CATH – a hierarchic classification of protein domain structures

CA Orengo[1], AD Michie[1], S Jones[1], DT Jones[2], MB Swindells[3] and JM Thornton[1,4]*

Background: Protein evolution gives rise to families of structurally related proteins, within which sequence identities can be extremely low. As a result, structure-based classifications can be effective at identifying unanticipated relationships in known structures and in optimal cases function can also be assigned. The ever increasing number of known protein structures is too large to classify all proteins manually, therefore, automatic methods are needed for fast evaluation of protein structures.

Results: We present a semi-automatic procedure for deriving a novel hierarchical classification of protein domain structures (CATH). The four main levels of our classification are protein class (C), architecture (A), topology (T) and homologous superfamily (H). Class is the simplest level, and it essentially describes the secondary structure composition of each domain. In contrast, architecture summarises the shape revealed by the orientations of the secondary structure units, such as barrels and sandwiches. At the topology level, sequential connectivity is considered, such that members of the same architecture might have quite different topologies. When structures belonging to the same T-level have suitably high similarities combined with similar functions, the proteins are assumed to be evolutionarily related and put into the same homologous superfamily.

Conclusions: Analysis of the structural families generated by CATH reveals the prominent features of protein structure space. We find that nearly a third of the homologous superfamilies (H-levels) belong to ten major T-levels, which we call superfolds, and furthermore that nearly two-thirds of these H-levels cluster into nine simple architectures. A database of well-characterised protein structure families, such as CATH, will facilitate the assignment of structure–function/evolution relationships to both known and newly determined protein structures.

Addresses: [1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK, [2]Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK, [3]Helix Research Institute, Yana 1532-3, Kisarazu T292, Japan and [4]Crystallography Department, Birkbeck College, Malet street, London WC1E 6BT, UK.

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

Key words: evolution, fold families, protein structure classification

Received: 4 April 1997
Revisions requested: 23 May 1997
Revisions received: 17 June 1997
Accepted: 17 July 1997

Structure 15 August 1997, 5:1093–1108
http://biomednet.com/elecref/0969212600501093

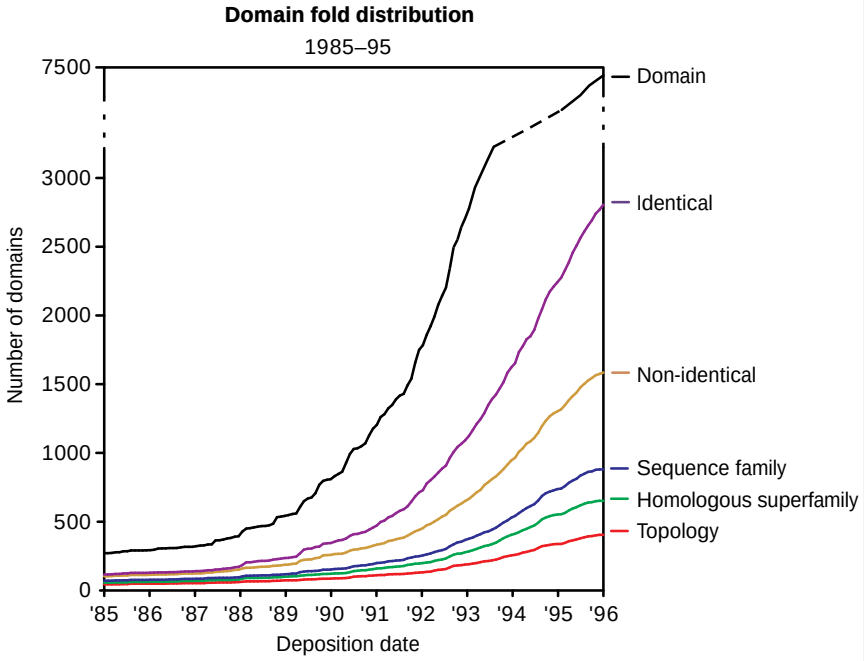
Introduction

As the number of sequences identified by the various genome projects increases at a phenomenal rate, it becomes correspondingly necessary to improve methods for predicting their three-dimensional (3D) structures. Understanding structural relationships between proteins, such as whether certain architectures occur more frequently than others, can inform these approaches but requires an appreciation of all the known structures. Because there are currently more than 5000 known structures, a number that increases by nearly 150 per month (see Figure 1) and with new structures appearing in the literature almost every day, manual inspection of all structures has become almost impossible. Therefore, fast and automatic methods are needed to evaluate the relationships between the known structures, particularly, as nearly three-quarters of the entries in the current Brookhaven Protein Database (PDB) are practically identical.

Any insights derived from grouping structures into families will depend on the criteria chosen for clustering them, which in turn should reflect the biological or physical causes for their similarity [1,2]. Global similarities between proteins may suggest an evolutionary relationship, which, because the structural core of a protein tends to be conserved, remains detectable even when sequences have diverged beyond any recognisable similarity. At 30% sequence identity, proteins will almost certainly have the same overall fold [3–5], but various studies have now uncovered many protein pairs with even lower sequence identities (e.g. <15% in the globin family) having both similar folds and functions [1,6–8]. In the absence of any clear evolutionary information, such structural similarity may simply be associated with preferred packing ensembles for secondary structure elements within a protein core.

Using standard sequence alignment methods [9] and the structure comparison algorithm, SSAP [10], we previously clustered the ~1800 well-resolved structures in the March 1993 release of the PDB [11,12] into 208 sequence-based homologous families, which further grouped into 112 unique structural or fold families. This analysis revealed

Figure 1



Annual increase in the numbers of protein domain structures in the PDB (top plot, [11,12]). The lower lines show the numbers of identical families (I-level, 100% sequence identity between structures within the family and 100% overlap), non-identical families (N-level, > 95% sequence identity, 85% overlap), sequence families (S-level, > 35% sequence identity, 60% overlap), homologous superfamilies (H-level, > 25% sequence identity, SSAP >80 and 60% overlap), and topological or fold families (T-level, SSAP >70), where SSAP is a structural comparison score.

some nine highly populated families ('superfolds' [1]), with important implications for prediction algorithms, and it illustrated the insights to be gained from ordering the data in this way.

Several other groups have also classified the known structures, focusing on a variety of local and global topological features and employing a range of algorithms (structure comparison algorithms and classification generally are reviewed in [13–16]). The SCOP database, developed by Murzin *et al.* [17], groups proteins having significant sequence similarity into homologous families, whereas more distant structural similarities are largely identified manually. This database places emphasis on evolutionary relationships and information from the literature relating to well-studied fold families is also incorporated (e.g. the β trefoils [18] and the OB fold [19]). By contrast Holm and Sander, use the structure comparison algorithm DALI to recognise structural neighbours, whether motif or fold based, without formally ordering proteins in the PDB into families [20]. The ENTREZ database of Hogue *et al.* [21], uses a similar approach to DALI, listing neighbours by a fast vector-based comparison algorithm (VAST).

The task of defining structural relationships is further complicated by the existence of multidomain proteins; more than 30% of non-identical structures in the current PDB contain two or more domains. A number of domain recognition algorithms have appeared recently to address this problem [22–26]. The 3Dee database of Siddiqui and Barton (http://snail.biop.ox.ac.uk:8080/3Dee) separates the constituent folds of multidomain proteins using the DOMAK algorithm. Similarly, Sowdhaminini *et al.* have constructed a database of single domain families [27], using the domain recognition algorithm DIAL [26] and the structural comparison procedure SEA [28]. Both databases contain data that is generated largely automatically, but is subsequently checked and where appropriate reordered manually.

In recognition of the need to regularly maintain and update data on structural relatives, we have further developed our automatic procedures for identifying and classifying structural families [6] to construct a database of single-domain fold families. Any multidomain proteins are first divided into their constituent domain folds by an automatic consensus procedure which is in agreement between three independent algorithms (SJ *et al.* unpublished data). As well as clustering proteins by sequence and structure, recognised families are also grouped according to similarity in protein class (i.e. secondary structure composition and contacts). Finally, the architecture (shape, defined by the assembly of secondary structures, regardless of their connectivity) adopted by each protein fold, is assigned manually. Although this is a somewhat subjective process, based largely on commonly used descriptions in the literature (e.g. sandwich, barrel and propellor), it is an essential first step towards ordering the known folds in a useful and practical way.

Subsequent analysis of these groupings will allow us to identify common structural features, which can be used to develop more automatic approaches for architecture classification in the future.

The structure classification procedure naturally results in a tree hierarchy outlining the relationships between folds. The data is stored in an hierarchical database (CATH), with each structure indexed by a CATH number akin to the EC nomenclature for enzymes. CATH has been made accessible in a hypertext form over the World Wide Web for use by text-based or graphical browsers (http://www.biochem.ucl.ac.uk/bsm/cath). The user can scan through the hierarchy of protein structures, with graphical representations at each level. Derived data such as structural alignments and protein family templates are also stored. A CATH lexicon and gallery describe different architectural levels and summarise data for each fold family, and a CATHserver will allow the user to scan a new protein structure against the CATH database of unique folds.

Results and discussion

Philosophy of the structural hierarchy

The five major levels in the CATH hierarchy — class, architecture, topology, homologous superfamily and sequence family (families with >35% sequence identity) — are described below. Each level is assigned a unique numeric label ('CATH number'). Numbers for different class levels were incremented by one (current range 1–4), whereas architecture (A), topology (T) and homology (H) numbers were incremented in steps of ten to allow new numbers to be assigned within these bins of ten and ensure that the CATH numbers remain constant as new structures are added to the database. Below the H-level, numbers are incremented in steps of one and therefore may change with new versions of the database.

