Acta Cryst. (1998). D**54**, 1155–1167

Classifying a Protein in the CATH Database of Domain Structures

C. A. Orengo,[a]* A. M. Martin,[a] G. Hutchinson,[b] S. Jones,[a] D. T. Jones,[b] A. D. Michie,[a] M. B. Swindells[b] and J. M. Thornton[a]

[a]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, Gower St, London WC1E 6BT, England, and* [b]*Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England. E-mail: orengo@biochemistry.ucl.ac.uk*

(*Received 13 March 1998; accepted 27 May 1998*)

Abstract

The CATH database of protein domain structures classifies structures according to their (C)lass, (A)rchitecture, (T)opology or fold and (H)omologous family (http://www.biochem.ucl.ac.uk/bsm/cath). Although the protocol used is mostly automatic, manual inspection is used to check assignments at some critical stages, such as the detection of very distantly related homologues and anologues and the assignment of novel architectures. Described in this article is a recently established facility to search the database with the coordinates of a newly determined structure. The CATH server first locates domain boundaries and then uses automatic sequence and structure comparison methods to assign this new structure to one or more of the domain families within CATH. Diagnostic reports are generated, together with multiple structural alignments for close relatives. The Server can be accessed over the World Wide Web (WWW) and mirror sites are planned to improve access.

1. Introduction

Several protein structure classifications have been established over the last five years. Other papers in this issue describe the methodology applied by Murzin *et al.* (*SCOP*, 1995), Sowdhamini *et al.* (*Ddbase*, 1996) and Barton *et al.* (3*Dee*, 1997) when clustering proteins into structural families. These vary in their degree of automation though most groups would acknowledge the need for some manual checking and indeed fold families in the SCOP database are largely contructed by visual inspection. However, with the huge increases in the Brookhaven Protein Data Bank (PDB, Abola *et al.*, 1987) expected (there are now over 150 structures deposited monthly and some estimates suggest that there will be as many as 20 000 structures in total by the millenium) it will become essential to automate the classification process as far as possible.

If estimates of the total number of folds are correct (1000, Chothia, 1993; a few thousand, Orengo *et al.*, 1994), there may soon be representatives of most of the folds in nature. Hopefully, as these fold families are further populated, we shall learn more about the structural constraints in each, and it will become easier to design robust criteria for automatically assigning relatives. In CATH, we already use an automatic structure comparison method for identifying homologues or analogues, with cutoffs established by empirical trials (*SSAP*, Taylor & Orengo, 1989; Orengo *et al.*, 1992).

However, there still exits a structural 'twighlight' zone where caution is needed and reasonable scores can be returned by unrelated structures, which have common motifs assembled differently in three-dimensional space. To some extent this reflects the fact that for some architectural arrangements there is a continuum of protein structures causing overlap between fold families, depending on the criteria used for clustering structures. Recognizing this problem, some groups have chosen not to cluster proteins formally into structural families and their approaches reflect this structural continuum by allowing proteins to be grouped according to different degrees of structural similarity (ranging from global through to local similarities based on common motifs). For example, both the *DALI* database (Holm *et al.*, 1993) and the *ENTREZ* database of Hogue *et al.* (1996) apply automatic structure comparison methods (*DALI*, Holm & Sander, 1993*a*; *VAST*, Hogue *et al.*, 1996, respectively) and generate lists of structural neighbours, ordered according to their similarity score. This is expressed as a Z score which takes into account how unusual the observed similarity is, across a data set of non-homologous structures. As well as providing ordered lists of structural relatives, the *DALI* server also generates multiple alignments of these relatives against the probe structure.

However, in the *CATH* database of protein domain structures we have chosen to group proteins into families according global similarities in the structures (Orengo *et al.*, 1997). The four major hierarchical levels in *CATH* are class, architecture, topology (or fold) and homologous family. There are currently only three major classes recognized in CATH (mainly α, mainly β and α–β) since our analysis of class, based on residue composition and secondary-structure packing (Michie *et al.*, 1996), found no clear distinction between the alternating α/β and $\alpha+\beta$ classes originally described by Levitt & Chothia (1976). Below class, the architecture level

ISSN 0907-4449

simply describes the orientations of the secondary structures in three-dimensions without regard to their connectivity. We currently recognize 28 well defined architectures in *CATH*. Examples of complex arrangements of secondary structures, defying classification, are found in each class and are assigned to a complex category.

Within a given architecture, the topology or fold is then determined by the connectivity of the secondary structures. Fig. 1 illustrates different architectures and topologies currently observed in the α–β class. Proteins are assigned to a given fold family if they are structurally similar to at least one member of that family. Similarity is determined by considering scores returned by the structure comparison algorithm SSAP (Taylor & Orengo, 1989) (see below). There are currently 600 fold families in CATH (see Fig. 2). Within each fold family, structures are further grouped into homologous families whenever there is sufficient evidence of an evolutionary relationship. This will be discussed in more detail below. There are now more than 800 homologous families identified.

1.1. *Population of fold families in CATH*

Fig. 2 summarizes the number of groupings currently identified in the CATH hierarchy and the population of the different fold families and architectural groupings is illustrated by the CATHerine wheel shown in Fig. 3. Numbers given are for release 1.3 of CATH (November 1997). A recent analysis of the population of different fold families is reported in Orengo *et al.* (1997) and Swindells *et al.* (1998). Results confirmed a previous observation that some families are very highly popu-

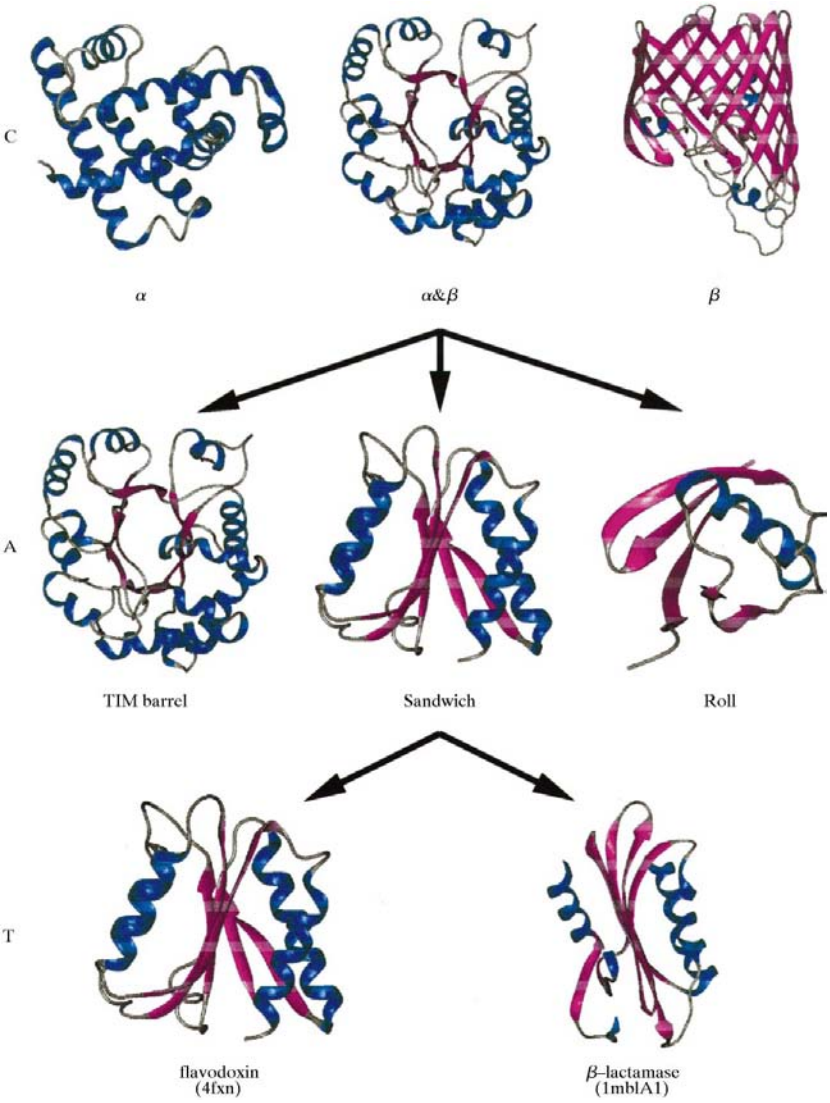


Fig. 1. Schematic representation of the (C)lass, (A)rchitecture and (T)opology/fold levels in the CATH database.

