of the diversity and ubiquity of protein families.

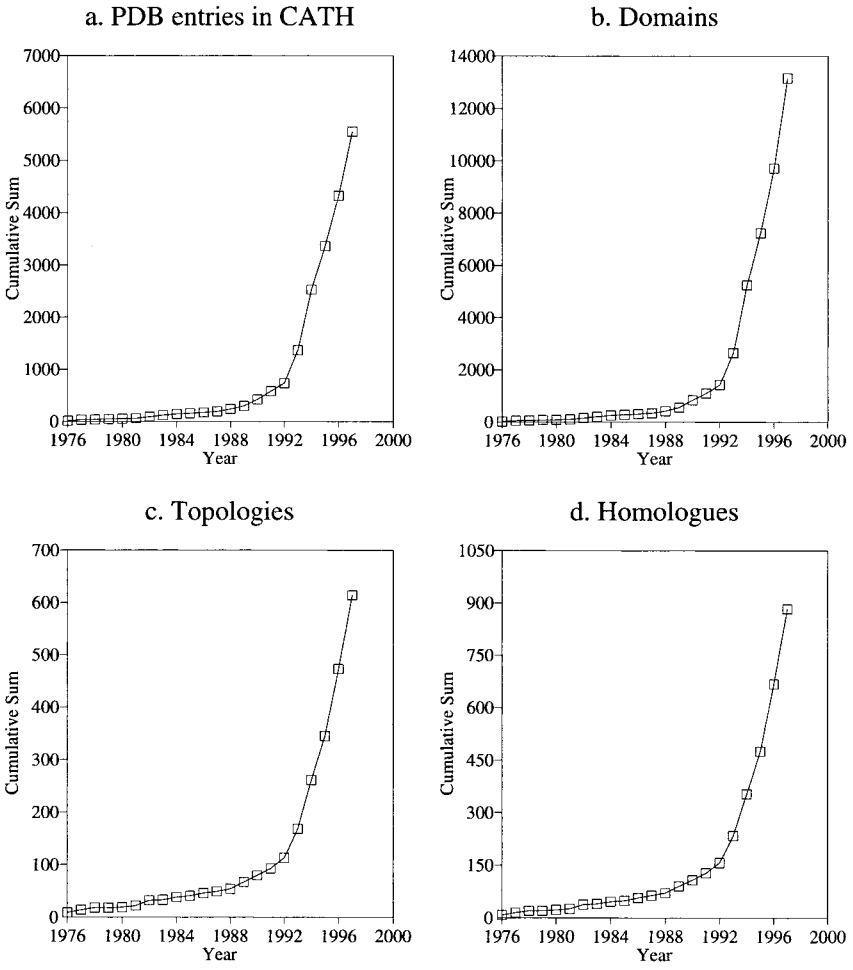

Figure 1. Database statistics for CATH: cumulative totals by year. (a) PDB entries; (b) domains; (c) topologies; (d) homologous families.

Bias in the PDB

The proteins analysed here are those for which structures have been determined. This selection inevitably introduces a bias, principally towards smaller proteins and those which crystallise. Historically there has also been a bias towards determining the structures of enzymes, which constitute 58% of homologous families in CATH. However the principles derived from such structures are expected to hold for all globular water-soluble proteins.

The size distribution of protein chains in PDB is shown in Figure 4. The mean chain length in PDB is 367 amino acid residues, compared to 287 in *Methanococcus jannaschi* and 484 residues in *Saccharomyces cerevisiae*. In our view, the size factor is unlikely to distort our understanding of the protein universe, since the larger proteins, whose sequences have been determined, clearly comprise smaller domains such as those typically found in the PDB. However, it is difficult to crystallise ''flexible'' proteins and membrane-bound proteins. Both these classes of proteins are known to be common, with membrane-bound structures representing over 20% of eukaryotic genomes. Flexible proteins fall into two categories: those which only adopt a unique structure when bound to a receptor or ligand, and those which comprise structured domains joined by flexible linkers.

