A systematic comparison of protein structure classifications: SCOP, CATH and FSSP

Caroline Hadley and David T Jones*

Background: Several methods of structural classification have been developed to introduce some order to the large amount of data present in the Protein Data Bank. Such methods facilitate structural comparisons and provide a greater understanding of structure and function. The most widely used and comprehensive databases are SCOP, CATH and FSSP, which represent three unique methods of classifying protein structures: purely manual, a combination of manual and automated, and purely automated, respectively. In order to develop reliable template libraries and benchmarks for protein-fold recognition, a systematic comparison of these databases has been carried out to determine their overall agreement in classifying protein structures.

Results: Approximately two-thirds of the protein chains in each database are common to all three databases. Despite employing different methods, and basing their systems on different rules of protein structure and taxonomy, SCOP, CATH and FSSP agree on the majority of their classifications. Discrepancies and inconsistencies are accounted for by a small number of explanations. Other interesting features have been identified, and various differences between manual and automatic classification methods are presented.

Conclusions: Using these databases requires an understanding of the rules upon which they are based; each method offers certain advantages depending on the biological requirements and knowledge of the user. The degree of discrepancy between the systems also has an impact on reliability of prediction methods that employ these schemes as benchmarks. To generate accurate fold templates for threading, we extract information from a consensus database, encompassing agreements between SCOP, CATH and FSSP.

Address: Protein Structure Group, Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

*Corresponding author.
E-mail: jones@globin.bio.warwick.ac.uk

Key words: evolution, protein folds, protein-structure classification, taxonomy

Received: 28 January 1999
Revisions requested: 12 March 1999
Revisions received: 1 April 1999
Accepted: 25 May 1999

Published: 27 August 1999

Structure September 1999, 7:1099–1112
http://biomednet.com/elecref/0969212600701099

0969-2126/99/$ – see front matter

Introduction

Since the creation of the Protein Data Bank (PDB [1]) over twenty years ago, more than 8000 protein structures have been deposited. With experimental techniques becoming more advanced and less time consuming for solving protein structures, the rate of growth in structural information is expected to rise even more rapidly. Although a great deal of information may be revealed by analysis of a single protein structure, it has long been understood that a more global, comprehensive view of proteins comes from a comparison of multiple structures, and investigations into their folding similarities and evolutionary relationships. A logical beginning to the comparison of protein structures is a system of classifying these structures in order to easily identify and group similar folds and families. An advantage of classifying proteins in this way is the prospect of introducing some sense of order to the growing volume of structural data available.

One complication that immediately arises in structure classification is the fact that protein structures are often composed of discrete globular domains. The commonly accepted definition of a domain is a compact, local, semi-independent folding unit built from secondary structure elements [2]. Because domains may function individually within a protein, with distinct functional and structural roles, proteins are usually separated into discrete domains before classification. The identification and delineation of domains within protein structures is a difficult and often subjective process. Although domains can often easily be distinguished by manual inspection, the automation of this process is not simple. Many algorithms exist for domain assignment, each relying on a different set of defined rules governing domain structure and packing, such as compactness [3,4], surface area [5], residue–residue contact maps [6] and hydrophobicity [7]. The difficulties in this first step of protein structure classification are an indication of the difficulty of this process in general.

Once defined, domains can then be classified at the levels of class, fold, and superfamily (and further subdivided into families). ‘Class’ is generally determined from the overall composition of secondary structure elements within a domain [8]. A ‘fold’ is determined from the number, arrangement, and connectivity (or topology) of these elements. A 'superfamily' consists of domains with similar folds and usually similar functions, suggesting common ancestry, often in the absence of detectable sequence similarity. Sequence similarity is usually taken into account in forming 'families', namely, groups of very closely related domains. Examples of families might be a group of the same proteins from different species, or perhaps different isozymes from the same species.

Several systems have arisen to address the need for structural classification, in particular SCOP (Structural Classification of Proteins [9]), CATH (Class Architecture Topology Homology [10]) and FSSP (Families of Structurally Similar Proteins [11]). One of the advantages to having these systems is that they represent three unique methods of classifying structural data: FSSP is based on a purely automated process, SCOP is almost completely manually derived, and CATH employs an intermediate process, using automated procedures along with human intervention.

SCOP organizes proteins in a hierarchy, from class down to fold, superfamily, and family [9]. A total of ten classes are defined (only the first four of which are considered here): all alpha, all beta, alpha and beta (α/β), alpha plus beta ($\alpha+\beta$), multidomain, membrane and cell-surface proteins and peptides, small proteins, peptides, designed proteins, and non-protein structures. Although the SCOP protein classification is essentially a manual process using visual inspection and comparison of structures, some automation is used for the most routine tasks such as clustering protein chains on the basis of sequence similarity. Proteins are usually (but not always) separated into domains, and most of these domains are classified into one of the first five classes noted above. Structural similarities of proteins at the fold level often represent favourable packing arrangements and chain topologies, although some distant evolutionary links may exist. Common ancestry (i.e. homology) is more clearly defined upon classification into superfamilies, where proteins with similar structure and/or functional features are believed to share a common evolutionary origin. Proteins with similar sequences, or very similar structures and functions that imply a solid evolutionary link, are grouped together as families. Thus, members of the same family or superfamily within SCOP share common ancestry.