lated, containing many homologous relatives having low sequence similarity (<20%) and also containing three or more analogous proteins having no functional similarity (Orengo *et al.*, 1994). These 'superfold' families, of which there are currently ten, account for approximately one third of non-homologous structures in *CATH* and in some of these families more than ten different functions are currently observed. For example, in the TIM-barrel family, there are 14 different functions, whilst the doubly wound fold is adopted by proteins exhibiting nearly 30 different functions. Interestingly, though, for both TIM and doubly wound families, substrates often bind in similar structural locations (Martin *et al.*, 1998).

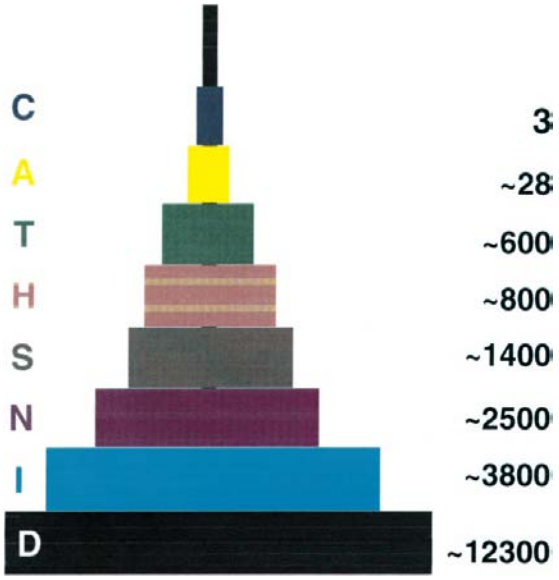


Fig. 2. Pyramid plot showing the number of groups identified at each level in the CATH database. Characters on the left-hand side gives the CATH levels: (C)lass; (A)rchitecture, (T)opology; (H)omologous superfamily; (S)equence family, 35% sequence identity; (N)ear-identical, 95% sequence identity; (I)dentical, 100% sequence identity; (D)omain entry.

Importantly, Fig. 4 shows that less than one quarter of new sequences (*i.e.* those which have less than 25% sequence identity to any structure in CATH), are found to adopt a novel fold. This suggests that we may now have representatives for many of the major folds occurring in nature. This is supported by recent attempts to find structural relatives for sequences from a number of microbial genomes. Several groups were successful in assigning structures to more than 35% (Huynen *et al.*, 1998) and 47% of the microbial genome sequences (Jones *et al.*, 1998). Obviously, non-globular proteins and membrane-bound proteins are poorly represented at the moment. However, these observations do emphasize the need to focus on solving the structures of new sequences for which no structural relatives can be found by database searches – either using robust sequence-alignment methods (*e.g.* PSI-BLAST, Altschul *et al.*, 1997; Intermediate Sequence Searching, Park *et al.*, 1997) or one-dimensional–three-dimensional fold recognition strategies, such as threading (Jones *et al.*, 1992).

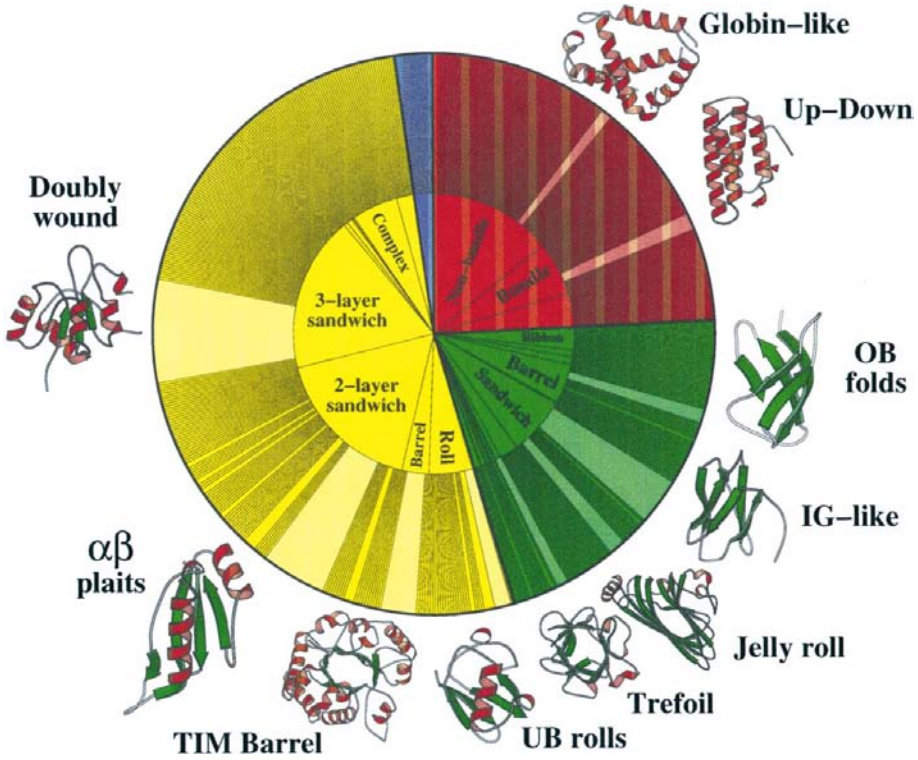


Fig. 3. CATH wheel plot showing the population of homologous families in different fold groups, architectures and classes. The wheel is coloured according to protein class (red, mainly α; green, mainly β; and yellow, $\alpha\beta$). The size of the outer wheel represents the number of homologous families in CATH whilst each band in the outer wheel corresponds to a single fold family. The size of each 'fold band' therefore reflects the number of homologous families having that fold. It can be seen that most fold families contain a single homologous family. The superfold families, are shown as paler bands, containing many homologous families. The inner wheel shows the population of homologous families in the different architectures.

§2, below, explains how CATH data is presented on the WWW and describes the CATH glossary and lexicon. Details are given on how to search for a particular structure in CATH and on the derived data available at different levels in the classification. We have recently set up a CATH server which allows the user to submit the coordinates of a newly determined structure for automatic classification in CATH. §3 describes how this works. The server also provides a list of structural neighbours and alignments are given for the five highest scoring matches. In §4 we briefly discuss some of the problems encountered in classifying structures and some planned future improvements.

2. CATH on the WWW – derived data and links to other databases

The CATH database of protein domains structures is a hierarchical classification of domains, generated by a combination of automatic and manual approaches. It currently contains nearly 9000 chains from the Brookhaven Protein Data Bank (Abola *et al.*, 1987). Only well resolved (<3.0 Å) structures are included and there are no models, or synthetic proteins. The procedure followed in classyifying domain structures is outlined in the flow chart shown in Fig. 5 and the concepts summarized below. For detailed descriptions of the methods applied at each stage of the classification see Orengo *et al.* (1997).