Clearly flexibility and conformational change are an integral part of most biological functions, and must not be forgotten in the analysis of a crystal structure in which gross conformational changes are impossible. There are also the highly repetitive low complexity sequences, whose structures and functions are often unknown.

Distribution of families in protein structure space

As more structures are determined, it becomes increasingly clear that the distribution of homologous families between the different architectures and folds is not even. As we first noted in 1994 (Orengo *et al*., 1994), certain folds are very common (which we called the superfolds), whilst others so far have only been seen for a single evolutionary family. Similarly, the distribution of folds between architectures is not even, with the three-layer αβα sandwich and two-layer αβ sandwich dominating the αβ class of structures. The current data are shown in Figure 5. Analysis of genome sequences

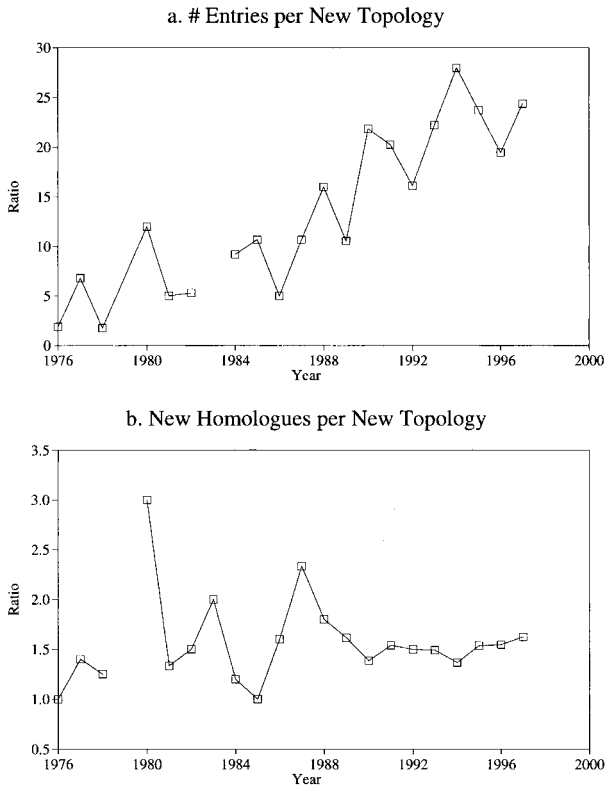

Figure 2. CATH ratios for structures deposited in each year. (a) Number of entries/Number of topologies. (b) Number of homologous families/Number of topologies.

(e.g. Gerstein, 1997) shows that folds which are very common from our filtered analysis of representatives of homologous families, are also often over-represented in genomes, presumably as a result of extensive gene duplication. For example, Teichmann *et al* (1998) found that there are 51 members of the P-loop nucleotide triphosphate hydrolase superfamily in the *Mycoplasma genitalium* genome. It may be that such duplications have allowed evolutionary divergence to proceed more rapidly and further than for other families.

It is interesting to speculate on why some folds/architectures dominate the universe of structures. We recently explored one possible explanation (Salem *et al*., 1999) by analysing the occurrence of supersecondary structure in the different folds. We found some evidence that more secondary structures adopt classic supersecondary motifs (i.e. αα, ββ hairpins, the αβα unit and Greek key structures) in common folds than in rarer folds. i.e. folds with more local interactions between secondary structures are more common. Another possible explanation relates to the mechanism by which new sequences and new folds are generated. The common folds often incorporate some element of repetition, e.g. the eight-fold βα repeat in the TIM barrel, which may reflect their evolution through repetitive duplication of a small stable unit. Such a mechanism to generate new structures may have dominated early stages of protein structure evolution.

Protein structure/function relationships

Protein function classification and evolution

The function of a protein can be defined at various levels, from biochemical through cellular to physiological function. Whilst the native structure is an absolute requirement for activity, it remains very difficult to predict even the biochemical function from structure except by recognition of similarity to a protein of known function. With the advent of structural genomics, structural information will increasingly be used to provide functional clues to guide experiments. Before considering the structural data, it is useful to review how new functions can be created (see Figure 6).