CATH is also a hierarchical system, which differs from SCOP in that it incorporates some automation in classifying protein structures. After extracting highly resolved structures from the Protein Data Bank (PDB), comparisons are made to group highly similar proteins on the basis of sequence similarity. A representative structure is taken from each sequence family, and is divided into domains using a consensus approach incorporating three automatic domain-assignment techniques [12]. Class (C-level) is defined first, to prevent unnecessary structure comparisons between different classes at a later stage. This step is primarily automated, although difficult cases may be dealt with manually. Domains are assigned to one of four classes (mainly α, mainly β, alpha beta ($\alpha\beta$), or few secondary structures) on the basis of composition, secondary-structure contacts and the proportion of parallel and antiparallel sheets [13]. Within each class, structure comparisons are made to produce fold groups (T-level) and then homologous superfamilies (H-level) [14,15]. The final stage is the manual assignment of architecture (A-level) using visual inspection and reference to literature. This stage is particularly important when considering novel folds. Together, these levels (C-A-T-H) produce an index or number for each domain; domains sharing C-A-T numbers have the same fold, whereas a shared H-level indicates a common evolutionary origin.

FSSP is known as both Families of Structurally Similar Proteins [16] and Fold classification based on Structure–Structure alignment of Proteins. Like SCOP and CATH, FSSP attempts to relate protein structures with respect to evolutionary relationships, although unlike CATH and SCOP, it is fully automated and does not assign proteins into classes, fold families or superfamilies. Instead, pairwise structural comparisons are made between proteins of a representative set (where no two proteins or domains have greater than 25% sequence similarity) and members of a sequence-homologue set (homologues with greater than 25% sequence identity) using the Dali program [17]. For each member of the representative set, a file is created containing all pairwise structural matches above a Z-score of 2.0 (pairs with values below this number are described as structurally dissimilar). Other information is presented in the file, along with the alignment information generated by Dali. Ultimately, a fold tree is constructed using hierarchical clustering methods; an indexing system is also incorporated by dividing the pairwise structural comparisons at Z-scores of 2, 3, 4, 5, 10 and 15 (creating a six-character index). These cut-offs are not an accurate distinction between protein folds or superfamilies.

Each of these three classification schemes has dedicated users, who tend to use one method consistently rather than try others with which they are unfamiliar. Recent papers demonstrate the extent to which structural classification databases are used in the bioinformatics field, from the analysis of protein structure [18] to the extraction of homologous structures [19,20], the testing of prediction methods [21–23] and the calculation of numbers of folds and families [24,25]. Other databases have applications in structural biology, such as VAST (Vector Alignment Search Tool, an algorithm that produces neighbourhoods of similar folds by performing structure–structure comparisons of all domains in the PDB

[26,27]) and HOMSTRAD (HOMologous STRuctural Alignment Database, which provides aligned three-dimensional structures of homologous proteins [28]). Databases of protein structural domain definitions, such as the DIAL-derived domain database (DDBASE [29]) and the Database of Protein Domain Definitions (3Dee [30]) can also be useful in structural investigations. However, SCOP, CATH and FSSP have the advantage of being the largest, most comprehensive and most frequently used classification databases available.

When a structural classification database is required for a specific purpose, such as (in our case) the production of fold recognition templates, it is not only important to choose the right database for the right reasons, but also to take into consideration the reliability of the classification scheme. The construction of fold templates aimed at the identification of distant superfamily members depends on highly accurate groupings of homologous sequences, such as would be expected in homologous families in SCOP and CATH, and FSSP pairwise matches with high Z-scores. However, with no indication as to the accuracy of this grouping in any of these databases, we found it necessary to investigate the three classification methods for content, reliability and accuracy. The widespread use of these databases in the field of bioinformatics certainly warrants an investigation into their reliability and structural agreement.

This analysis of SCOP, CATH and FSSP reveals a large percentage of agreement between the three databases, and highlights a number of important issues regarding these specific resources, and protein-structure classification in general. Because of the subjective nature of classification, using a combination of databases may be best; the benefit of a consensus structural-classification database for the production of reliable threading templates is currently being assessed.

Results and discussion

Shared-codes set

The comparison of SCOP, CATH and FSSP data produced 6875 common chains. When domains were subsequently taken into account, the SCOP set contained 8498 domains (74% of 11,515 domains in SCOP [March 1998, v 1.37]), and the CATH set contained 9874 domains (74% of 13,338 domains in CATH [April 1998, v 1.4]). The 6875 shared chains represented approximately 78% of the 8805 chains taken from FSSP. The resulting set of shared PDB codes is referred to as pset3.