Class (C-level): secondary structure composition and contacts

The first, most general level of the classification, class, describes the relative content of α helices and β sheets in a similar way to that described by Levitt and Chothia [29], except that we only define three major classes — mainly α, mainly β and α–β. Although the latter class can be subdivided into alternating α/β and α+β, our analysis of structural class [30] suggested that this can only be achieved automatically by taking into account secondary structure connectivity. In CATH, this information is considered at a lower level describing topology. At this level, CATH additionally groups all those structures having very low secondary structure content into a separate class.

Architecture (A-level): description of the gross arrangement of secondary structures, independent of connectivity

This level distinguishes structures in the same class with different architectures, but does not distinguish between different topologies (connectivities). The architectural groupings can sometimes be rather broad as they describe general features of protein-fold shape, for example, the number of layers in an α-β sandwich. A given architecture will contain structures with diverse connectivities (see Figure 2) which will be distinguished at the next level down (topology). For example, in the α-β class (C = 3), there are two common architectures each containing a large number of different fold families. One is the barrel-like architecture (A = 20) adopted, for example, by the TIM-barrel folds. These have an inner β barrel and an outer layer of α helices (Figure 2). Alternatively, the three-layer α-β sandwich architecture (A = 40) consists of a central β sheet which is covered by a layer of α helices on both sides of the sheet (Figure 2).

Topology (T-level): fold families

Structures which are grouped at the T-level have the same overall fold, which means that they have a similar number and arrangement of secondary structures and that the connectivity linking their secondary structure elements is the same. In this paper, the words fold and topology have the same meaning. Proteins with the same CAT numbers have the same class, architecture and topology but do not necessarily belong to the same homologous superfamily.

Topological description is given by reference to previously observed structures and well-known folds. Within a given topology level, the structures are similar, but may have diverse functions. Where possible the name chosen for a given T-level is either the name of the first structure in the family to be solved or the common name for the family (e.g. the globin fold or the immunoglobulin fold).

Homologous superfamily (H-level): highly similar structures and functional similarity

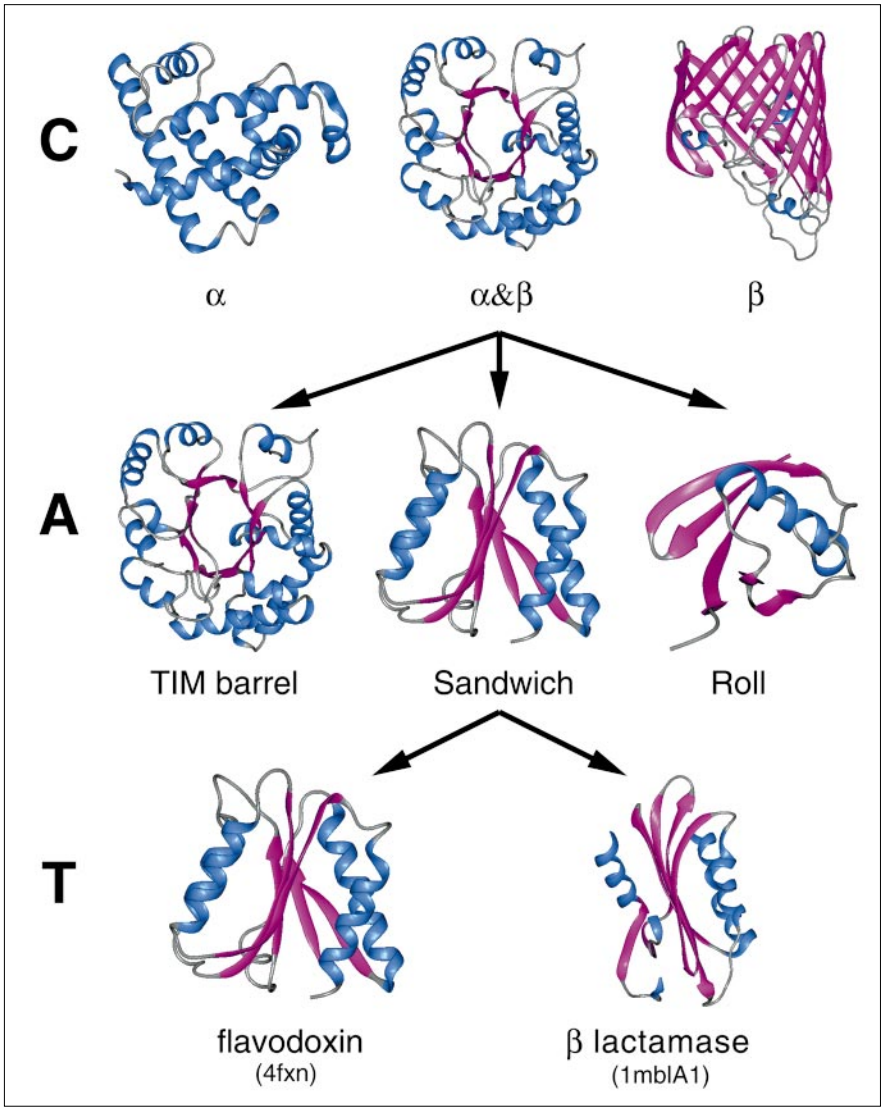
At the H-level, structures are grouped by their high structural similarity and similar functions, which suggest that they may have evolved from a common ancestor, particularly, where there are resemblances in core packing or putative active sites. Using the example of the mainly α.non-bundle. globin-like folds — the erythrocruorins, colicins, phycocyanins and domain 1 of diptheria toxin — all have the same CAT number (1.10.340), but are differentiated by their H numbers 10, 20, 30 and 40, respectively (see Figure 3).

Sequence family (S-level): significant sequence similarity and thus a high probability of having similar structure/function

Members which are clustered at this level (having the same CATHS number) have sequence identities >35% and as such are presumed to have extremely similar structures and functions — they may be slightly different examples of the same protein from different species belonging to the same sequence superfamily.

A detailed description of the CATH classification procedure is given in the Materials and methods section.

Figure 2



Schematic representation of the class (C), architecture (A) and topology (T) level in the CATH database. Helices are drawn in blue and strands are drawn as magenta arrows. The barrel, three-layer sandwich and roll architectures (A-level) are shown for the α–β class. Two representatives from fold families in the three-layer sandwich architecture are shown.

The structural universe as revealed by CATH

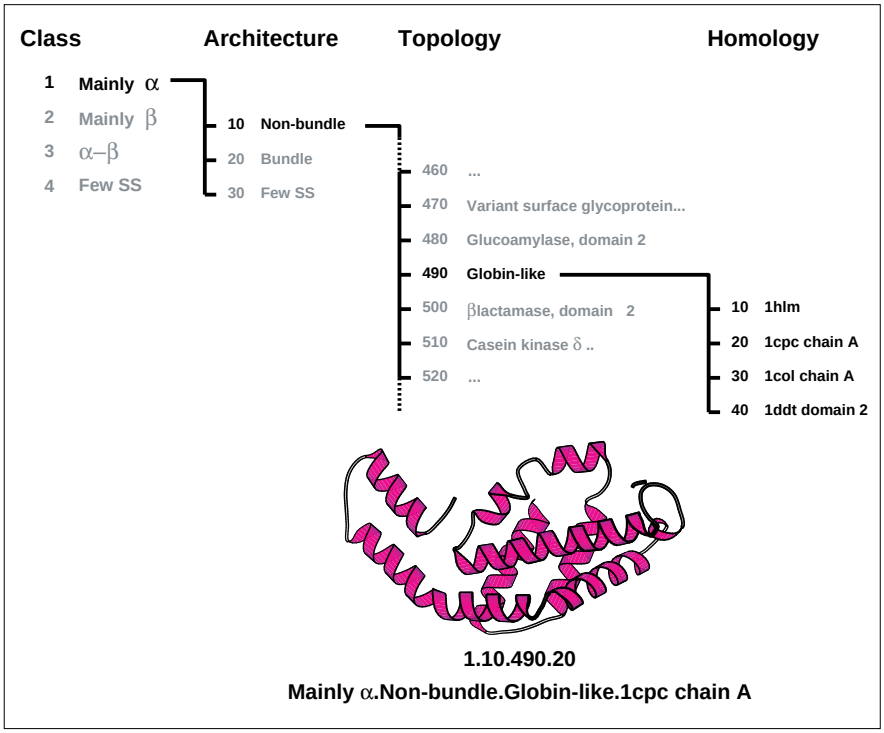
Although 5993 protein chains (8078 domains) were selected for CATH from the September 1996 release of the PDB, sequence comparisons showed that approximately three quarters of these were nearly identical (see Figure 1, Tables 1 and 2). CATH grouped the 8078 domain structures into 1821 non-identical families (N-level). By assuming that proteins with more than 35% sequence similarity adopt the same fold and share a common evolutionary ancestor, this number can be reduced to 1068 sequence families (S-level). More distant relatives were added to these families by searching for significantly high structural similarity and related biological functions, thereby reducing the number of homologous superfamilies to 645 (H-level). If a lower degree of structural similarity is allowed, these further group to give a total of 505 fold families (T-level), within which similarity may be a result of divergent or convergent evolution. These fold families are further grouped within CATH, automatically according to class and manually according to their architecture giving a total of 3 major classes and 28 different architectures (see Table 1). Figure 1 shows the annual increase in structures and structural families for each level in the CATH hierarchy.