Conceptually, the simplest route to create a new function is to create a new protein *ab initio*. However there are many alternative routes available during evolution which are summarised below. Increasingly evidence is presented that many gene products have more than one biochemical function, depending on their biological context (Jeffery, 1999). The functions of these ''moonlighting'' proteins may be modulated by location, pH, ligand availability etc. Such multi-functional proteins will cause complications for genome annotation. During evolution, possibly following gene duplication, one of the gene products may be released from functional constraints, to evolve a new function by the accumulation of local mutations. There are many examples where enzymes have modulated their functions in this way. Other possibilities involve oligomerisation (homo- or hetero-), duplication, or molecular construction of a new gene containing two or more domains created from the pool of fundamental domains.

If there is a rather small number of basic protein folds, probably no more than several thousand, the need for many more biological functions suggests that modulation of function has been a major route to evolve new functions. Analysis of structure/function relationships is hampered by the availability of functional information and an ''agreed'' functional classification scheme. Below we concentrate on data for enzymes in the PDB, for which the enzyme classification (E.C.) numbers (NC-IUBMB, 1992) provide a fairly complete framework.

The structures and functions of enzymes

There are currently 556 enzyme families in CATH, comprising 58% of all families. Almost half (45%) of all enzyme families in the PDB are multi-domain, often comprising a catalytic domain fused with a nucleotide-binding domain. The distribution

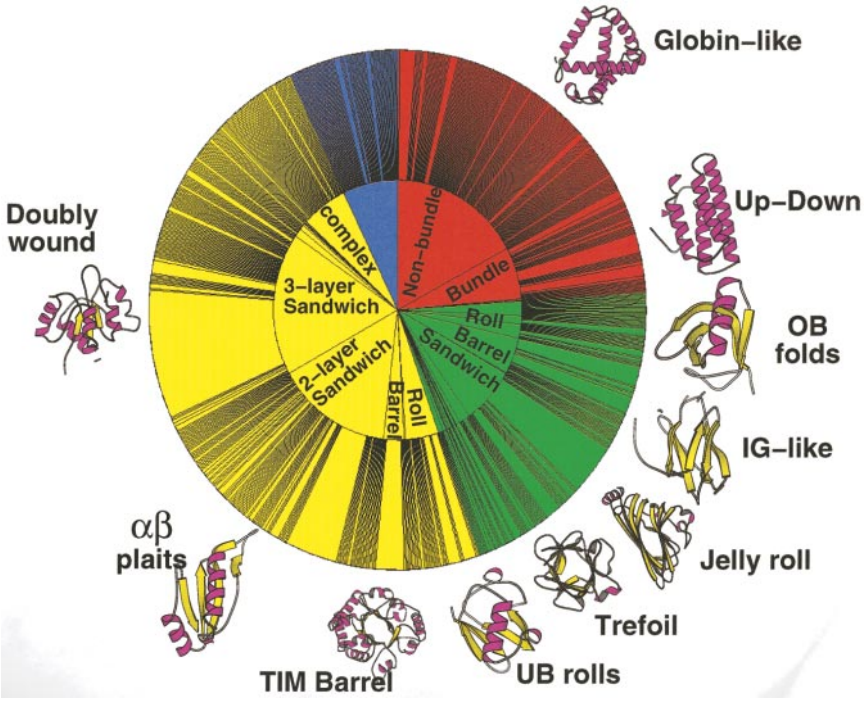

Figure 3. The protein structure universe in the PDB (1997) as illustrated by a CATH wheel, showing the population of homologous families in different fold groups, architectures and classes. The wheel is coloured according to protein class (red, mainly α; green, mainly β; yellow, αβ; blue, few secondary structures). The outer wheel shows the population of families between the different fold groups, whilst the inner wheel shows the population in the different architectures. The protein cartoons were drawn using MOLSCRIPT (Kraulis, 1991).