Database comparison

Comparing FSSP pairwise matches to both SCOP and CATH

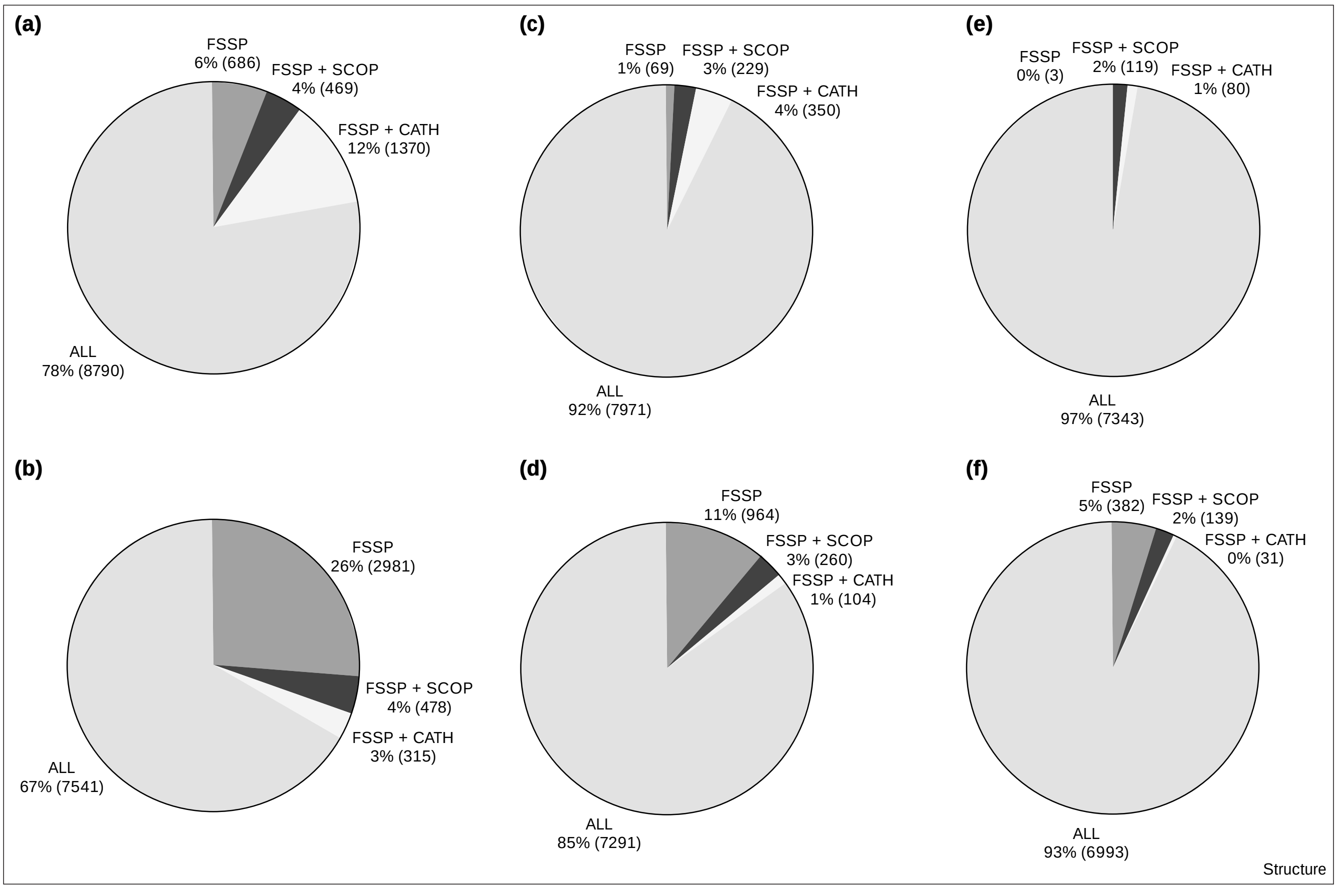
The comparison of FSSP pairwise matches against SCOP and CATH is shown in Figures 1a–f. These pie charts show that even at a relatively low Z-score of 4.0, the three databases have a high percentage of agreement, especially at the fold level (78% of the FSSP matches at this Z-score are found in both SCOP and CATH). As Z-score increases from 4.0 to 6.0 and then to 8.0, agreement at both the fold and homology levels steadily increases. Beyond Z-score 8.0, this trend continues. Table 1 further subdivides the FSSP pairwise matches by sequence identity along with Z-score: this more clearly shows that agreement between databases increases both with Z-score and with sequence identity. For the most part, a combination of Z-score and sequence identity can be used to determine the likelihood of a structural pairwise match being found in all three databases (indicating a fairly undisputed match). As might be expected, a higher Z-score is needed at low sequence identities, and vice versa. At low sequence identities (0–19.9%), agreement between databases is rarely high. On the other hand, with a high enough sequence identity, Z-score is not necessarily very important for structural agreement between databases. Even at a Z-score below 4.0, sequence identity of 30% or greater results in complete agreement between the three databases.

It is not possible to state a Z-score within FSSP for which all three databases will completely agree on fold and/or homology for pairs of structures. This is certainly not what the authors intended; however, the vast majority of users may not have sufficient structural or biological knowledge to ascertain the importance of the data presented in an FSSP file. Determining Z-score cut-offs at which certain generalizations about the data could be made would facilitate the use of this resource.