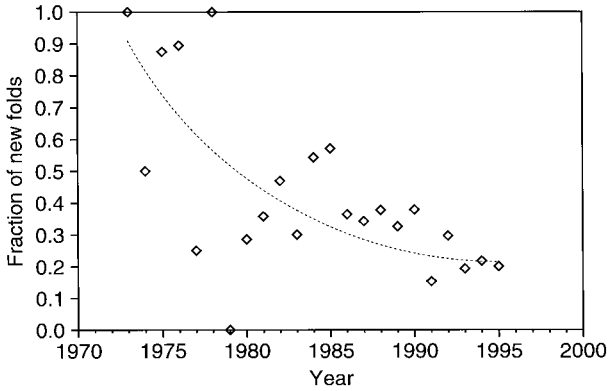


Fig. 4. Annual fraction of newly determined structures, having less than 25% sequence identity to any protein in the CATH database, which are found to have a novel fold.

CATH may be viewed over the WWW (http://www.biochem.ucl.ac.uk/bsm/cath) together with an associated glossary which defines all the structural terms used (*e.g.* β-hairpin) and a lexicon which provides a description of each class and architecture, together with *Rasmol* pictures for representative structures (Sayle & Milner-White, 1995). Each entry in CATH is linked to the PDBSUM database (Laskowski *et al.*, 1997) which generates WWW pages showing summary information derived from the PDB file *e.g.* secondary structure, supersecondary motifs (*PROMOTIF*, Hutchinson & Thornton, 1996), hydrogen-bonding plots (*HERA*, Hutchinson & Thornton, 1990), ligand interaction plots (*LIGPLOT*, Wallace *et al.*, 1995) and validation data (*PROCHECK*, Laskowski *et al.*, 1993).

There are also links from CATH entries to SWISS-PROT (Bairoch & Boeckman, 1992) and *PRINTS* (Attwood *et al.*, 1994). For each fold family, there are two-dimensional matrices giving the pairwise SSAP scores and sequence identities between all non-identical relatives. There are also summary tables showing the secondary structures observed in each relative, lengths of the proteins and other general information. Multiple structural alignments are currently being generated in each family using the program *CORA* (Orengo, 1998) (see below) and will be available shortly.

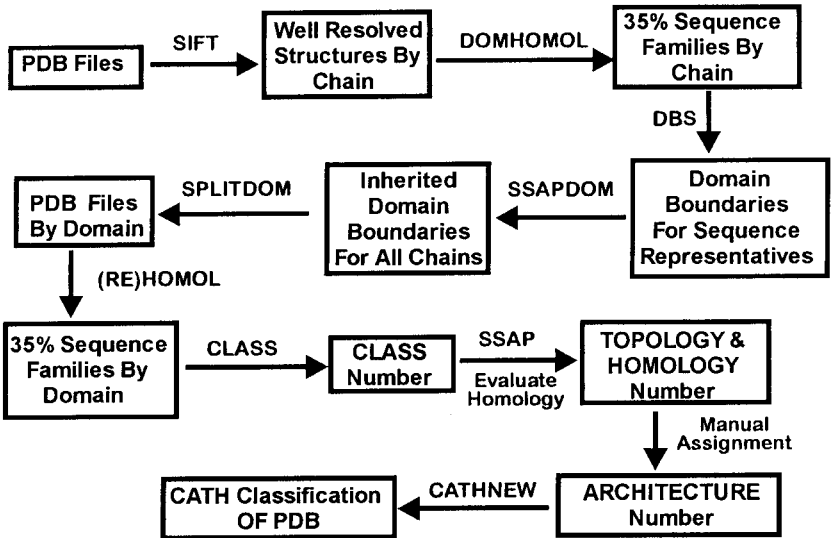


Fig. 5. Flow chart of the procedures used in generating the CATH database of structural domains. Each box summarizes a different step and any programs used, are given above the arrows.

2.1. *CATH numbers*

Each level in CATH has an associated numeric identifier (see Fig. 6). At the architecture, fold and homologous family level, numbers are currently incremented in bins of ten to allow for expansion of the database. Numbering entries in this way is essential for efficient data managment but can also be useful for research and diagnostics. For example, the hits from a threading trial can be quickly assessed for fold similarity, by inspecting the CATH numbers.

3. The CATH classification procedure and the CATH server

The CATH server is a relatively simple WWW-based tool which scans a new protein structure against the CATH database. It currently runs on a four-processor Origin 200, obtained by funding from the Medical Research Council. The procedures used in the server, mirror all the automatic stages in the CATH classification. Domains are first assigned using the *DETECTIVE* program (Swindells, 1995). The sequences of all domains are then compared using a global dynamic programming method (*domhomol*, Orengo *et al.*, 1997). Subsequently, protein class is assigned. Architecture and a set of possible fold families having similar topology are then identified using a fast topology-scanning program (Martin, 1998). Finally, non-identical structures from these families are aligned against the new structure, more carefully, using the structure-comparison method SSAP (Taylor & Orengo, 1989), to identify the correct fold family.

3.1. *Assignment of domain boundaries in CATH*

Since CATH is a domain-level database, a submitted structure must first be split into domains. We have decided to identify similarities between domain folds, because we consider the domain to be a fundamental level in the structural hierarchy and probably an evolutionary unit. It is reasonable to assume that it will be easier to predict structure at this level. Furthermore, by separating multidomain proteins into their constituent folds, we hope to make it easier to subsequently study domain interactions.

However, assigning domains from the three-dimensional coordinates is not trivial. Although there are now more than 30 different methods available, most adopt a common philosophy of searching for large hydrophobic clusters or 'cores' and separating putative domains in such a way as to maximize the number of internal contacts or compactness measured for each domain. However, a survey of some of the most robust methods available, revealed that none could be applied comple-

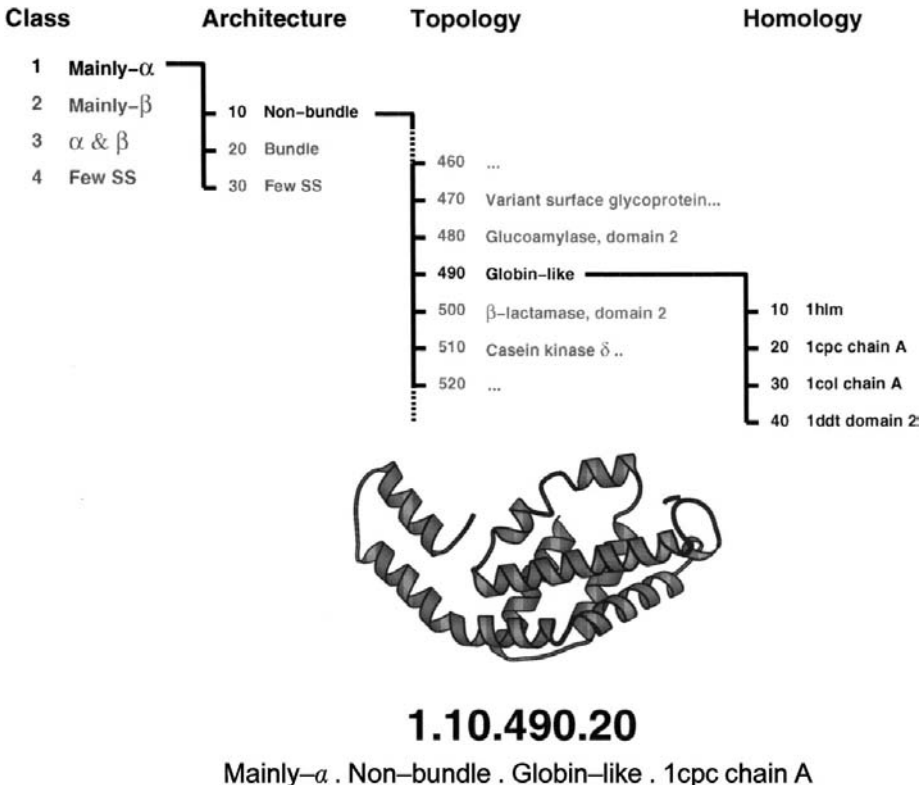


Fig. 6. Representative example of the use of numeric identiers at each level in CATH.

tely automatically as they only gave the correct answer about 70% of the time (Jones *et al.*, 1998).