Overview of architectures

The CATH architecture level is a subjective grouping of folds having similar shape, regardless of differences in scale or numbers of secondary structures. For example, both the 5-stranded barwin-like β-barrel folds and the 17-stranded porin-like β-barrel folds are assigned to the same general β-barrel architecture. Although this is a

Figure 3

CATH numbering scheme for representative structures from the globin-like fold family in the mainly α class. Four of the seven levels within the CATH database are shown, associated with Class, Architecture, Topology, and Homology. Each level is associated with a unique number. The (A), (T) and (H) levels are numbered in bins of ten to allow expansion of the database.



somewhat broad category, the fold families within this barrel architecture all share a common structural feature comprising a single β sheet. Similarly, the three-layer α–β sandwich architecture also represents a large grouping of folds of varying sizes (containing β sheets having from 4 to 17 β-strands). Again, all the folds within this architecture can be simply and usefully described as having a central β sheet with layers of α helices on each of its sides (see Figure 2).

Wherever possible, we have used architectural descriptions commonly cited in the literature. For a majority of the folds (>80%) this was a simple and straightforward process and the architectural categories assigned agreed well with those given in other publicly available databases (e.g. SCOP [17]). For more complex shapes, no architecture was assigned, and these folds were all placed in a single 'complex' bin until alternative assignment methods are developed. Such methods will probably describe shape according to the diverse motifs contained within the fold and the ways in which these motifs are combined in 3D space.

The variety of architectures that can easily be assigned by visual inspection is shown, for each class, in Figure 4. For mainly α proteins only the familiar four-helix bundle architecture is easily distinguishable. Other helix arrangements

Table 1

The numbers of families identified at different levels in the CATH hierarchy is shown for the mainly α, mainly β and α–β classes.

	A		T		H		S		N		I		Domains	
Class	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
Mainly α	3	9.7	145	28.7	157	24.3	232	21.7	380	20.9	837	26.4	1793	22.2
Mainly β	17	54.8	102	20.2	137	21.2	266	24.9	585	32.1	891	28.1	2625	32.5
α–β	10	32.3	244	48.3	337	52.2	556	52.1	829	45.5	1411	44.5	3562	44.1
Few SS*	1	3.2	14	2.8	14	2.2	14	1.3	27	1.5	32	1.0	98	1.2
Total	31	100.0	505	100.0	645	100.0	1068	100.0	1821	100.0	3171	100.0	8078	100.0

*The number of families for proteins having few secondary structure (SS) elements is also shown at each level in the hierarchy.

Table 2

The numbers of fold families (T-level), homologous superfamilies (H-level) and domain structures in different architectures are shown for the mainly α, mainly β and α–β classes.

Class	Architecture	Number of T-levels	Percentage of all T-levels*	Number of H-levels	Percentage of all H-levels*	Number of domains	Percentage of all domains*
Mainly α	Non-bundle	86	17.03	93	14.42	1455	18.01
	Bundle	34	6.73	39	6.05	226	2.80
	Few SS	25	4.95	25	3.88	112	1.39
Mainly β	Ribbon	17	3.37	17	2.64	114	1.41
	Single sheet	5	0.99	6	0.93	56	0.69
	Roll	6	1.19	6	0.93	55	0.68
	Barrel	22	4.36	29	4.50	861	10.66
	Clam	1	0.20	1	0.16	1	0.01
	Sandwich	21	4.16	43	6.67	1236	15.30
	Distorted sandwich	14	2.77	14	2.17	83	1.03
	Trefoil	1	0.20	4	0.62	49	0.61
	Orthogonal prism	1	0.20	1	0.16	4	0.05
	Aligned prism	1	0.20	2	0.31	3	0.04
	Four-propellor	1	0.20	1	0.16	3	0.04
	Six-propellor	1	0.20	1	0.16	37	0.46
	Seven-propellor	2	0.40	2	0.31	11	0.14
	Eight-propellor	1	0.20	1	0.16	2	0.02
	Two-solenoid	2	0.40	3	0.47	5	0.06
	Three-solenoid	1	0.20	1	0.16	1	0.01
	Complex	5	0.99	5	0.78	104	1.29
α–β	Roll	24	4.75	33	5.12	469	5.81
	Barrel	8	1.58	20	3.10	365	4.52
	Two-layer sandwich	77	15.25	112	17.36	957	11.85
	Three-layer (αβα) sandwich	78	15.45	115	17.83	1396	17.28
	Three-layer (ββα) sandwich	3	0.59	3	0.47	11	0.14
	Four-layer sandwich	4	0.79	4	0.62	12	0.15
	Box	1	0.20	1	0.16	2	0.02
	Horseshoe	1	0.20	1	0.16	1	0.01
	Complex	34	6.73	34	5.27	253	3.13
	Few SS	14	2.77	14	2.17	96	1.19
Few SS	Irregular	14	2.77	14	2.17	98	1.21

*The percentages of total fold families, total homologous superfamilies and total domain structures adopting a particular architecture are shown.

appear less distinct and may reflect the tolerance of helix packing modes that allows diverse combinations of two- and three-helix motifs. This gives rise to a continuum of folds within which helix packing angles range from aligned through to orthogonal. Despite this variety, certain motifs appear to recur frequently — the aligned α hairpin and the two-helix and three-helix orthogonal motifs common in the repressor and globin-like folds. Therefore, in this class, it may ultimately be more appropriate to separate fold families into architectural groups that reflect specific combinations of these common motifs.

By contrast to the mainly α class, in the mainly β class, the constraints on β strands to be hydrogen bonded within sheets and also on sheet–sheet packing gives rise to some very distinct and easily recognisable architectures. In particular, the β prisms, β propellors and β solenoids demonstrate the symmetry and regularity of structures satisfying these preferred packing constraints. In contrast to the few architectures observed within the mainly α class, at least 16 different, relatively simple, architectures can be discerned in the mainly β class.

The diversity of the mainly β class is not currently observed within the α–β class, in which only eight regular architectures are apparent to date. This may simply reflect a bias in the structures determined or could suggest that in this class the preferred motifs are more constrained in the ways in which they combine. The βαβ motif appears to be highly favoured and is observed within a large proportion of folds. In some topologies, the β strands are adjacent in space (classic motif) but in others they are separated by a third antiparallel strand, forming a three-stranded β sheet (split motif) [31]. Although both the classic and the split βαβ motifs are most commonly found in two and three-layer architectures, the classic motif is also found to recur within barrel and semi-barrel or horseshoe architectures (Figure 4).

The structures that fall outside these rather simple layer architectures tend to be quite complex. Compared to the

mainly β class, which has only three folds too complex to be assigned architectures, there are 12 complex folds within the α–β class. The irregularity of these structures appears to be a consequence of the heterogeneity of the motifs found within them which prevent regular packing of secondary structures. The α–β proteins often contain a mixture of motifs borrowed from the mainly β (e.g. antiparallel β meanders) and mainly α (e.g α hairpin) classes, together with βαβ motifs and α–β meanders. Disparity in the sizes and packing requirements of these motifs gives rise to a plethora of different combinations and 3D shapes, which cannot be easily described.

Population of fold space

Using CATH, we have reexamined the the distribution of known protein structures into different fold families (i.e. the distribution of structures in 'fold space'). Our previous classification of the June 1994 release of the PDB, using similar criteria for recognising structural families, revealed that single domain structures particularly favoured some nine T-level families or folds; these nine families comprised approximately 43% of the non-homologous structures. We described these families as 'superfolds' to indicate their unusually high popularity and the existence within them of many protein pairs with no significant sequence similarity or functional similarity. At least three different functions could usually be discerned for a given superfold family.

Analysis of the distribution of structures using the September 1996 version of CATH shows similar trends, revealing that the OB folds also demonstrate superfold-like qualities (i.e. high population of non-homologous structures, multiple functions exhibited across the family). Additionally, some eleven other T-level (fold) families contain at least two homologous superfamilies (having different functions) and have low sequence similarity (<20%) between many members, indicating that they too could have superfold like qualities (see Figure 5).

At the H-level, further possible evolutionary relationships between sequence families can be identified by cross-checking the literature and by reference to the SCOP database [17], which contains evolutionary data extracted from a variety of sources and derived by expert consideration. After merging possible relatives into the same superfamily (H-level), we observe that a lower 31% of non-homologous structures now belong to any one of the superfold families. The decrease in this value from our previous analysis (43%), as well as being due to a broader consideration of evolutionary relatedness, is due to the much larger dataset examined (~8000 structures compared to ~1000 in 1994). The proportion of structures in superfolds will also have decreased because our current analysis includes many domains from multidomain proteins — only a small proportion of these are found to match single domain proteins or other multidomain proteins (see below).

The proportion — nearly one third — of non-homologous proteins adopting one of the superfolds is still significant. Nevertheless, the underlying reasons for the popularity of these folds is still unclear, though they share similar characteristics. In all the families, the architectures ensure extensive contacts between adjacent secondary structure elements, guaranteeing a well-packed hydrophobic core and possibly facilitating folding by directing the protein along a pathway of locally folded intermediates. All contain one or more of the common or preferred motifs for their class and these motifs have often been repeated and combined to give symmetric and regular arrangements of secondary structures in two or three layers. Three of the folds are barrels or barrel-like, four are simple two or three layer sandwiches, the layers of which are always composed of the same type of secondary structure. The alpha superfolds (globin-like and four-helix bundle) can also be viewed as two-layer architectures in which the two pairs of α hairpins form separate layers and pack against each other in either an aligned or orthogonal fashion [32]. This point is further illustrated by Figure 6, which is a 'CATHerine wheel' plot showing the population of the different fold families and architectures within CATH. The outer radius of the wheel corresponds to the total number of superfamilies (H-level) in CATH. Each class is coloured separately and the number of superfamilies in each architecture is revealed by the size of the segment in the inner circle. The superfold families are illustrated as paler segments with the MOLSCRIPT [33] representations drawn adjacent to the segment. It can be seen that the majority of the superfolds occur within architectures which are also highly populated and could perhaps be referred to as 'superarchitectures' or major architectures. Superfold does not necessarily imply superarchitecture, however. For instance, the β-trefoil fold is an example of a superfold as there are more than three different functions exhibited within this family. The trefoil architecture, however, is not adopted by any other known fold families (see Figure 7).