of these domains amongst the major structural classes, compared to that observed for all non-enzymes in the PDB, is shown in Figure 7(a) and (b). From these plots, it can be seen that the majority of enzymes are αβ proteins, with the mainly-α and mainly-β proteins being under-represented relative to the whole PDB. A more complete analysis is presented by Martin *et al*. (1998) and more recently by Hegyi & Gerstein (1999). We also show the same distribution for the nucleotide-binding domains alone, and for the 11 enzymes in the glycolytic pathway (Figure 7(c) and (d)). The nucleotide-binding domains, which provide critical coenzyme binding, are dominated by the αβ proteins, as are the enzymes of the glycolytic pathway, which only utilise three architectures of the αβ class. At the extreme, these two observations could be interpreted as evidence that evolutionary factors have dominated both the development of ancestral proteins with coenzyme binding functions (whose specificity and structure have been modulated over time) and the development of complex biochemical pathways.

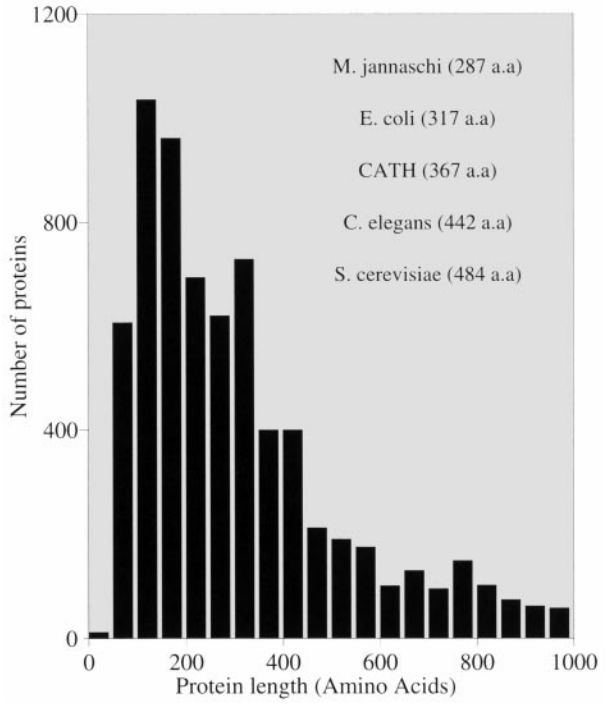

Figure 4. Length distribution of all polypeptide chains in the PDB.

Multiple enzyme functions in one homologous family: a problem for genome annotation

Detailed inspection of the structural and functional data quickly reveals that many enzyme families include members with different catalytic activities, as illustrated in Figure 8(a) and (b). Of the 190 enzyme homologous families with multiple members in the PDB, about half (91) have members with different E.C. classifications, and