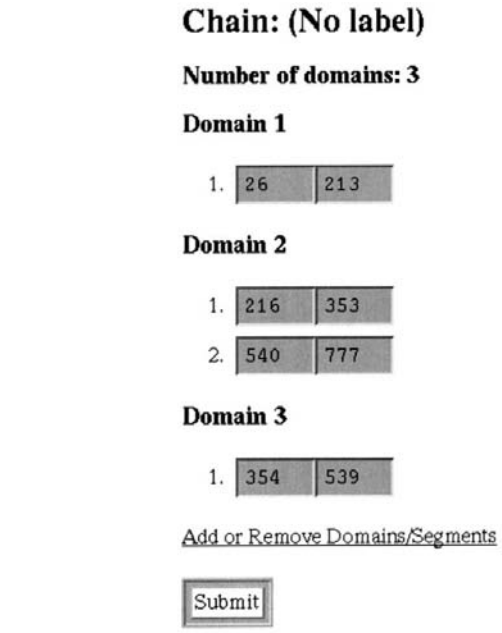
Therefore, for our classification, we chose to adopt a consensus approach (Jones *et al.*, 1998) whereby three independent methods are applied (*DETECTIVE*, Swindells, 1995; *PUU*, Holm & Sander, 1993*b*; *DOMAK*, Siddiqui & Barton, 1995). When these methods agree in the number of boundaries assigned and provided these boundaries are within five residues, the assignments are automatically accepted. When they disagree, assignments from each method are visually inspected and the most reasonable selected.

Recent application of this consensus approach showed that the majority (>95%) of single-domain proteins could be assigned automatically. However, for multidomains we can only assign automatically about 20% of the time but one of the methods always returns a good result. Since the methods are fast this usually represents a considerable saving in time over assigning boundaries by eye. Since nearly one-third of all proteins determined are currently multidomain, time is an important consideration. Especially as the proportion of multidomains will probably increase as techniques for solving larger structures, improve.

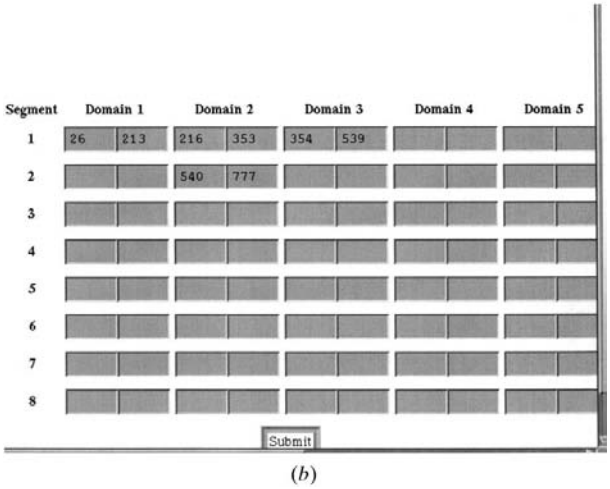
In CATH, all proteins which are not single domains according to our consensus method, are divided into two or more separate domains. This is in contrast to the approach taken by some other groups. For example, although Holm & Sander apply an automatic algorithm to search for boundaries (PUU, Holm & Sander, 1993*b*, see also http://www2.ebi.ac.uk/dali/domain), domains are only separated provided they are observed to have occurred also as single domains. This evolutionary consideration may ensure that boundaries are assigned more reliably and consistently. However, it may also under-represent the number of separate domain folds in nature. In our recent analysis of CATH (Orengo *et al.*, 1997), we found quite a low level of domain recurrence (<10%), though this may also reflect some distortion of domains when assembled into a multidomain protein and may suggest that greater tolerance should be used when searching for structural similarities between domains.

In the SCOP database, a more pragmatic approach is adopted as all domain boundaries are assigned manually. Evolutionary similarities are recognized. However, where visual inspection reveals clear boundaries, unique domain folds are also recognized and classified. Classifications of structures in the 3*Dee* database (Barton, 1997) and Ddbase (Sowdhamini *et al.*, 1996) follow the same approach as in CATH and rely on automatic algorithms to identify separate domains, which are subsequently checked manually, where necessary.

3.1.1. *Domain assignment in the CATH server*. Since, not one of the domain-identification programs is, in itself, reliable enough to perform this task in an automated fashion, we have used the *DETECTIVE* program (Swindells, 1995) which is relatively good at identifying multi-domain proteins even when it does not define the domain boundaries accurately. The results from *DETECTIVE* are returned to the user in less than a minute (see Fig. 7*a*) and domains may be further split or merged or the boundaries may be moved using a WWW form (see Fig. 7*b*). A simple consistency check on the entered data is made using routines written in JavaScript before the form is submitted.



(*a*)



(*b*)

Fig. 7. (*a*) Domain table generated by the CATH server, showing domain assignments by the *DETECTIVE* program (Swindells, 1995). (*b*) A definition table provides the user with the option to modify the automatically generated domain boundaries or supply alternative ranges.

3.2. *Assessing sequence similarity*

Protein sequences can be aligned an order of magnitude faster than their structures. Provided at least 30% of the sequences are identical, the alignments will be reliable and several studies have confirmed that 30% or more sequence identity clearly indicates that the proteins are homologous and will have highly similar structures (Chothia & Lesk, 1986, Sander & Schneider, 1991; Flores, Orengo *et al.*, 1993). Therefore, in CATH, the sequences are always compared before comparing the structures directly. A global alignment method is used (*domhomol*, Orengo *et al.*, 1997), based on the algorithm of Needleman & Wunsch (1970).

Single linkage clustering is then applied to group proteins having 35% (sequence-level), 95% (near-identical level) and 100% (identical-level) sequence identity. To prevent proteins being grouped on the basis of local similarity, we check that at least 60% (S-level), 85% (N-level) and 100% (I-level) of residue positions in the larger protein are equivalent to positions in the smaller protein.

In the CATH server, if a sequence match with percentage identity greater than 95% is found and at least 85% of the larger protein is equivalent to the smaller, then we assume that the domain is nearly identical to one in CATH and a link is provided to the CATH entry for that hit. If a sequence hit is found with percentage identity greater than 30% then we assume that proteins are homologous (*i.e.* the CATH number will be the same) and we run structure comparisons (see below) against representatives from all the sequence families (S-level) for this homologous family (H-level).