A high percentage of agreement exists between the three databases at the fold level. Above 25% sequence identity, a threshold commonly used to distinguish between homologous or randomly related proteins, the agreement between the three databases is almost always 100% (the agreement between FSSP and SCOP is 100%). There are some exceptions, which represent 0.3% of the total number of FSSP pairs in the region included in Table 1 (which itself only represents 31% of the total FSSP matches). Between 25% and 29% sequence identity, five of the eight mismatches present are due to domain-assignment discrepancies, usually where only one portion of the protein is included in one of the databases, but another portion is classified in the other. The two mismatches between 30–34% are also due to this problem. Between 45–49%, five mismatches occur, all of which contain a phaseolin seed storage 7S protein domain paired with a canavalin 7S vicilin protein of the same family (according to SCOP classification, which agrees). The canavalin protein is less than half the size of the phaseolin, and is considered as one domain in CATH; the phaseolin protein is separated into four domains in CATH, and only two in SCOP.

At 100% sequence identity, only four mismatches contribute to the loss in agreement between the three databases. Two

Figure 1



Pie charts reflecting the agreement between pairwise matches in FSSP, CATH and SCOP. FSSP pairwise matches are compared to both CATH and SCOP: they are found in FSSP only (i.e. in neither SCOP nor CATH), in FSSP and SCOP (missed in CATH), in FSSP and CATH (missed in SCOP), or in all three databases. **(a,b)** FSSP pairwise matches (Z-score ≥ 4.0) compared to CATH and SCOP matches at the fold and homology level, respectively. Numbers in parentheses indicate the number of pairwise matches in question. At this Z-score, agreement between the three databases is already high at both the fold and homology level. **(c,d)** Pairwise matches (Z-score ≥ 6.0) compared to CATH and SCOP as before. Agreement between the databases has increased by at least 15% at both the fold and homology levels. The difference between FSSP + SCOP and FSSP + CATH agreement has also reduced. **(e,f)** Pairwise matches with Z-score ≥ 8.0. Already, agreement between the databases is as high as 97% at the fold level. Pairwise matches found in FSSP only are limited to three (see text for description), and the numbers of FSSP pairwise matches found in either SCOP or CATH (but not both) are very low.

of the mismatches pair two chains of restriction endonuclease *Bam*H1 (1bhm, chains A and B) with another *Bam*H1 structure (1bam0). CATH has the 1bam0 domain in a different three-layer (αβα) sandwich fold from the 1bhm domains; the former is classified in the collagenase (catalytic domain) fold, with the latter in the restriction endonuclease domain 2 fold. Both folds have the three-layer (αβα) sandwich architecture. Because of a different arrangement of helices, and the addition of small β strands in 1bam0 domain, CATH considers the geometry significantly different to assign a non-endonuclease fold, despite the function of the protein.

Another mismatch in the 100% sequence identity region concerns two thermolysin structures; because one is a fragment (1trlA), it is classed as a thermolysin fragment fold in CATH, whereas the other structure (1hyt) is classed as a neutral protease fold (which encompasses thermolysin). As they share obvious common ancestry, SCOP understandably considers both domains to have the same fold.

Among the mismatches above the 25% identity level, there is only one case where both SCOP and CATH disagree on the pairwise match: at 25% identity (now 26% in the current version of the FSSP database), the C-terminal domain of ribosomal protein L7/12 is aligned to the Taq DNA polymerase with a root mean square deviation (rmsd) of 3.9 Å. Both databases consider the ribosomal protein as one domain, but classify the polymerase as six (in CATH) or three (SCOP) domains. There is a small region within the large polymerase structure that resembles the small β-sheet structure of the ribosomal protein;

Table 1

Percentage agreement of FSSP pairwise matches with SCOP and/or CATH.

	Z-score of FSSP pairwise matches						
Percentage identity of FSSP pairwise matches	2–3.9	4–5.9	6–7.9	8–9.9	10–11.9	12–13.9	14–19.9
0–9	5.6	30.7	56.7	77.1	84.3	88.0	70.0
10–14	6.4	24.3	52.5	70.1	80.4	85.0	92.3
15–19	11.0	35.4	61.1	83.6	89.1	88.9	97.9
20–24	35.0	73.1	70.8	89.5	94.1	88.0	91.7
25–29	66.7	80.0	100	100	100	80.0	100

Percentage of FSSP pairwise matches (separated by Z-score and percentage identity) found in both SCOP and CATH; percentage agreement improves with increases in both Z-score and percentage identity. See text for additional discussion.

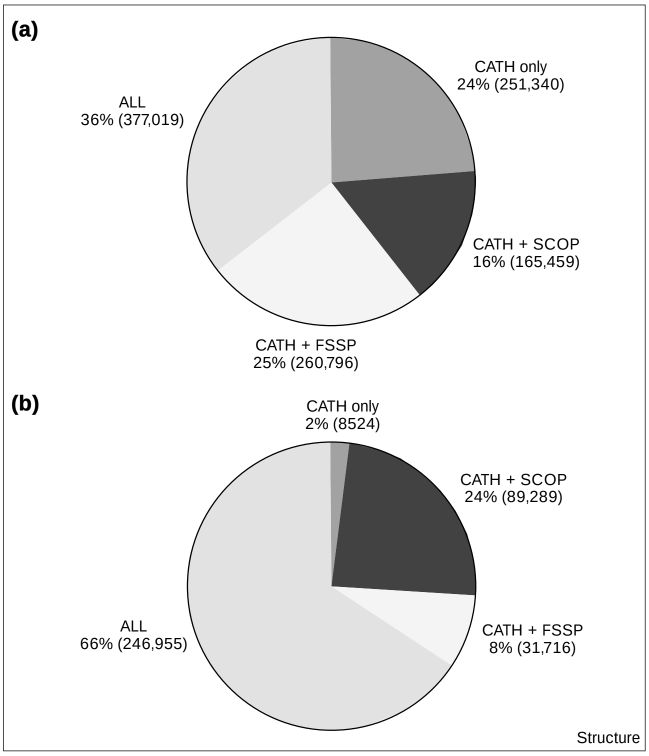
the structural-alignment program in FSSP would pick up this small similarity, while there may not be an evolutionary relationship between the two domains. This is one example of a match between two protein domains that is probably not indicative of common ancestry, but is simply a chance match between similar regions.