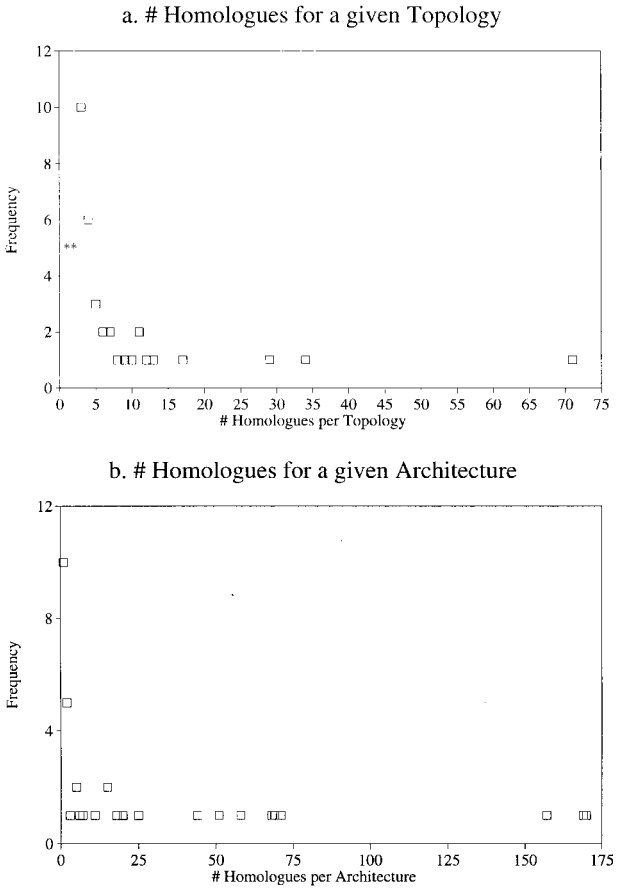

Figure 5. Distribution of families between the different topologies and architectures. (a) Number of homologous families in each topology; an asterisk (*) represents data which has been omitted from the plot. There are 531 topologies which include only one homologous family and 36 topologies with two families. (b) Number of homologous families in each architecture.

28 families include members which are not enzymes. Figure 8 shows the conservation of E.C. number within the enzyme families within CATH. In some families (17/190) even the primary E.C. number is not conserved, whilst others show minor changes in the third or fourth digit of the enzyme classification. Some families include proteins which are not enzymes at all and others have different numbers of E.C. ''functions''. Figure. 8(b) summarises these data by plotting the number of families with a given number of enzyme functions (defined as any change in E.C. number). This highlights the fact that enzyme function is often modulated or even changed completely during protein evolution. The reverse scenario, in which the same (E.C.) function is catalysed by two unrelated enzymes, as typified by the subtilisin and trypsin-like serine proteinases, is also quite common. Galperin *et al*. (1998), using sequence analysis with some structural data, found 105 examples in Genbank. The structural data reveal that, at least in some

- New gene - New protein - New function.
- Moonlighting - The same gene product may perform different functions in different environments, i.e. function is context sensitive, dependent on cell location, pH, available ligands etc.
- Incremental mutational evolution - The protein evolves through local mutational events to perform a different, though often related, function.
- Oligomerisation - The gene products form oligomers with new functions. These can comprise identical, related or completely different sub-units.
- Duplication - Gene duplication can give rise to a novel gene product with novel function.
- Modular construction (mix and match) - Separate gene products may be joined to generate proteins with novel activities.

NB. These alternative routes are often combined in practice to generate new biochemical functions. The biological functions *in vivo* are always context sensitive.

Figure 6. The evolution of a new protein function.

of these examples, the fold may be different but the catalytic mechanism is the same (see Wallace *et al*., 1997). Such observations suggest that the number of catalytically competent arrangements of the 20 amino acid residues may be limited,

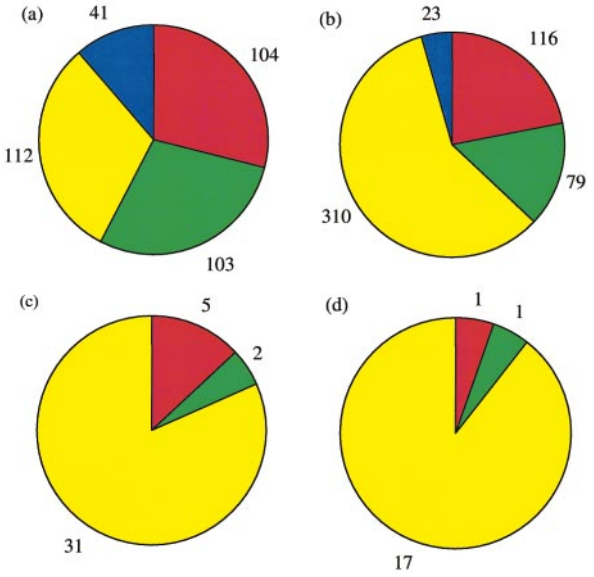

Figure 7. Class distribution of enzymes in the PDB. The wheels are coloured according to protein class (red, mainly α; green, mainly β; yellow, αβ; blue, few secondary structures). (a) all non-enzyme domains; (b) all enzyme domains; (c) nucleotide-binding domains; (d) the 19 domains in the 11 enzymes of the glycolytic pathway.