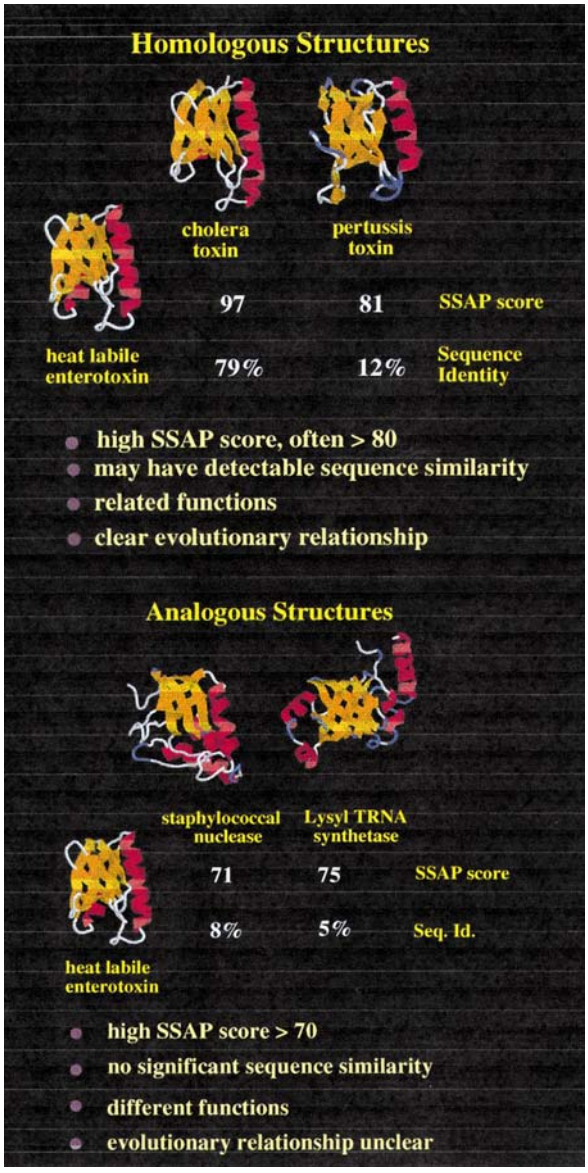


Fig. 8. (*a*) *Molscript* (Kraulis, 1991) representatives of homologous proteins in the OB fold family and criteria used for assigning homologues in CATH. (*b*) *Molscript* representatives of analogous proteins in the OB fold family and criteria used for assigning analogues in CATH.

3.3. *Assessing structural similarity*

Structure comparison algorithms are applied to detect more distantly related homologues and analogues, having no significant sequence similarity. There are now many examples of proteins having the same fold despite very low sequence identities. Fig. 8 shows how structure is conserved for representatives from the OB fold family, some of which have similar functions and are therefore homologues. Other members of the family having very low sequence identities and exhibiting different functions are described as analogues, since the evolutionary relationship is unclear. These proteins may be very distant evolutionary relatives which have diverged to a point where no sequence similarity remains and the functions too have changed. Alternatively, they may be examples of convergent evolution, whereby the possible number of arrangements of secondary structures in three-dimensions are limited by physical constraints, so that proteins adopting the same fold may be the result of a common solution to these constraints rather than examples of divergent evolution.

In CATH, structural similarity is assessed by using the SSAP comparison algorithm (Taylor & Orengo, 1989; Orengo *et al.*, 1992) which aligns proteins by comparing internal residue geometry expressed as 'views' or sets of vectors from the Cβ atom to all other Cβ atoms in the same protein. Dynamic programming is applied to handle insertions/deletions but since these also affect the comparison of views, dynamic programming is applied at two levels (double dynamic programming, see Orengo & Taylor, 1996, for a review). Early empirical trials using extensive SSAP comparisons between

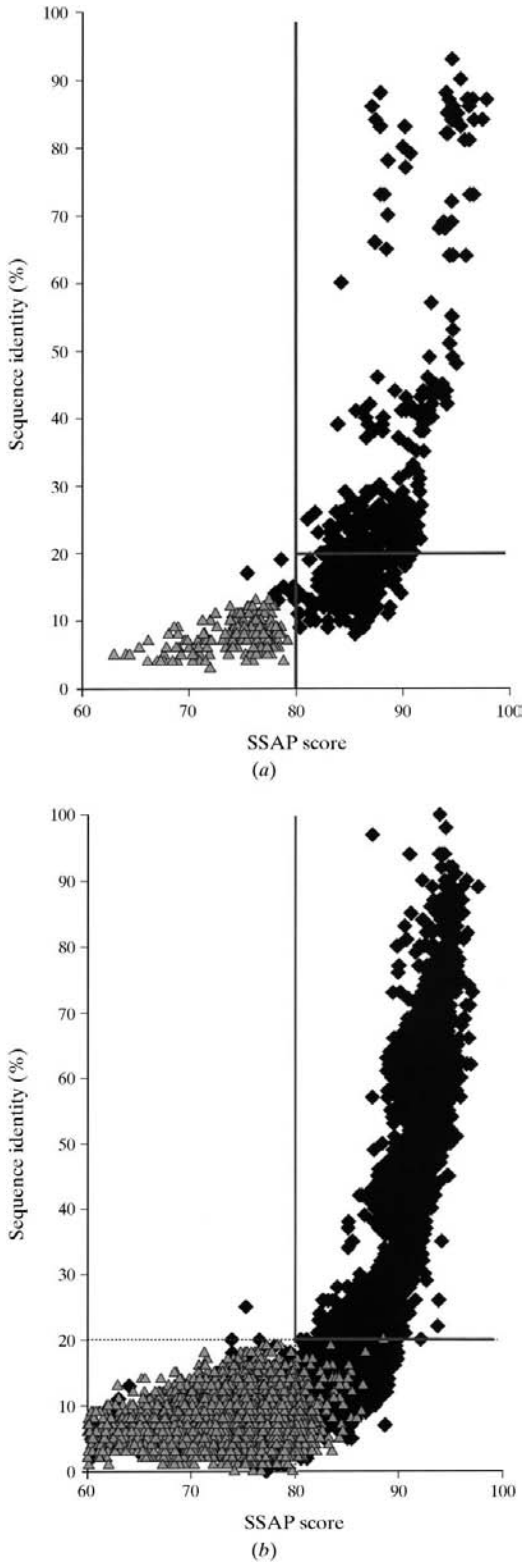


Fig. 9. Plots showing structure comparison (SSAP) scores and sequence identities (%) for homologous (black dots) and analagous (grey dots) protein pairs from the (*a*) mainly α, globin-like fold and (*b*) mainly-β, immunoglobulin-like fold.

structures in the PDB, were used to identify cutoffs on the score, for related proteins (Orengo *et al.*, 1992).

At the time of establishing these cutoffs, there were fewer than 2000 structures in the PDB. Therefore, we have recently re-examinined these criteria using the current much larger data set of structures in CATH and checking that cutoffs can be applied consistently for all families in CATH, particularly the highly populated 'superfolds'. SSAP scores range from 0 to 100 for complete identity. Tests showed that structural pairs returning a SSAP score greater than 80 and having 20% or more sequence identity (measured from the structural alignment), could automatically be assigned as homologues, even within superfold families (see Fig. 9). Whilst lower values of 70–80 were occasionally returned by very distant homologues as well as by analogous pairs having the same fold but no sequence or functional similarity.