Nearly two-thirds (65%) of the currently known non-homologous structures adopt one of these simple and repetitive layer-based superarchitectures (see Table 1). These major architectures include the α bundles, α two-layer-orthogonal, β rolls, β barrels, β sandwiches, αβ rolls, αβ barrels, αβ two-layer sandwiches, αβ three-layer sandwiches. In many architectures, the structures can easily be extended by one or more motifs (e.g. additional β hairpins or βαβ motifs) without changing the overall shape of the fold. This ease of extensibility may help to stabilise the fold during evolution. Structures possessing irregular or complex contacts between motifs might be less tolerant to mutational changes as these would induce

Figure 4

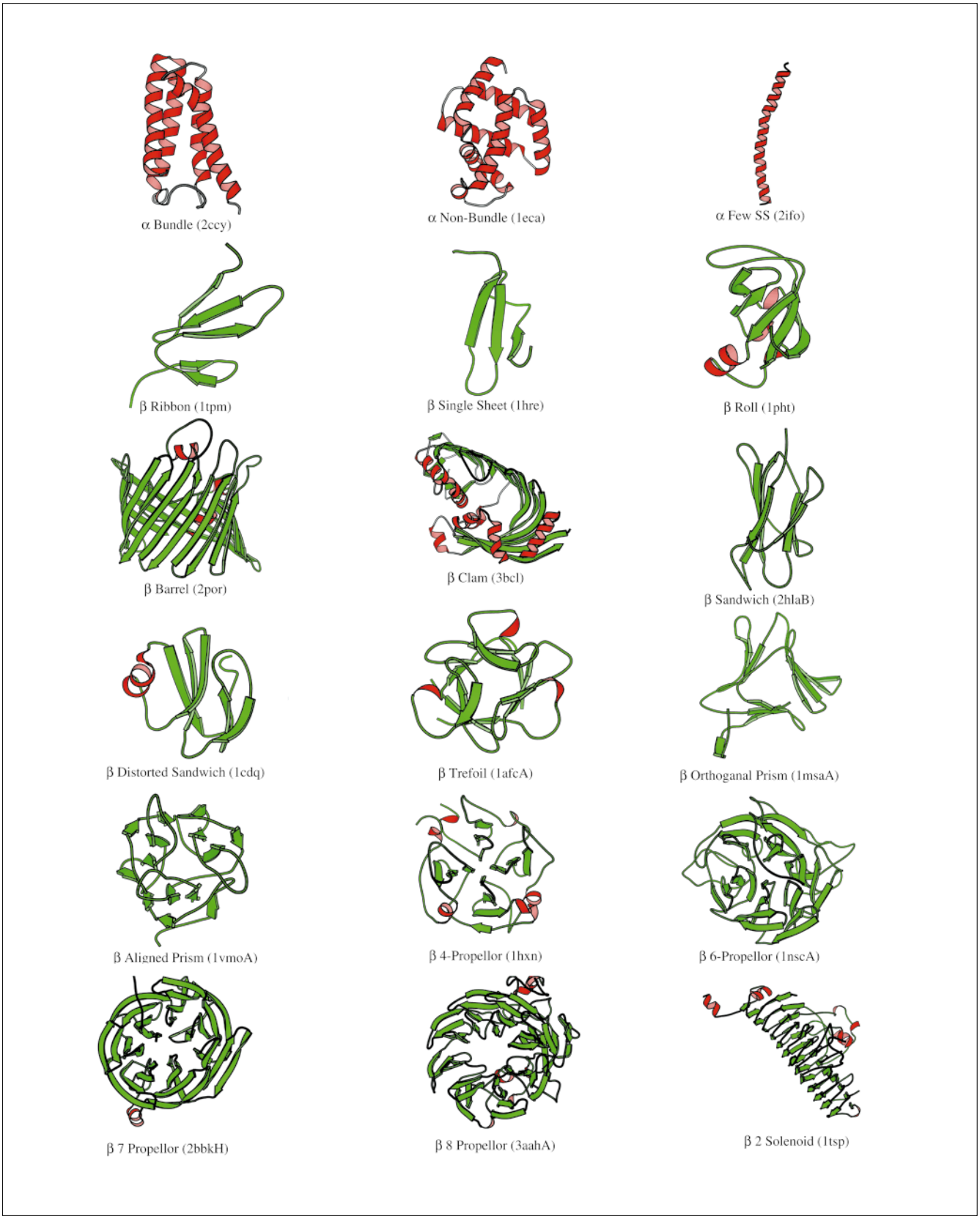
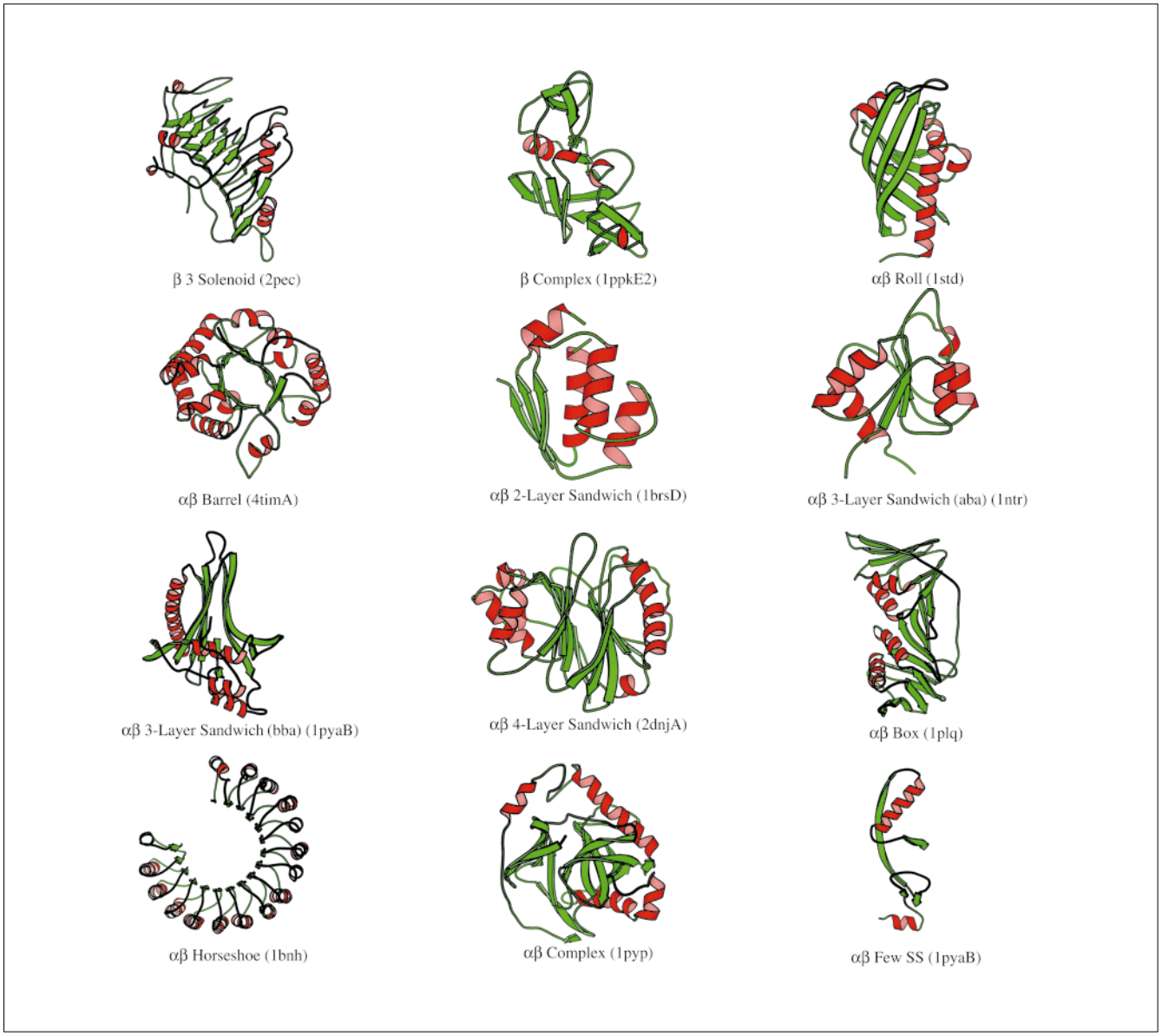

α Bundle (2ccy)
α Non-Bundle (1eca)
α Few SS (2ifo)
β Ribbon (1tpm)
β Single Sheet (1hre)
β Roll (1pht)
β Barrel (2por)
β Clam (3bcl)
β Sandwich (2hlaB)
β Distorted Sandwich (1cdq)
β Trefoil (1afcA)
β Orthoganal Prism (1msaA)
β Aligned Prism (1vmoA)
β 4-Propellor (1hxn)
β 6-Propellor (1nscA)
β 7 Propellor (2bbkH)
β 8 Propellor (3aahA)
β 2 Solenoid (1tsp)