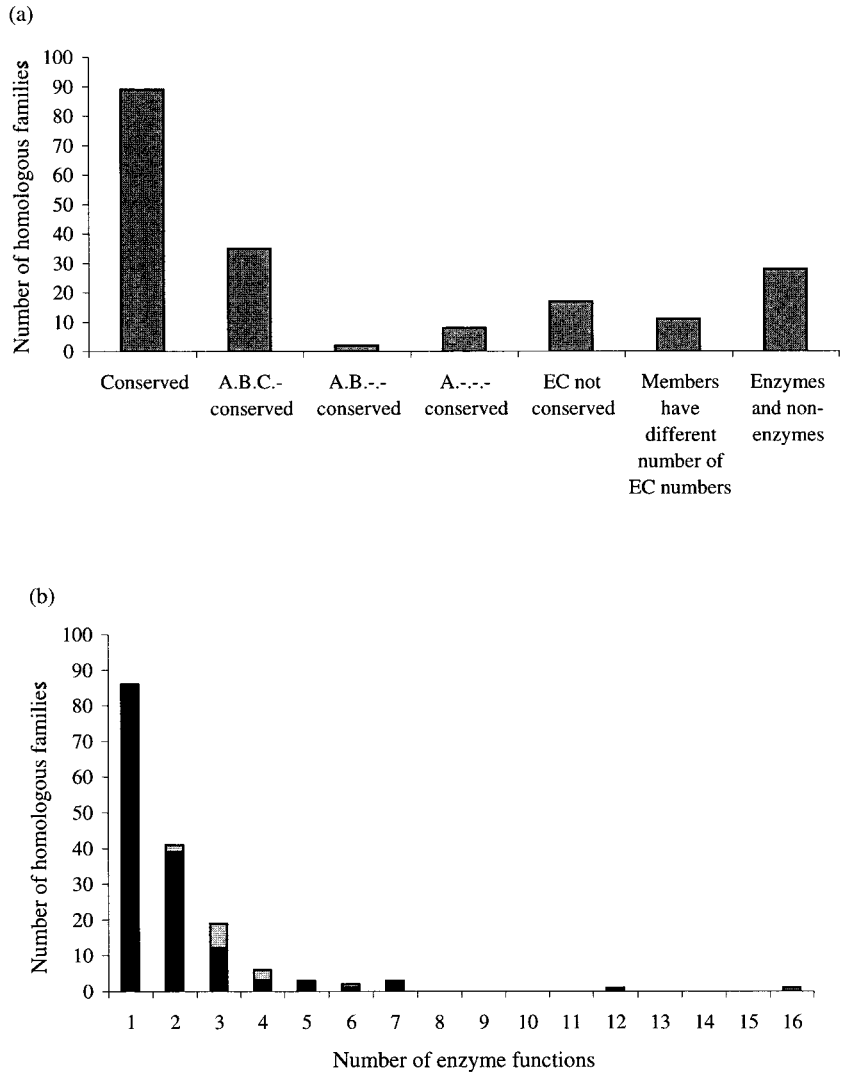

Figure 8. The conservation of enzyme function as defined by E.C. number within homologous protein families in CATH. A total of 190 enzyme families were identified in the PDB, each containing more than one non-identical entry (i.e. each family has at least two members). The E.C. number for a given PDB entry was extracted *via* the corresponding SWISSPROT file, identified using the EBI list (see http//:www2.ebi,ac.uk/msd/3Dseq) or in-house software. (a) Conservation of E.C. number; (b) number of E.C. functions within homologous families. Grey shading represents homologous families which contain enzymes with more than one associated enzyme function.

since the same mechanism has evolved independently several times during evolution. Combining these data with gene recruitment, in which the function is context sensitive, it is clear that functional annotation from sequence, even at the biochemical level, will require much more than just identification of homology. It will be necessary to consider specific mutations within sequences and methods must be developed to try to predict the modulation of function with sequence changes. The annotation of sequence data by structure information is one way forward (see Milburn *et al.*, 1998).