Occasionally, visual inspection of pairs scoring in the low 70's did not reveal sufficient global similarity in the folds for the proteins to be grouped into the same family, and scores appeared to be the result of high recurrence of common motifs (*e.g.* $\beta\alpha$ motifs in the $\alpha-\beta$ class). Therefore, in current updates of CATH, all pairs scoring in these ranges (SSAP 70–80 and sequence identity <20) are visually inspected. Functional information is also considered when assigning homologues. Functional keywords are automatically extracted from SWISS-PROT (Orengo *et al.*, 1997) and by reference to the

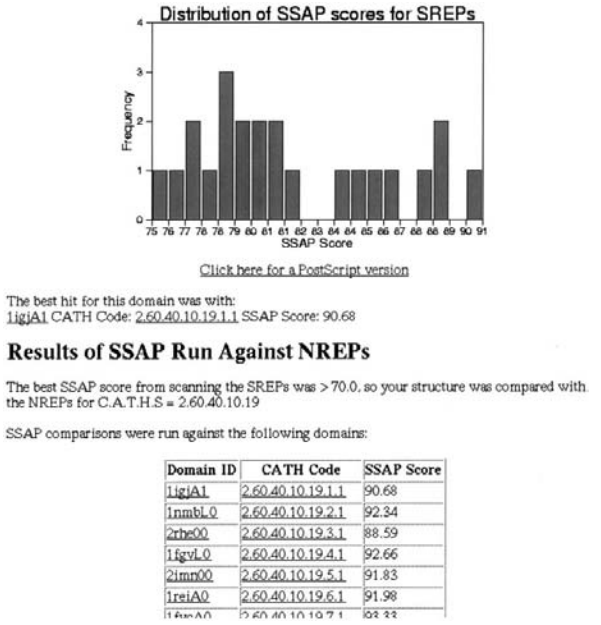


Fig. 10. Diagnostic report from the CATH server, giving a histogram of SSAP scores obtained by scanning the new structure against representatives from CATH. A list of top scoring structural neighbours is provided, together with associated SSAP scores, sequence identities and links to relevant CATH families.

litererature and other databases containing relevant information (*e.g.* SCOP, Murzin *et al.*, 1995; ENTREZ, Hogue *et al.*, 1996). Common ligand interactions can also be used to suggest homologous proteins and are automatically detected in ligand-bound structures using the program *GROW* (Milburn *et al.*, 1998). Fig. 8 summarizes the criteria currently applied for grouping proteins into fold families (T-level) and homologous families (H-level).

Since the SSAP comparison is relatively time-consuming, the class of the protein is first determined automatically by a method which measures both residue composition and secondary-structure contacts and applies empirically derived cutoffs (Michie *et al.*, 1996). Subsequently, proteins assigned to the mainly α class are not compared with those in the mainly β class and *vice versa*. Furthermore, a fast topology scanner (*TOPSCAN*, Martin, 1998) is currently being implemented which will compare secondary-structure characteristics for the new structure against those of all non-identical structures in each fold family, to identify possible fold families to which the new structure may belong.

3.3.1. *Assigning structural family in the CATH server*. After assigning protein class to the new structure, a list of posible fold families is generated using TOPSCAN (Martin, 1998). Subsequently, a fast version of SSAP (Orengo *et al.*, 1992) scans representatives from all the

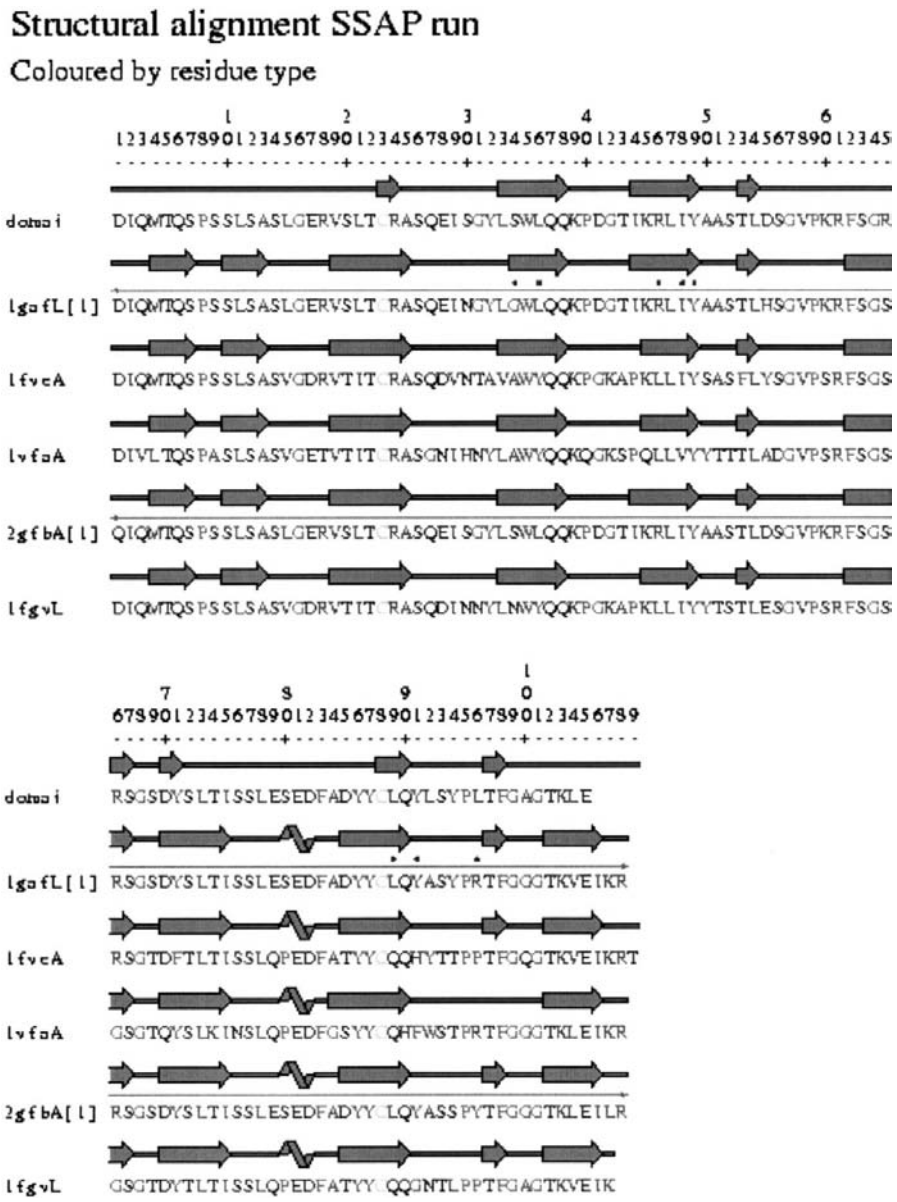


Fig. 11. Diagnostic report from the CATH server showing the multiple structure alignment of the probe structure with the five highest scoring structures from the SSAP scan. The aligment plot is generated using graphical software in the SAS package (Milburn *et al.*, 1998).

sequence families (S-level) in those possible fold families. The normal (slow) version of SSAP is then used to scan the non-identical representatives of the S-families returning the top ten scores.

In all cases where SSAP has been run, barcharts showing distributions of scores are displayed together with tables of scores and hits which are clickable links back to the appropriate CATH family and derived data (see Fig. 10). Finally, the SSAP structural alignment between the submitted domain and the top five matching structures is displayed using a graphical display package (SAS, Milburn *et al.*, 1998) (see Fig. 11).