Figure 4 continued



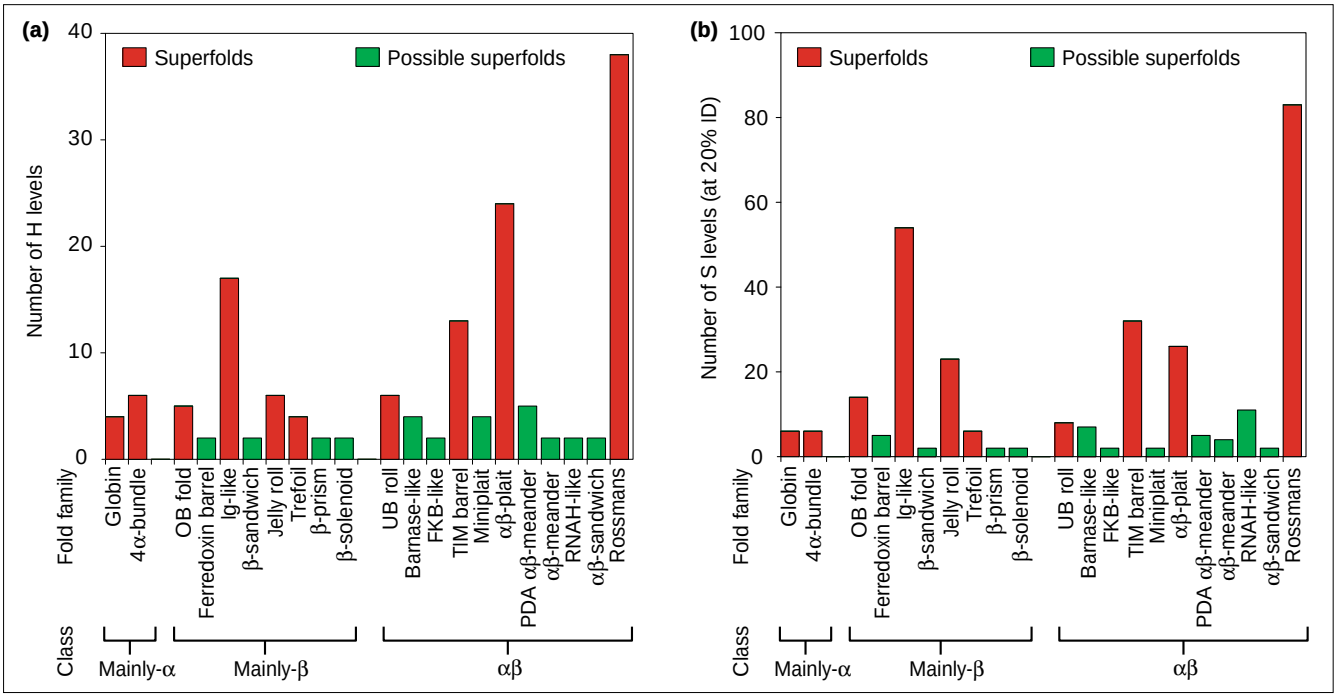
MOLSCRIPT [33] representations of the architectures identified for the mainly α, mainly β and α–β classes (PDB codes are given in parentheses).

more profound disturbances to the architecture. This is because, in addition to the ability to expand by adding on similar motifs (which may arise from gene duplication), layer-based architectures can further accommodate evolutionary changes by allowing the layers to slide relative to each other. This mechanism would not be available to complex architectures possessing a more diffuse hydrophobic core.

Overlap between fold families: the Russian doll effect

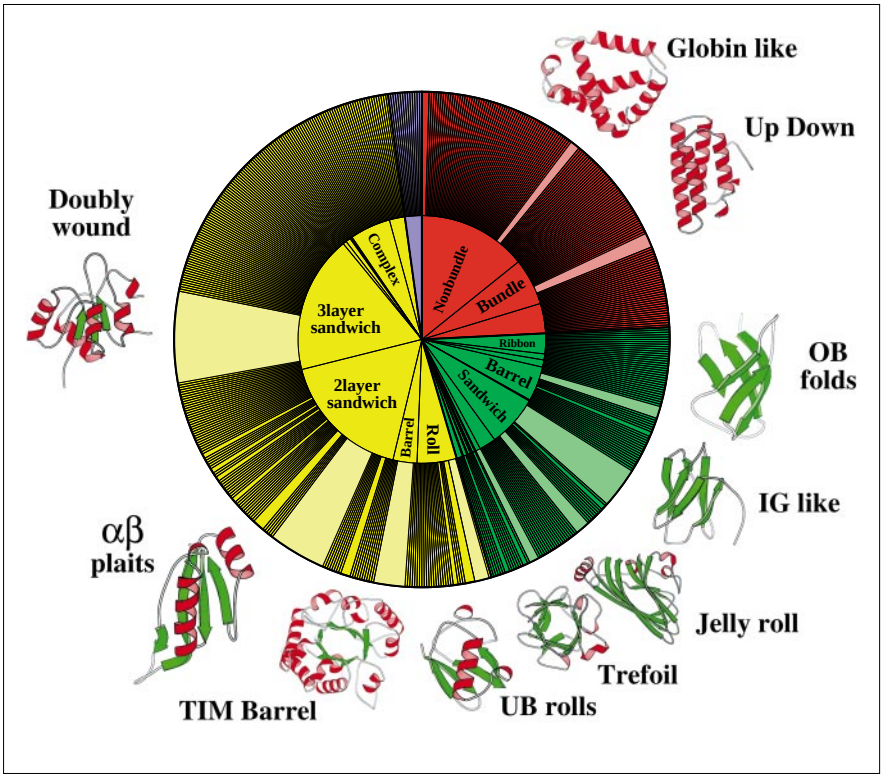
The recurrence of common motifs within many of the superfolds and major architectures gives rise to an overlap of structures in these regions of fold space. This means that it becomes harder to distinguish between structural families for these architectures and it is perhaps more appropriate to consider a continuum of protein folds. This is particularly apparent in the layer-based sandwich architectures of the mainly β and α–β classes. For example, within the α–β three-layer doubly wound architectures, it is possible to generate a very large family of structures using the simple criteria of a good structural comparison (SSAP score $>=70$) and reasonable overlap ($>=60\%$). Each new structure added to a family will be related to the last by a simple extension of one or more βαβ motifs and

Figure 5



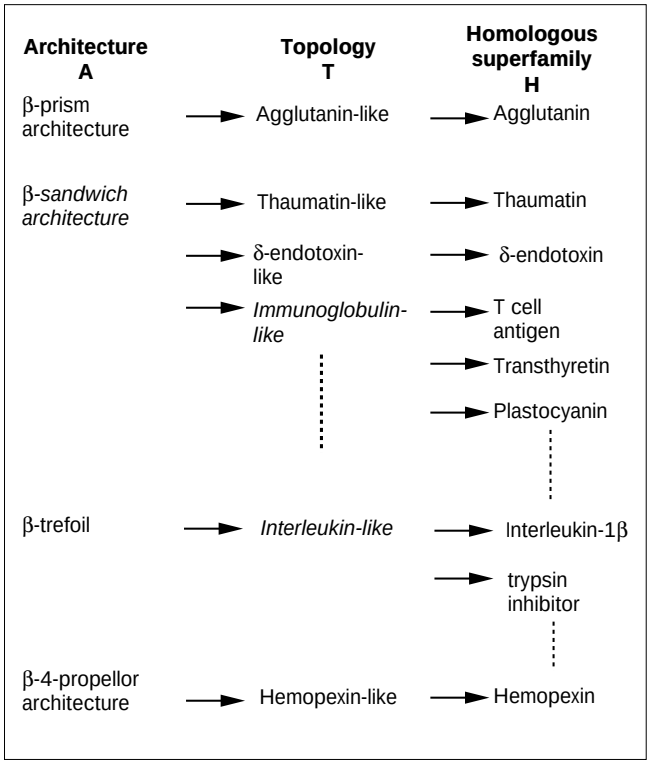
(a) Histogram showing the numbers of homologous superfamilies in each of the nine previously identified superfold families [1], the probable OB superfold and the 11 possible superfolds identified in the September 1996 release (version 2.0) of CATH. (b) The numbers of sequence families (proteins clustered with ≥ 20% sequence identity) in each of the same fold families.