Protein families and protein evolution

The next five years will see the determination of many new sequences and structures. Together these data will allow us to identify and characterise the basic set of protein families, to assess their diversity and distribution amongst the kingdoms of life and to understand better how they perform their biochemical and biological functions. At this stage, many questions remain unanswered. How many protein families exist? How did complex pathways evolve? How do proteins fold? How were the first structures formed, before conservation of structure became such a powerful constraint? In all possible sequence space, what fraction of sequences fold into a unique native structures? Were the basic set of structures evolved before the three kingdoms of life separated? Are new folds being made today?

The fundamental problem of predicting structure from sequence remains an intellectual challenge, though it will be overtaken for practical purposes by knowledge-based recognition approaches, such as threading. As more structures are determined, our understanding of how function is modulated by sequence should improve, allowing the design of proteins with novel functions. The other fundamental challenge of calculating free energy of binding, which must be at the heart of the successful structure-based design of novel ligands as thera-

peutics, will undoubtedly benefit from the flood of structural data, combined with accurate biophysical measurements. Whilst the sequence data give us the ''blueprint'' of life, the new structural data we can expect in the next five years will provide the stepping stones to begin to understand biological function at the molecular level.

Acknowledgements

We thank Andrew Harrison for his help with generating Figures. We acknowledge the support of the UK BBSRC and MRC. A.E.T. is supported by a BBSRC Case studentship in collaboration with Oxford Molecular Ltd.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci.* **26**, 6073-6078.

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

Galperin, M. Y., Walker, D. R. & Koonon, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome. Res.* **8**, 779-790.

Gerstein, M. (1997). A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.

Hogue, C. W. V., Ohkawa, H. & Bryant, S. H. (1996). WWW-Entrez and the molecular modelling database. *Trends. Biochem. Sci.* **21**, 226-229.

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

Holm, L. & Sander, C. (1993b). Parser for protein folding units. *Proteins: Struct. Funct. Genet.* **19**, 256-268.

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

Holm, L. & Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins: Struct. Funct. Genet.* **33**, 88-96.

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

Jeffery, C. J. (1999). Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8-11.

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.

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

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

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. & Thornton, J. M. (1998). Protein folds and functions. *Structure,* **6**, 875-884.

Michie, A. D., Orengo, C. A. & Thornton, J. M. (1996). Analysis of domain structural class using an automated class assignment tool. *J. Mol. Biol.* **262**, 168-185.

Milburn, D., Laskowski, R. A. & Thornton, J. M. (1998). Sequences annotated by structure: a tool to facilitate use of structural information in sequence analysis. *Protein. Eng.* **11**, 855-859.

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

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.

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

Nomenclature Committee of the International Union of Biochemistry Molecular Biology (NC-IUBMB) (1992). *Enzyme Nomenclature*, Academic Press, New York, NY.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **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.

Overington, J. P., Zhu, Z. Y., Sali, A., Johnson, M. S., Sowdhamini, R., Louie, G. V. & Blundell, T. L. (1993). Molecular recognition in protein families - a database of aligned three-dimensional structures of related proteins. *Biochem. Soc. Trans.* **21**, 597-604.

Salem, G. M., Hutchinson, E. G., Orengo, C. A. & Thornton, J. M. (1999). Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.* **287**, 969-981.

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 representative family members into similar folds. *Fold. Design*, **1**, 209-220.

Swindells, M. B. (1995). A procedure for detecting structural domains in proteins. *Protein Sci.* **4**, 103-112.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science,* **278**, 631-637.

Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 208-229.

Teichmann, S., Park, J. & Chothia, C. (1998). Structural assignments to the proteins of *Mycoplasma genitalium* show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA,* **95**, 14658-14663.

Wallace, A. C., Borkakorti, 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.