Comparing CATH pairwise matches to both SCOP and FSSP

At the fold level, the percentage of agreement with both SCOP and FSSP is small, at only 36% (see Figure 2). Alone, SCOP contains 51% of the CATH matches, and FSSP contains 60%. As compared to the SCOP comparison (see below), a much larger percentage of CATH matches is missed in SCOP or FSSP, but a higher percentage of matches is found in FSSP. The majority of the 49% of CATH matches missed in SCOP arise not from classification mistakes, but from problems such as domain assignment and fold overlap. Much of the 40% of CATH matches missed in FSSP may be due to changes in FSSP data (discussed below).

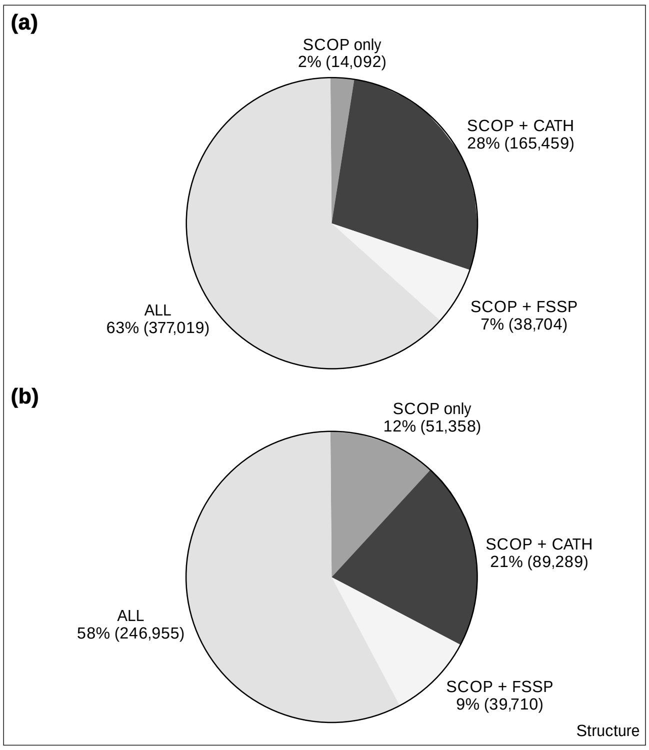
The most commonly occurring CATH pairwise matches missed by SCOP occur in the three-layer (αβα) sandwich Rossmann fold and the immunoglobulin fold. Most mismatches stem from the ‘fold-overlap’ problem, where a fold within CATH encompasses more than one fold within SCOP, and vice versa. When a domain is classified within CATH as being a three-layer (αβα) sandwich Rossmann fold, there are several SCOP folds to which it could conceivably belong. The same occurs with the immunoglobulin fold within CATH: several SCOP folds, such as the immunoglobulin-like β sandwich (SCOP code: 2.1), the prealbumin-like fold (2.3), and the cupredoxin fold (2.5) may contain these domains. Thus a domain classified as one SCOP fold will not be paired with a domain in another; although the structures are deemed by CATH to be geometrically similar, SCOP separates them to reflect an evolutionary or topological distinction.

Figure 2



Comparing CATH pairwise matches to SCOP and FSSP. **(a)** At the fold level, only 36% of the pairwise matches found in CATH are found in both SCOP and FSSP. Note that the number of CATH matches found in SCOP and FSSP is the same as the number of SCOP matches found in CATH and FSSP: the differing percentages reflect the total number of pairwise matches, which is much higher in CATH than in SCOP. A large percentage of these matches is found only in CATH. **(b)** A smaller number of pairwise matches are found at the homology level, so the overall agreement between the databases is higher, and the number of pairwise matches confined solely to CATH is lower. SCOP (and FSSP) still includes additional CATH matches that the other database does not.

Figure 3



Comparing SCOP pairwise matches to CATH and FSSP. (a) At the fold level, almost two-thirds of the SCOP pairwise matches are also found in both FSSP and CATH. CATH agrees with a further 28% of the SCOP matches, whereas FSSP includes only an extra 7%. Only a small percentage of the pairwise matches is unique to SCOP. (b) Fewer shared matches are found at the homology level in comparison to the fold level. Because of the difficulties inherent in assigning homology, there is a higher percentage of SCOP matches at this level that is not found in the other two databases.