Figure 6



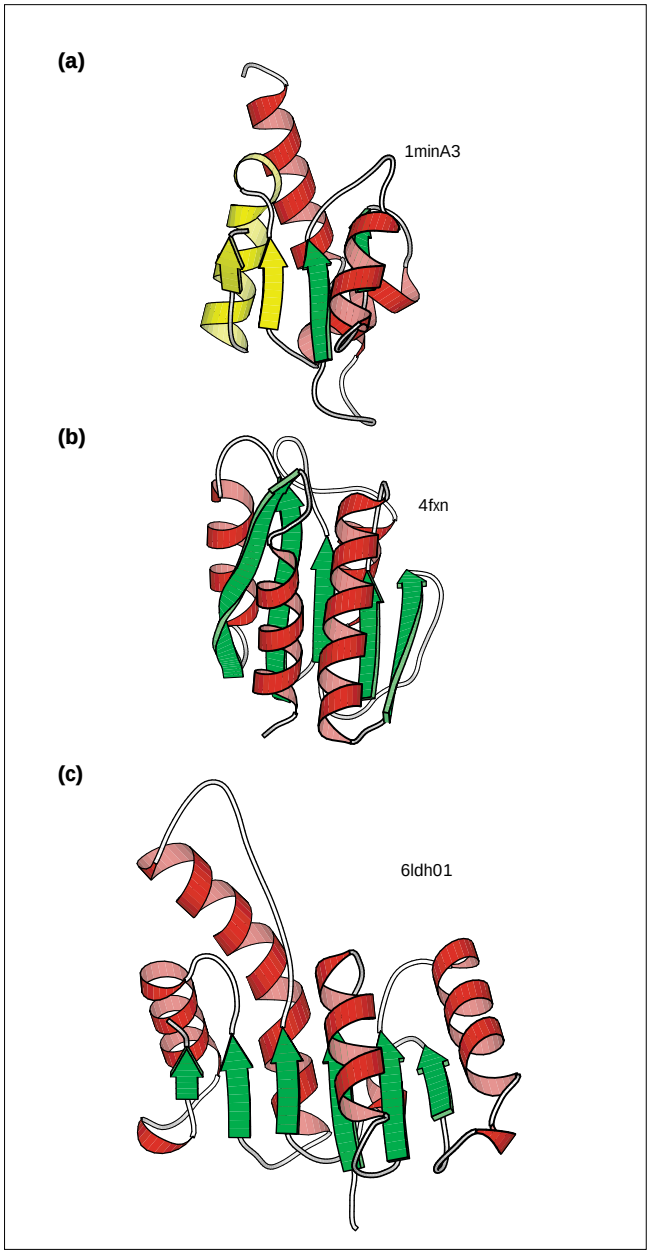
'CATHerine wheel' plot showing the distribution of non-homologous structures (i.e. a single representative from each homologous superfamily at the H-level in CATH) amongst the different classes (C), architectures (A) and fold families (T) in the CATH database. Protein classes are shown coloured as red (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). The superfold families are indicated in paler colours and illustrated with a MOLSCRIPT drawing of a representative from the family.

Figure 7



Schematic illustration of the population of architecture, topology and homologous superfamily levels within the CATH database, showing the existence of 'superarchitectures' (containing many different fold families) and 'superfolds' (containing many different homologous superfamilies). Superarchitectures and superfold families are shown in italics.

Figure 8



The 'Russian doll' effect for the flavodoxin fold family. (a–c) MOLSCRIPT diagrams for representatives from different homologous superfamilies (H-level) of the flavodoxin fold family (CAT number 3.40.50) in the α–β class. All members of the family contain recurring βαβ motifs, coloured yellow in (a), and the progression from four-stranded β sheet in 1minA3 through to five β strands in 4fxn to six β strands in 6ldh01, by addition of βαβ motifs, illustrates the Russian doll effect for this fold family. The smaller protein (1minA3, 110 residues) has two thirds the number of residues as the largest shown (6ldh01, 162 residues).

the structures are then embedded within each other in a 'Russian doll' like effect (see Figure 8). Similarly, the mainly β class (Figure 9) shows the overlap of β hairpin and greek-key motifs between the immunoglobulin-like telokin fold and the jelly-roll fold of the tobacco necrosis virus protein, which also results in significant structure comparison scores. According to the simple criteria used for generating fold families (T-level) within CATH, the immunoglobulin-like folds and the jelly-roll folds should merge into a single T-level family. Structural matches that cause a merging of different fold families (T) are checked manually before updating CATH and the families that are more commonly described in the literature as separate folds (as with the jelly-roll and immunoglobulin folds) are not clustered, but the relationship between them is expressed in the pairwise SSAP score matrix for the β-sandwich architecture.

Because the superarchitectures are so highly populated, the Russian doll effect can give rise to some extremely large and diverse structural families and it is pertinent to consider whether such a grouping would be valuable. One of the main motivations for identifying families in the CATH database is to allow better analysis of sequence–structure relationships thereby improving structure

Figure 9

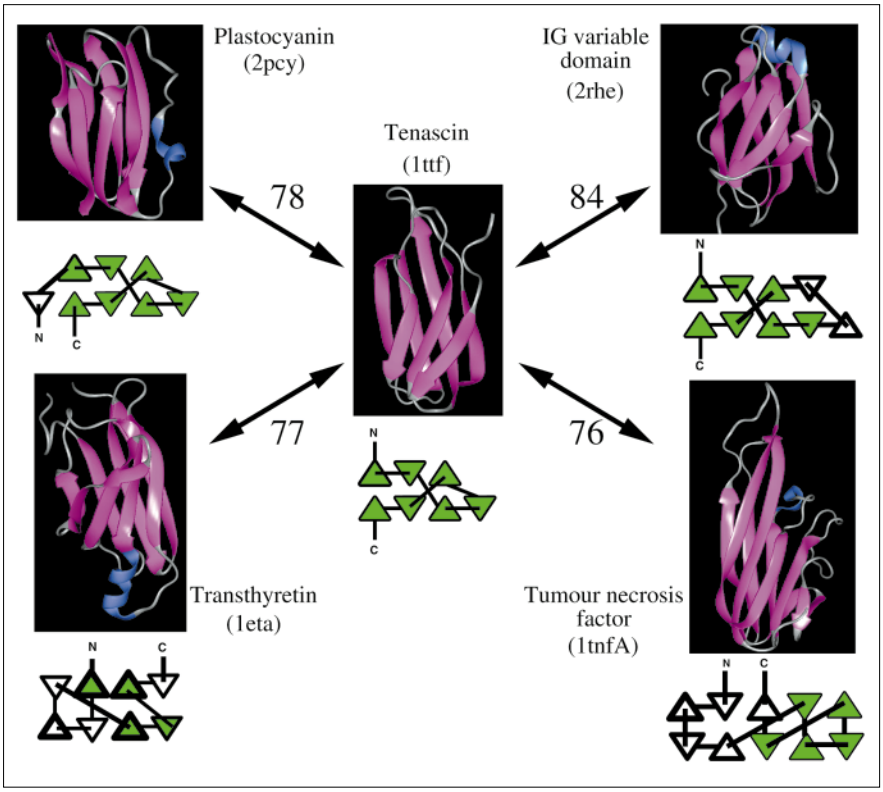


Illustration of 'motif' overlaps in the mainly β sandwich architecture. Each structure shown can be related to the central tenascin structure by a motif containing at least four β strands (although these are not sequential in the transthyretin structure) up to seven β strands in plastocyanin and the immunoglobulin variable domain structures. It can be seen that this results in the possible merging of the immunoglobulin fold family (2rhe) and the jelly-roll fold family (1tnfA) through overlap of a large motif containing five β strands. This is not currently done in CATH, as both families are commonly referred to as separate folds in the literature.

prediction (from sequence) methods. It would therefore seem more useful to subclassify these regions of fold space according to more sophisticated criteria for structural similarity, in order to generate smaller families containing closer relatives. These criteria would be based on recognising similarities within the cores of the protein structures belonging to a particular family and setting tolerances on the number of allowed secondary structure indels. With this aim, a suite of programs (CORA, COnsensus Residues Attributes) for analysing structural families has been developed (CAO, unpublished data) and will be applied to all the superfold families within the CATH database. It is planned that future releases of CATH will assign structures to fold families within the superarchitectures according to the diagnostics generated by CORA.

Identification of multidomains and recurrence of folds within multidomain proteins

By applying the consensus domain boundary assignment procedure to all N-representatives (787 proteins) in the September 1996 version of CATH, 74% of the total number of single domains (515) could be assigned automatically and 21% of the total number of multidomain proteins (272). Of those remaining unassigned, at least one of the methods gave acceptable boundary assignments, needing no or only minor adjustments.

Interestingly, an analysis of the distribution of domain structures in non-homologous multidomain proteins showed that only 8.1% of these domains occur also as single domain proteins and only 5% recur in other multidomain proteins.

Future developments: automatic architecture assignment

The CATH architectural groupings are currently broad, general, categories that represent a preliminary classification which should significantly aid a future, more detailed analysis of common architectural features. Although, these groups are assigned manually, other publicly available classifications have adopted a similar pragmatic approach, using a combination of automatic and manual approaches where appropriate (SCOP, DIAL [17,27]).

Until we improve our understanding of structural constraints on secondary structure packing, the ideal of a completely automatic approach generating self-consistent and reproducible hierarchies at all levels is not feasible. Some approaches avoid this problem by clustering proteins on the basis of overlapping helices, strands or small common motifs, regardless of 3D arrangement. Such motif-based classifications, however, are generally less useful for revealing global structural relationships between evolutionary related proteins. We chose, therefore, to use an

initial manual approach based on visual recognition of protein architecture, akin to the early strategies for biological classification of organisms. By performing this 'preliminary' architecture classification, we hope to improve our understanding of protein structural constraints with the ultimate aim of developing a more automated approach. Any features identified can be subsequently encoded in automatic algorithms and tested for their suitability in generating consistent and reproducible classification schema. This will inevitably be an iterative approach whereby initial groupings suggest improved algorithms which subsequently lead to more appropriate clusters. However, the current observation that the majority of protein folds adopt very simple layer-based shapes, which could be expected to be amenable to automatic recognition methods, offers considerable hope for achieving this aim and justifies this initial manual step in architecture classification.

Biological implications

Although there are more than 5000 known protein structures deposited in the Brookhaven Data Bank, classification of these proteins into structural families using the CATH database suggests that they adopt only ~500 different folds. Ten of these folds are very highly populated (superfolds) — they are seen in nearly one-third of all non-homologous structures currently known. Within these superfold families, sequences and functions can differ quite considerably. For the majority of folds outside these superfold families, however, we can be reasonably confident that proteins assigned to a given structural family will possess a similar function to other proteins within the family.