3.4. *Assigning architecture in CATH*

This stage in the CATH update is performed manually, though automatic methods are currently being researched for assigning some of the simpler architectures. Architectural groups represented in CATH build on the early work of Richardson (1981), in describing common secondary-structure arrangments. They also include descriptions given by crystallographers and NMR experts, for recently identified architectures (*e.g.* β-helix, Kobe & Deisenhofer. 1993; β-horseshoe, Yoder *et al.*, 1993), together with architectural terms applied by other structural analysts for commonly observed arrangements (*e.g.* $\alpha\beta\alpha$ sandwich, Richardson, 1981).

CATH is unique in assigning an architectural descriptor to each fold family in the database. Other classifications such as SCOP, although they often describe secondary-structure arangements (*e.g.* β-sandwich, β-barrel), provide no formal grouping of folds into their respective architectures. We consider architecture to be a useful level in the hierarchy, not just from a consideration of data management, though grouping common architectures does help with validation. Also, because our most recent analysis of CATH fold families, demonstrated that over 70% of folds could easily be assigned to very simple architectures. In particular, the β-sandwiches, β-barrels, $\alpha\beta$-sandwiches (two- and three-layer) and $\alpha\beta$-barrels are highly favoured (see Fig. 3). Therefore, it is hoped that by grouping structures in this way, we shall learn more about the physical constraints governing the packing of helices and strands in these groups and this will enable us to automate the recognition of common regular architectures in the future.

4. Problems and future developments

4.1. *The Russian Doll Effect – difficulties in using single linkage clustering for assigning proteins to structural families*

The majority of fold families in CATH (>620 out of 630) contain only homologous proteins and these often have significant sequence similarity (>25%) which is

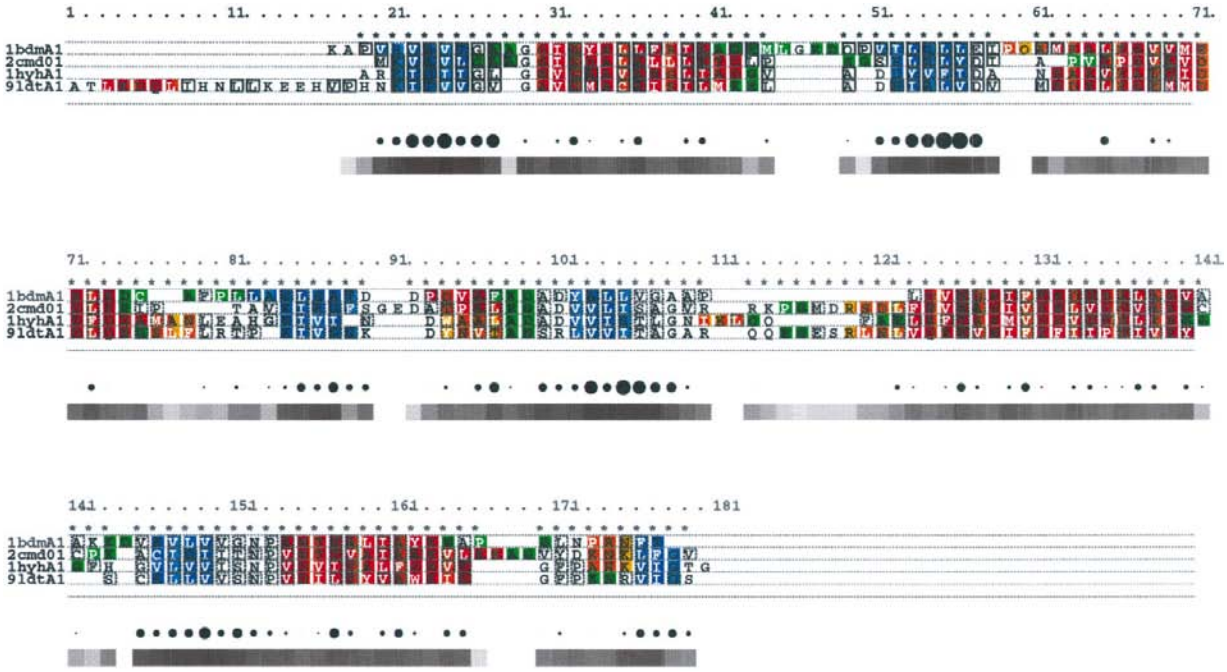


Fig. 12. Multiple structure-derived sequence alignment of a set of diverse relatives from the NAD-binding Rossmann-fold family. Red boxes show residues which adopt α-helical conformation, blue, β-strand; orange, 3_{10} helices; and green, turn. The shaded bar shows the relative structural conservation at each position in the alignment, with black for the most highly conserved positions. Black dots label positions involved in multiple contacts in the buried core, and are scaled according to the relative number of contacts.

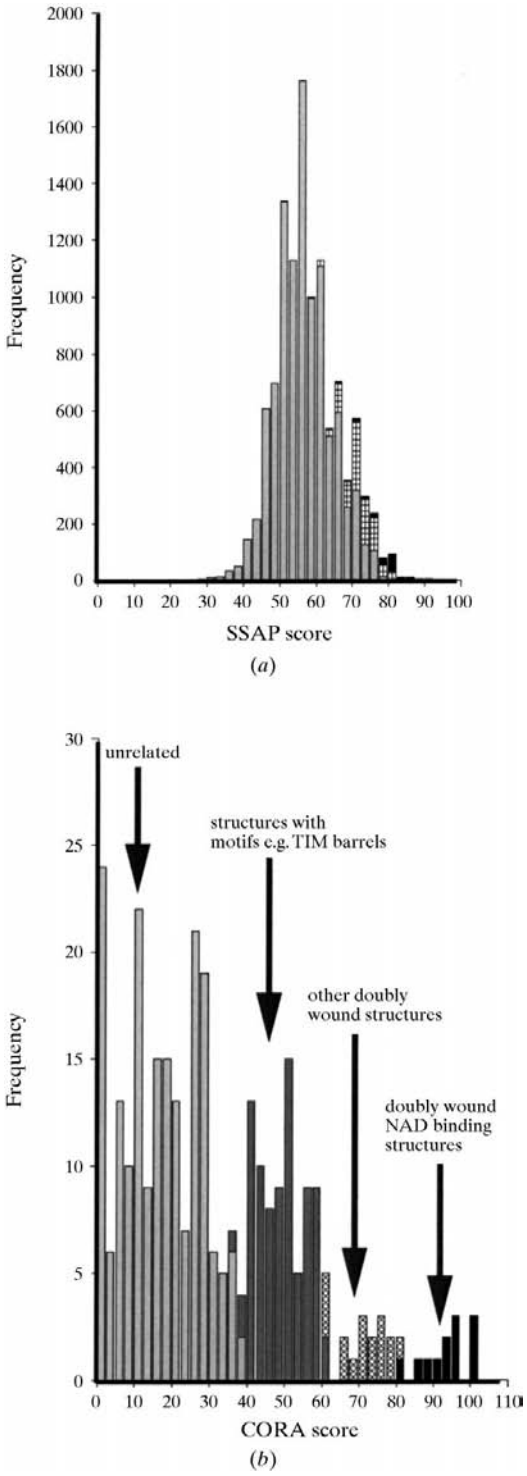


Fig. 13. Distributions of structure comparison scores returned by structural relatives and non-relatives. (*a*) Scores obtained by scanning non-identical structures from the PDB, with relatives from the NAD-binding Rossmann fold family, using the pairwise structure comparison algorithm SSAP. (*b*) scans through the same data set using the *CORA* template for the NAD-binding Rossmann fold family.

accompanied by considerable structural similarity. Recognizing new relatives of these families is relatively easy as any member of the family can be expected to give a good pairwise SSAP score against a new structural relative.