The problems inherent with domain assignment also affect this comparison. Obviously, any protein separated into a different number of domains within SCOP and CATH will probably be classified into completely different folds as well. There are cases of proteins not completely classified by one of the databases: a group of MHC (major histocompatibility complex) class II proteins has only the N-terminal region included in SCOP, but both the N- and C-terminal regions are found in CATH. The C-terminal region is an immunoglobulin fold; each of these MHC proteins will thus be paired with every other immunoglobulin fold domain within CATH, but will be missed in SCOP. This is one example that illustrates the impact one discrepancy may have on the rest of the database; these four protein domains affect over 1000 pairwise matches within CATH.

Comparing SCOP pairwise matches to both CATH and FSSP

Approximately two-thirds of the SCOP matches at the fold level are found in both CATH and FSSP (Figure 3). Because SCOP has a much smaller number of pairwise matches than CATH, the percentage of matches agreed by all three databases is higher (64%) than in the CATH comparison. CATH alone agrees with over 90% of the SCOP matches; FSSP agrees with over 70%. Only 2% of the SCOP pairwise matches are absent in both CATH and FSSP. It is likely that the same problems discussed for the CATH comparison account for many of the mismatches and disagreements between the databases in this case. Of course, many of the comparisons (such as SCOP matches found in CATH, and CATH matches found in SCOP) are simply duplicated data.

At the homology level, fewer SCOP pairwise matches are found in both CATH and FSSP. Taken individually, both CATH and FSSP match fewer pairs than at the fold level (79% and 67% respectively). Understandably, the percentage of matches not found in either database has increased. These values reflect the difficulties in assigning homology between pairs of similar structures.

Differences and discrepancies between SCOP and CATH

Domain assignment

As mentioned previously, the separation of proteins into domains is a difficult and often subjective process. Table 2 highlights the difference between the methods of SCOP and CATH for distinguishing domains. Of the 6875 protein chains in pset3, 1194 (~17%) are assigned different numbers of domains in SCOP and CATH. In general, CATH assigns more domains than SCOP. This is due to the fact that CATH employs a purely structural definition for domains (essentially based on compactness), whereas SCOP takes into account whether or not a domain is observed as recurring in another superfamily, or observed as a separate single-domain fold. Occasionally, protein chains are classified as multidomain in SCOP until a more thorough classification on individual domains can be completed; for this reason, a chain could seem to consist of only one domain (i.e. as yet undivided) in SCOP while having several domains in CATH.

Problems with domains account for several groups of discrepancies between SCOP and CATH. An obvious domain problem is the exclusion of one part of a protein. In the case of the MHC class II chains (1iea(A–D)), only the N-terminal domain is included in SCOP. CATH includes both the N- and C-terminal domains, so any protein matching the C-terminal domain of 1iea(A–D) in CATH will not have an equivalent match in SCOP. The definition of domain obviously leaves some room for interpretation, and, in some cases, dividing a protein along a possible structural-domain boundary may in fact divide one active-site region into two or more nonfunctional segments. Such is the case with papain (1ppo). SCOP treats the protein as one domain, leaving the catalytic cysteine, histidine and asparagine together to form

Table 2

Comparison of domain assignment methods: SCOP and CATH.

	CATH (number of domains)						
SCOP (number of domains)	1	2	3	4	5	6	7
1	4475	817	51	31			
2	80	1007	61	104	14	3	
3		7	159	18		3	
4			3	38	2		
5					1		
6							
7							1

The number of domains into which each chain is separated in SCOP and CATH is compared. For the most part, the two classification schemes agree on the number of domains per chain (5681 of 6875 chains is ~82% agreement). However, in the case of chains split into two domains in CATH, almost half are considered as only one domain within SCOP. Examples and possible reasons for this are discussed in the text.

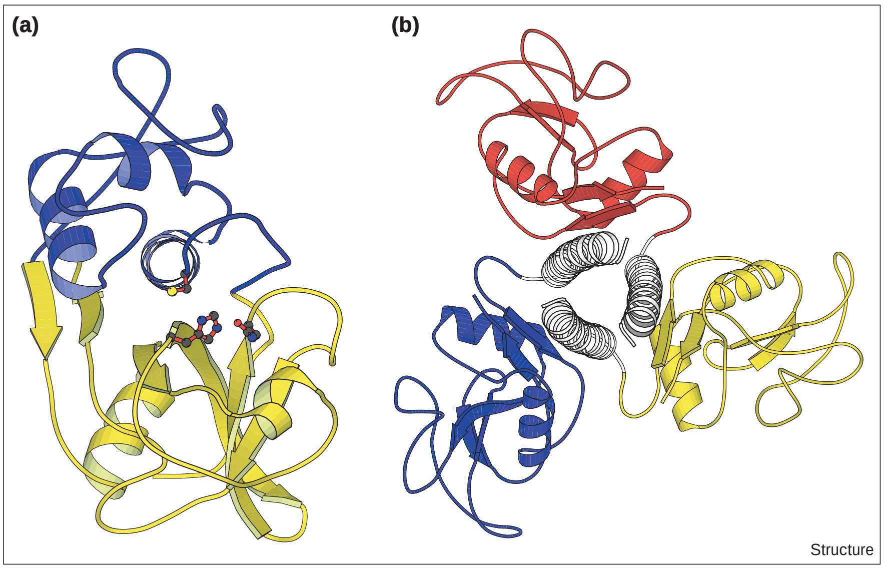
the active site. CATH, however, splits the protein into two domains, separating the cysteine from the asparagine and histidine, and rendering each domain effectively functionless (Figure 4a). CATH does the same for the trypsin-like serine peptidases and the aspartic acid pepsin peptidases. The decision is in many respects a philosophical one: whereas those interested in the biochemical aspects of protein structure may see the structure as a complete functional unit, others with interests in the dynamics of protein folding may argue that the functional unit can be separated into smaller, commonly occurring structural domains. Interestingly, the opposite occurs with mannose-binding protein A (1afb), where CATH categorizes each trimer subunit as a single domain, while SCOP separates the triple coiled-coil helix from the mainly β region of each monomer, and classifies the domains individually (Figure 4b).