Analysis of the structural families generated by the CATH database also revealed that nearly two-thirds of non-homologous structures adopt one of nine simple architectures. Comparison of the fold families suggests that for some architectures there is a continuum of structures traced through a spectrum of favoured motifs, occurring in many different combinations, with some regions in this continuum being much more highly populated.

It also suggests that the most favoured protein folds are mostly composed of particularly symmetrical arrangements of common motifs. Although these motifs may represent favoured folding pathways or preferred nucleation sites for folding, the assembly of such recurring motifs into regular symmetric architectures might also be associated with energy minima for the complete tertiary structures, achieved by optimising contacts between neighbouring secondary structures.

The organisation of proteins by global structural similarity should not only improve prediction algorithms based on fold recognition but will also allow the distribution of common motifs to be explored more easily, giving insights into which combinations of motifs generate stable protein architectures. Importantly, such a database of well-characterised fold families allows newly determined structures to be easily examined for recognisable folds. This is desirable as any similarity to a known structure may have important functional and evolutionary implications.

Clearly, it is most important to relate structure to function and it would be very useful to have a functional classification system for proteins. For enzymes, the E.C. numbers provide a useful starting point, but there is no systematic equivalent for other functional types. Correlation between fold type and other factors, such as cellular location, are also interesting. Developing such a classification scheme will facilitate the analysis of the human genome and the recognition of distant relationships between proteins.

Materials and methods

Overview of procedures used for identifying structural families

A flowchart for constructing the CATH database is shown in Figure 10. Proteins are initially grouped according to sequence similarity, after which domain boundaries are assigned for any multidomain families and the proteins belonging to them chopped into their separate domains. The sequences of the chopped domains are subsequently recompared and reclustered. Class is then assigned, before comparing representative structures from each sequence-based family, using a structure alignment program [10,34]. This prevents unnecessary cross-class comparisons between the mainly α and mainly β proteins, although the mainly β versus α–β and mainly α versus α–β comparisons are performed. Structures are then automatically merged into homologous superfamilies (H-level) and fold families (T-level) on the basis of structural similarity. Finally, the architecture of each fold is determined manually. A description of the methods and validation procedures used at each stage is given below.

Step 1: selection of structures for CATH database

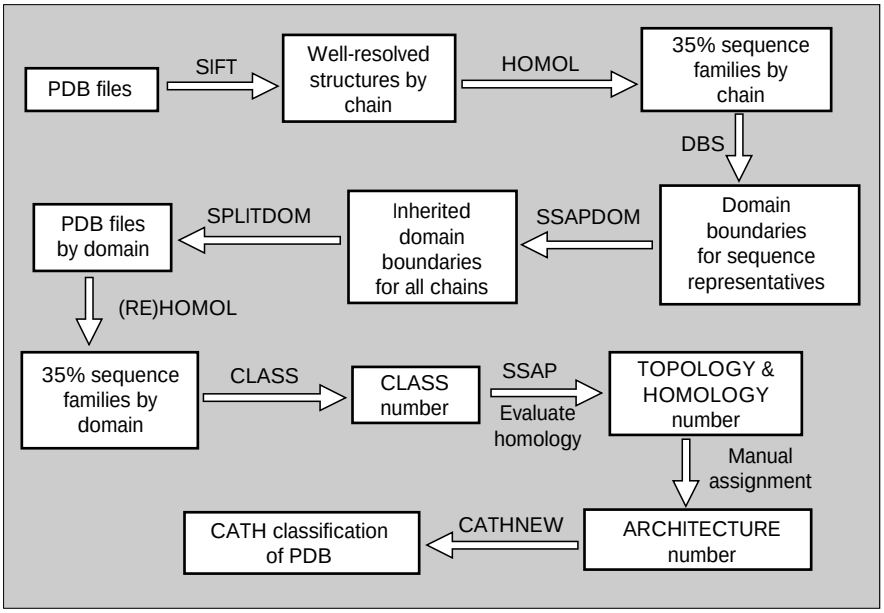
CATH contains only well-resolved crystal structures (3.0 Å resolution or better) and NMR structures from the Brookhaven Protein Data Bank (PDB) [11,12]. Structures are automatically labelled according to the method of determination and whether or not they are native structures: 1, native (X-ray); 2, mutant (X-ray); 3, native (NMR); 4 mutant (NMR); 5, Cα only; 6, model; 7, design protein; and 8, non-protein. Proteins labelled 5–8 are removed from the list of allowed PDB structures, which is then sorted according to the label and resolution of the structure (proteins solved by NMR and therefore having labels greater than 2 are assigned an artificial resolution of 999.99). The data considered here include all the structures in the PDB up to September 1st, 1996. Out of 8260 entries 5993 are selected for classifying within CATH.

Step 2: sequence comparisons (S-level)

Much of the information in the databank is highly redundant as more than three-quarters of the structures have nearly identical sequences (see Figure 1). Therefore, most classification procedures initially perform sequence comparisons between all the proteins [6,20,35–37].

Pairwise comparisons between the sequences of all the proteins selected for CATH are performed using a standard Needleman and Wunsch algorithm [9], scoring 1 for matching identical residues, 0 otherwise and charging a gap penalty of 4. Scores are normalised on the length of the smallest protein to give a value in the range of 0 to 100 for identical proteins. To check the significance of any pair score, the associated sequences are jumbled 100 times and if any comparisons

Figure 10



Flow chart of the procedures and programs used to update the CATH database. The names given above the arrows are those of the programs used in the database. The boxes describe the data generated by each program.

of jumbled sequences score more highly the original score is reset to 0. Sequence identities are calculated as the number of identical residues as a percentage of the smaller protein.

Subsequently, single linkage cluster analysis, on the matrix of pairwise sequence scores, groups proteins into sequence-based families. Initially all completely identical proteins are grouped together into identical (I) families (100% sequence similarity and 100% overlap of structures). Subsequently near-identical (N) families are created (>95% sequence similarity, at least 85% of larger protein equivalent to smaller). The N and I levels are implicit groupings in CATH. The S-level in CATH is then generated by clustering proteins having 35% or more sequence identity (at least 60% of larger protein equivalent to smaller). Although this is a rather stringent cutoff, it ensures that there are no false positives. Any distant similarities between proteins having less than 35% sequence identity are recovered at the next stage of structure comparison.

Proteins in each sequence-based family are sorted according to label and resolution and the first entry in the list, which is therefore generally the best resolved crystal structure, selected as a representative for the family. Because representative structures may change with new releases of CATH, a paradigm for the family is also chosen which is never changed and which is typically the first well-resolved structure to be determined or a commonly known example of the fold (e.g. erythrocruorin 1ECA for the globin fold).

Step 3: assignment of domain boundaries for multidomain proteins

One representative from each near-identical sequence family (N-level, >95% sequence identity) is analysed to determine the number of domains and corresponding domain boundaries. A consensus approach is used whereby the assignments given by three automatic methods are compared (DETECTIVE [22], PUU [23], DOMAK [24]). If they agree in the number of domains identified and there is at least 85% overlap in residues assigned to a given domain, the boundaries given by DETECTIVE are used to chop the structure into its constituent domains. Where the algorithms disagree, the boundaries are examined by visual inspection and by reference to assignments in other databases, SCOP [17], 3DEE, Siddiqui and Barton. (http://speed.biop.ox.ac.uk8080/3Dee), and the literature.

Once domain boundaries are established for each N-representative, they are inherited by every member of that family using a modified version of the structure comparison algorithm, which identifies the equivalent boundary positions in the aligned protein. A validation procedure checks that, once boundaries are assigned for each member within an S-level family, the number of residues unassigned to a domain is less than 30 residues. This ensures that any structures potentially containing more domains than the N-representative are flagged, so that they can be checked manually. This can sometimes occur in a family of large multidomain proteins (e.g. >400 residues), where a small domain (e.g. ~50 residues) would still allow an 85% overlap between structures, even if the small domain was absent in some members of the family.

The sequences of the resulting single domains folds are compared against each other and all other single domain structures in CATH, using the same procedure as for step 1. Any which match at the 100%, 95% and 35% sequence identity levels and required overlap, are added to the respective families. Information about domain boundaries for multidomain structures, is stored within CATH and can be accessed from the web page for each constituent domain (see Table 2). The multidomain sequence families (I, N and S level) are also stored for the purpose of identifying related multidomain proteins in subsequent releases of the databank.

Step 4: automatic assignment of class

The increasing number of structurally related proteins with insignificant sequence similarities means that at some stage direct structural comparison methods must be applied. As structure comparisons are compute-intensive, the structural class of the protein is assigned next (C-level). This speeds up the classification by preventing any cross-class comparisons. An automatic procedure [30] is applied to each S-level representative, which examines the composition, secondary structure contacts and proportion of parallel/antiparallel β sheet. Approximately 90% of structures can be confidently assigned to one of the

three major structural classes (i.e. mainly α, mainly β, and α–β) using this method. Those structures lying on the boundaries between different classes are assigned by visual inspection. Where the class is ambiguous, structures may be placed in more than one class.

Step 5: structure comparisons (H- and T-levels)

Structure comparisons are next performed between the N-level representatives within each class to merge related folds (T and H levels). A fast version of the program SSAP [34] is used. This first aligns secondary structure vectors between the proteins. For high scoring pairs of proteins, SSAP subsequently compares residue 'structural environments', returning a normalised similarity score (0–100). For protein pairs scoring below 75, the structures are re-aligned using a slower and more sensitive version of SSAP [10], which only compares residue structural environments and performs at least tenfold more of these than the fast version of SSAP.