However, for the superfold families, relatives exhibit far more structural variation. This means that whether a new relative is recognized will depend on how similar it is to the CATH relatives with which it is compared. As increasing numbers of structures are solved and CATH expands correspondingly, this could result in new relatives being missed because they are too distant from the representatives chosen for the family. One solution would obviously be to compare new structures against all the proteins in these families, but this would be too time-consuming for CATH to keep pace with the growth in the PDB.

Another problem encountered when classifying proteins is that of deciding when structures have sufficient global similarity to be grouped into the same fold family. This problem is particularly apparent for some families in the highly populated architectures in the mainly β and $\alpha\beta$ classes (*e.g.* sandwiches and barrels). In both classes, proteins adopting these architectures tend to exhibit high recurrence of common motifs (*e.g.* β-hairpins, classic and split $\beta\alpha\beta$ motifs). Because proteins are grouped in CATH using single linkage clustering, this can result in families growing by the addition of structures having increasing numbers of similar motifs. This problem has been described as the Russian Doll Effect (Orengo *et al.*, 1997) and reflects the fact that for some architectures, protein fold space can be better described as a continuum, in which it is sometimes difficult to distinguish distinct folds, because families can often be merged by addition or deletion of one or more motifs.

The approach which has now been adopted to tackle both these difficulties and to enable more coherent definitions of structural families, is to derive consensus three-dimensional templates for each structural family in CATH. Using the program *CORA* (Orengo, 1997, 1998), a multiple structural alignment is generated using the most diverse representatives from the family, in order to identify features which have been highly conserved during evolution and are, therefore, the most essential to the stability and/or function of proteins in the family. Once identified these highly conserved structural features are described in the form of a consensus template. For each conserved residue position, templates contain average vectors between Cβ atoms, calculated over all the representative structures, also average accessibility, average torsional angles *etc.* Average properties of secondary-structure interactions (*e.g.* midpoint separations, pairwise angles) are also calculated and stored in the template, together with information about highly conserved residue contacts observed for the family.

CORA templates are automatically derived for both fold families and homologous families in CATH. New structures are then compared to templates for each structural family using the alignment program *coralign* (Orengo, 1998), which uses a similar double dynamic programming method to SSAP, but includes information about consensus residue contacts in the family, to improve sensitivity. Fig. 12 shows the *CORA* multiple alignment for representatives in the NAD-binding doubly wound Rossmann fold family and Fig. 13 demonstrates the improved performance of the *CORA* templates in identifying homologues and analogues. *Coralign* also returns diagnostic information on whether structures matching the template contain a sufficient proportion of highly conserved family characteristics, to be assigned to the family. This means new structures will only be assigned provided they show significant global similarity over a range of critical structural properties, rather than on the basis of a single SSAP score. This should guarantee consistent classification in CATH, despite the huge increases expected in the number of known structures, over the next five years. *CORA* templates and multiple structural alignments will be made available over the WWW, for each structural family.

4.2. *Multidomain information in CATH*

Often a protein's functional unit can only be studied by examining the multidomain protein. The serine and aspartyl proteins are good examples. Therefore, we now have an additional 'class' in CATH consisting of multidoman proteins grouped according to sequence similarity into families (S-, N-, I-levels) as for the single domain proteins. Structural similarity is also considered using SSAP comparisons and proteins grouped into homologous families (H-level) using the same criteria as for single domains. However, there is no formal architecture assignment or fold grouping.

Christine Orengo acknowledges the Medical Research Council. Andrew Martin acknowledges Oxford Molecular, David Jones the Royal Society and Alex Michie, the BBSRC.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein Data Bank, in *Crystallographic Databases – Information Content, Software Systems, Scientific Applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn/Cambridge/Chester: IUCr.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. (1994). *Protein Eng.* **7**, 841–848.

Bairoch, A. & Boeckman, B. (1992). *Nucleic Acids Res.* **20**, 2019–2022.

Barton, G. J. (1997). *3-Dee Database of Protein Domain Definitions*. http://circinus.ebi.ac.uk:8080/3Dee/help/help_intro.html

Chothia, C. (1993). *Nature (London)*, **357**, 543–544.

Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5**, 823–826.

Flores, T. P, Orengo, C. A. & Thornton, J. M. (1993). *Protein Sci.* **7**, 31–37.

Holm, L., Ouzonis, C., Sander, C., Tuparev, G. & Vriend, G. (1993). *Protein Sci.* **1**, 1691–1698.

Holm, L. & Sander, C. (1993*a*). *J. Mol. Biol.* **233**:123–138.

Holm, L. & Sander, C. (1993*b*). *Proteins*, **19**, 256–268.

Hogue, C. W. V., Ohkawa, H. & Bryant, S. H. (1996). *Trends Biochem. Sci.* **21**, 226–229.

Hutchinson, E. G. & Thornton, J. M. (1990). *Proteins*, **8**, 203–212.

Hutchinson, E. G. & Thornton, J. M. (1996). *Protein Sci.* **5**, 212–220.

Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C. A., Sunyaev, S., Yuan, Y. & Bork, P. (1998). *J. Mol. Biol.* Submitted.

Jones, D. T., Taylor, W. R. & Thornton. J. M. (1992). *Nature (London)*, **358**, 86–89.

Jones, S., Swindells, M. B., Stewart, M., Michie, A. D., Orengo, C. A. & Thornton, J. M. (1998). *Protein. Sci.* **7**, 233–242.

Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.

Kobe, B. & Deisenhofer, J. (1993). *Nature (London)*, **366**, 751–756.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). *Trends Biochem Sci.* **22**, 488–490.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.

Martin, A. C. R. (1998). In preparation.

Martin, A. C. R., Orengo, C. A., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J., Taroni, C. & Thornton, J. M. (1998). *Nature Struct. Biol.* Submitted.

Michie, A. D., Orengo, C. A. & Thornton, J. M. (1996). *J. Mol. Biol.* **262**, 168–185.

Milburn, D., Laskowski, R. A. & Thornton, J. M (1998). *Protein Eng.* Submitted.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453

Orengo, C. A. (1997). *Methods Enzymol.* **266**, 617–635.

Orengo, C. A. (1998). *J. Mol. Biol.* Submitted.

Orengo, C. A., Brown, N. P. & Orengo, C. A. (1992). *Proteins*, **14**, 139–167.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). *Nature (London)*, **372**, 631–634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Orengo, C. A. & Taylor, W. R. (1996). *Methods Enzymol.* **266**, 617–635.

Park, J., Teichmann, S. A., Hubbard, T. & Chothia. C. (1997). *J. Mol. Biol.* **273**, 349–354.

Richardson, J. S. (1981). *Adv. Prot. Chem.* **34**, 167–339.
Sander, C. & Schneider, R. (1991). *Proteins*, **9**, 56–68.
Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem Sci.* **20**, 374–376.
Siddiqui, A. S. & Barton, G. J. (1995). *Protein Sci.* **4**, 872–884.
Sowdhamini, R., Rufino, S. D. & Blundell, T. L. (1996). *Folding Des.* **1**, 209–220.
Swindells, M. B. (1995). *Protein Sci.* **4**, 103–112.
Swindells, M. B., Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Hutchinson, G., Martin, A. & Thornton, J. M. (1998). *Bioessays*. In the press.
Taylor, W. R. & Orengo, C.A. (1989). *J. Mol. Biol.* **208**, 1–22.
Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). *Protein. Eng.* **8**, 127–134.
Yoder, M. D., Lietzke, S. E. & Kurnak, F. (1993). *Structure*, **1**(4), 241–251.