Our exclusion of certain classes within this study has also generated some discrepancies between CATH and SCOP, as the larger number of SCOP classes makes it difficult to compare some proteins. For example, the C-terminal domain of the regulatory chain of aspartate carbamoyltransferase (1acmB/D:101–153) is classed as a small protein in SCOP (denoting structures usually dominated by metal ligand, heme, and/or disulphide bridges) (Figure 5a). However, CATH describes the same domain as being in the mainly β class. By removing the small protein class within SCOP, we are forced to ignore any pairwise matches containing this domain. Similarly, the haematopoetic cell kinase (hck) structure has one region classed as multidomain within SCOP (1ad5A/B:249–531), but approximately the same region is divided in CATH and presented as two domains, one in the αβ and one in the mainly α class (Figure 5b). When members of the multidomain class in SCOP are disregarded, corresponding pairwise matches are lost.

Class assignment

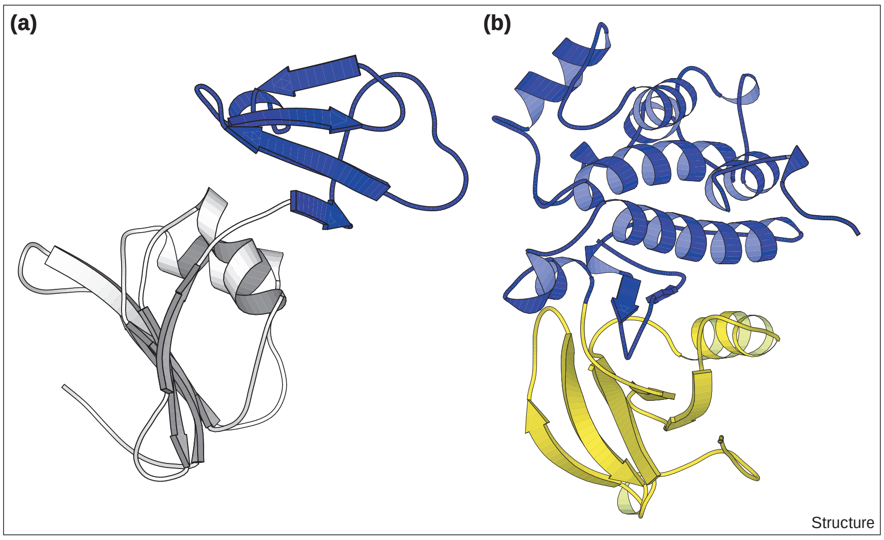
One category of discrepancy between SCOP and CATH arises from differences in class assignment, and the majority of disagreement arises from the presence of the two classes encompassing α/β domains in SCOP. However, domains within each class are allocated consistently in SCOP, and there are no cases of pairwise matches produced within both FSSP and CATH that SCOP has missed

Figure 4



Examples of domain-assignment disagreements between CATH and SCOP. (a) Structure of papain (1ppo) with catalytic histidine, asparagine and cystine shown as ball-and-stick residues. SCOP classifies the structure as one domain (SCOP code: 4.3.1), whereas CATH splits the structure into two, as shown by blue (CATH code: 1.10.190.10) and yellow (3.10.160.10) colouring. The cartoon figures were prepared using MOLSCRIPT [36]. (b) Structure of mannose-binding protein (1afb). CATH treats each monomer in the trimer as one domain (coloured red, blue and yellow), whereas SCOP separates the coiled-coil extension (uncoloured) from the rest of the structure, and classifies both domains individually.