SSAP scores for protein pairs are stored in a two-dimensional matrix and structure pairs that have a sufficiently high SSAP score (and a significant proportion of the larger fold equivalent to the smaller – at least 60%) are merged into structure-based families using single linkage clustering. Two cutoffs on the SSAP score are applied, 70 to generate the T-levels and 80 the H-levels of the CATH database (see above; [6]).

In addition to satisfying structural criteria (i.e. SSAP ≥80, overlap ≥60%), proteins are only assigned to a particular homologous superfamily if they possess a similar function to those exhibited by other proteins within the family. Functions are determined by reference to SWISSPROT [38] entries, where available, or using information from the PDB file or the literature. Where the evidence is unclear, the protein is assigned to a fold family (T-level) on the basis of the SSAP score but placed in a separate homologous superfamily until more information becomes available.

Step 6: assigning architecture

Finally, the architecture (A-level) or arrangement of secondary structures in the protein fold is determined manually. This is done using the classification of Richardson [39] and by reference to well-known groups reported in the literature (e.g. β propellors [40], and β trefoils [18]). Architecture describes the general shape of the fold (e.g. barrel, sandwich, roll) or, where this is less amenable to description, the general packing of the secondary structures (e.g. non-bundle and bundle in the mainly α class). Complex arrangements which cannot easily be described are placed in a general 'complex' architecture.

Within each architecture (A-level), the fold (T-level) families are sorted in order of increasing size, as determined by the average number of secondary structure elements. Averages are calculated by summing over the N-level representatives for the family.

Step 7: data on individual structures

For each structure in CATH, a number of graphical representations can be displayed, including topology diagrams (TOPS [41]) where available, hydrogen-bonding plots (HERA [42]) and MOLSCRIPT representations [33]) and where appropriate plots of ligand–protein interactions (LIGPLOTS [43]).

In addition, each entry has an associated summary report (http://www.biochem.ucl.ac.uk/bsm/pdbsum), generated from information given in the PDB file (Laskowski *et al*., personal communication) and also containing domain boundary data for multidomain proteins. Functional data from SWISS-PROT [38] is displayed where available.

Step 8: assigning CATH numbers

CATH numbers are automatically assigned at each level in the hierarchy (see Figure 3). Besides improving data management and updating, CATH numbers are useful for interpreting the results of fold prediction, fold recognition or search algorithms by quickly revealing agreement in class, architecture or overall fold family within a list of suggested structures.

Fold lexicon, fold gallery and glossary

In order to allow widespread access to the classification, CATH has been represented as a interlinked network of hypertext pages that can be viewed remotely from any suitably equipped computer system. CATH can be accessed via the URL – http://www.biochem.ucl.ac.uk/bsm/cath. The pages allow access to extra information and diagrams about given folds and their functional groups in the form of downloadable text or PostScript files. CATH is mirrored at the Brookhaven National Laboratory (Virginia, USA), the Weizmann Institute (Jerusalem, Israel) and the Helix Research Institute (Tokyo, Japan) to improve access.

A fold lexicon describes each architecture within CATH (i.e. unique CA numbers) and an associated fold gallery contains MOLSCRIPT representations for representatives from each architecture and unique fold family. Together the lexicon and gallery allow the user to browse through the universe of protein structures examining relationships at different levels in the CATH hierarchy. Additionally, we will provide a CATH server facility (ADM, unpublished data) for searching through the current structure database with a newly determined protein structure. The search procedure uses both sequence and SSAP comparisons to identify the fold family to which the new structure should be assigned.

For each fold family (T-level), pairwise sequence matrices and SSAP matrices are stored and accessible, together with a summary table which displays average data (e.g. number of residues and secondary structures) and DSSP [44] secondary structure strings for representatives from each S-level family. Multiple structure alignments and templates will be generated for each fold family using the program CORA (Orengo, unpublished data). Population statistics are also available for different levels in the hierarchy (see for example Table 1).

Acknowledgements

We are very grateful to Roman Laskowski, Andrew Wallace and Gail Hutchinson for providing derived data for individual protein structures in the PDBSUM files used in CATH. We would also like to thank Jane Richardson for many helpful discussions and suggestions. We are indebted to the crystallographers and NMR groups who have deposited structures in the Brookhaven Databank, thereby enabling this classification of protein domain folds. CAO acknowledges financial support from the Medical Research Council, DTJ from the Royal Society, SJ from BBSRC grant 31/MOL04573 and AM from the BBSRC.

References

1. Orengo, C.A., Jones, D.T., Taylor, W. & Thornton, J.M. (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631–634.
2. Chothia, C. (1993). One thousand families for the molecular biologist. *Nature* **357**, 543–544.
3. Chothia, C. & Lesk, A.M. (1986). The relation between divergence of sequence and structure in proteins. *EMBO J* **5**, 823–826.
4. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and structural meaning of sequence alignments. *Proteins* **9**, 56–68.
5. Flores, T.P, Orengo, C.A. & Thornton, J.M. (1993). Conformational characteristics in structurally similar protein pairs. *Protein Sci.* **7**, 31–37.
6. Orengo, C.A., Flores, T.P., Taylor, W.R. & Thornton, J.M. (1993). Identifying and classifying protein fold families *Protein Eng.* **6**, 485–500.
7. Orengo, C.A., Flores, T.P., Taylor, W.R. & Thornton, J.M. (1993). Recurring structural motifs in proteins with different functions. *Curr. Biol.* **3**, 131–139.
8. Holm, L., Ouzonis, C., Sander, C., Tuparev, G. & Vriend, G. (1993). A database of protein structure families with common fold motifs. *Protein Sci.* **1**, 1691–1698.
9. Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443.
10. Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.

11. 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.
12. 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* (Allen, F.H., Bergerhoff, G. & Sievers, R., eds), pp. 107–132.
13. Orengo, C. (1994). Classification of protein folds. *Curr. Opin. Struct. Biol.* 4, 429–440.
14. Holm. L. & Sander, C. (1996). Mapping the protein universe. *Science* 273, 595–602.
15. Holm, L. & Sander, C. (1994). Searching protein structure databases has come of age. *Proteins* 19, 165–173.
16. Brown, N.P., Orengo, C.A. & Taylor, W.R. (1996). A protein structure comparison methodology. *Comp. Chem.* 20, 359–380.
17. 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.
18. McLachlan, A.D. (1979). Three-fold structural pattern in the soyabean trypsin inhibitor (Kunitz). *J. Mol. Biol.* 133, 557–563.
19. Murzin, A.G. (1993). OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* 12, 861–867.
20. Holm, L, & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acid Res.* 22, 3600–3609.
21. Hogue, C.W.V., Ohkawa, H. & Bryant, S.H. (1996). WWW-Entrez and the molecular modelling database. *Trends. Biochem. Sci.* 21, 226–229.
22. Swindells, M.B. (1995). A procedure for detecting structural domains in proteins. *Protein Sci.* 4, 103–112.
23. Holm, L & Sander, C. (1993). Parser for protein folding units. *Proteins* 19, 256–268.
24. 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.
25. Islam, S.A., Luo, J. & Sternberg, M.J.E. (1995). Identification and analysis of domains in proteins. *Protein Eng.* 8, 513–525.
26. Sowdhamini, R. & Blundell, T.L. (1995). An automatic method involving cluster analysis of secondary structures for identification of domains in proteins. *Protein Sci.* 4, 506–520.
27. 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. Des.* 1, 209–220.
28. Rufino, S.D. & Blundell, T.L. (1994). Structure-based identification and clustering of protein families and superfamilies. *J. Comp. Aided Mol. Des.* 8, 5–27.
29. Levitt, M. & Chothia, C.(1976). Structural patterns in globular proteins. *Nature* 261, 552–558.
30. Michie, A.D., Orengo, C.A. & Thornton, J.M. (1996). Analysis of domain structural class using an automated class assignment protocol. *J. Mol. Biol.* 262, 168–185.
31. Orengo, C.A. & Thornton, J.M. (1994). Alpha plus beta folds revisited: some favoured motifs. *Structure* 1, 105–120.
32. Ptitsyn, O.B. (1974). Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding. *J. Mol. Biol.* 88, 287–300.
33. Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24, 946–950.
34. Orengo, C.A., Brown, N.P. & Taylor, W.R. (1992). Fast structure alignment for protein databank searching. *Proteins* 14, 139–167.
35. Boberg, J. Salakoski, T. & Vihinen, M. (1992). Selection of representative set of structures from Brookhaven Data Bank. *Proteins* 14, 265–276.
36. Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng.* 5, 121–137.
37. Overington, J.P., *et al*., & Blundell, T.L. (1993). Molecular definition in protein families: a database of three-dimensional structures of related structures. *Biochem. Soc. Transac.* 21, 597–604.
38. Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucleic Acid Res.* 20, 2019–2022.
39. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* 34, 167–339.
40. Murzin, A.G. (1992). Structural principles of the propellor assembly of beta sheets: the preference for seven-folded symmetry. *Proteins* 14, 191–201.
41. Flores, T.P., Moss, D.S. & Thornton, J.M. (1993). An algorithm for automatically generating protein topologies. *Protein Eng.* 7, 31–37.
42. Hutchinson, G.H. & Thornton, J.M. (1990). HERA: a program to draw schematic diagrams of protein secondary structure. *Proteins* 8, 203–212.
43. Wallace, A.C. Laskowski, R.A. & Thornton, J.M. (1995). LIGPLOT: a program to generate schematic diagrams of protein ligand interactions. *Protein Eng.* 8, 127–134.
44. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure. *Biopolymers* 22, 2577–2637.