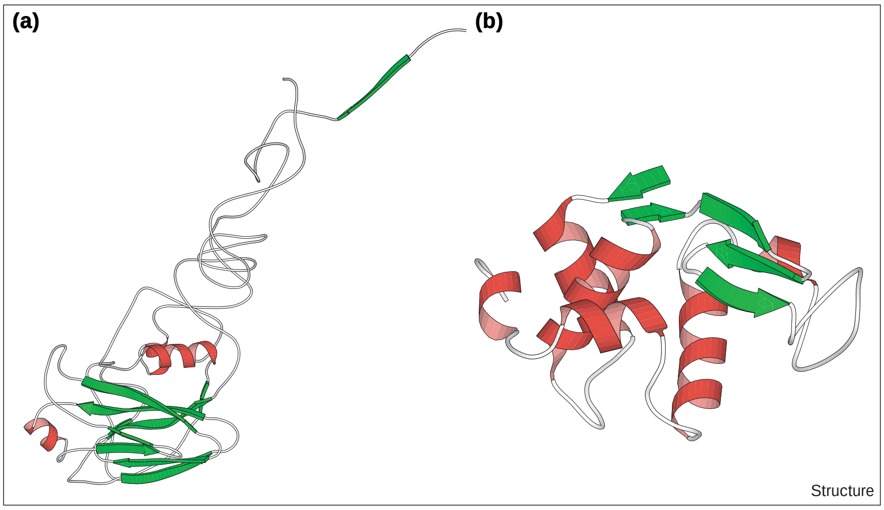
Figure 5



Problems associated with the definition of additional classes within SCOP. (a) The C-terminal domain of the regulatory chain of aspartate carbamoyltransferase (1acmB:101–153; coloured blue) is classed as a 'small protein' in SCOP, whereas CATH classifies it as a single-sheet mainly β structure of the monellin (subunit A) fold (2.20.30.60). (b) The haematopoetic cell kinase structure has one region classed as 'multidomain' within SCOP (1ad5A/B:249–531). Approximately the same region is presented as two domains in CATH: one domain is an αβ two-layer sandwich G4-amylase fold (1ad5A/B:259–344; 3.30.200.20; coloured blue), and the other is a mainly α non-bundle casein kinase I delta (subunit A, domain 2) fold (1ad5A/B:345–519; 1.10.510.10; coloured yellow).

because of a class mix-up (that is, defining one domain as class α+β, and one as class α/β). As with domain identification, class assignment can be dependent on subjective rules. An example is the haemagglutinin domain (1hgg, chains A, C and E), a domain which CATH considers to be in the αβ class because of the presence of two small helices amongst several β strands (Figure 6a). The SCOP authors ignore these small helical elements, as they are not consistently present across all available haemagglutinin structures and play no significant role in the function of the protein. The domain is thus classed as all β. The two methods are clearly relying on a different set of definition rules for classifying their entries: one may take small percentages of secondary structural elements into account, whereas the other disregards them. The situation is reversed with the case of the lysozyme superfamily: SCOP classifies these proteins as α+β, whereas CATH disregards the presence of small β strands, and opts instead to classify them as mainly α (Figure 6b). In this case, however, the evolutionary importance of these strands is a crucial factor in SCOP's determination of the overall structural class. Thus, for both SCOP and CATH the rules of classification are dependent on the protein family in question, and are not consistent throughout the classification database. Reassuringly, there are no cases where one scheme defines a mainly α domain that the other considers mainly β, or vice versa.

Figure 6



Examples of class assignment disagreements between CATH and SCOP. (a) SCOP ignores the small helical elements in the haemagglutinin structure (1hgg, chains A, C and E) and classifies the domain as mainly β, whereas CATH takes the helices into account and considers the structure αβ. (b) CATH disregards the presence of small β strands in the lysozyme superfamily (e.g. 1lys) and considers the protein mainly α, whereas SCOP takes into account the functional and evolutionary importance of these strands, and calls the lysozymes α/β.

Fold assignment

SCOP classifies pset3 into 286 separate folds, whereas CATH uses 447 folds to classify the same set. Of the total of 429 folds in SCOP, 323 exist in the major structural classes 1–4 (~75% of total). CATH defines a total of 590 folds, 527 of which are present in classes 1–3 (~89% of total). The definition of fold is thus somewhat arbitrary and left up to the creators of each classification method. As such, similar folds may have different names, or one method may encompass a subset of proteins under one general fold, whereas another method separates the set into more specific, less-populated folds. Surprisingly, most of the highly populated fold families in CATH are classified into more than one fold family in SCOP (the 'fold-overlap' issue mentioned previously). A good example of this is the Rossmann fold family (α/β class, three-layer αβα sandwich architecture: CATH no. 3.40.50). Proteins within this fold family in CATH are classified in several different fold families in SCOP, such as: the β subunit (capsid) of the lumazine synthase/riboflavin synthase complex (fold family 3.9; 1rvv, 30 chains), Flavodoxin-like (3.13; 1ofv0), NAD(P)-binding Rossmann fold domains (3.19; 1fmcA), N-carbamoylsarcosine amidohydrolase (3.22; 1nbaA), P-loop-containing nucleotide triphosphate hydrolases (3.25; 1ukz0), CheB methylesterase domain (C-terminal residues 152–349) (3.27; 1chd0), Subtilases (3.28; 1selA), Phosphotyrosine protein phosphatases I-like (3.31; 1phr0), anticodon-binding domain of Class II aminoacyl-tRNA synthetase (aaRS) (3.37; 1adyA), IIA domain of mannose transporter, IIA-Man (3.40; 1pdo0), phosphoglycerate mutase-like (3.43; 3pgm0), phosphoribosyltransferases (PRTases) (3.44; 1sto0), integrin A (or I) domain (3.45; 1lfaA), glycinamide ribonucleotide transformylase (3.46; 1cddA), S-adenosyl-L-methionine-dependent methyltransferases (3.47; 1vid0), and α/β-hydrolases (3.50; 1tib0). All these folds are described as three-layer αβα folds, with mostly parallel β sheets consisting of between four and eight strands. The SCOP authors have either made a topological distinction between the folds, or are more conservative in their fold assignment, choosing to keep folds separate until sufficient evidence warrants their unification. CATH on the other hand focuses on the geometric aspects of structural similarity, and thus encompasses all these SCOP folds into one large fold family, as they share the common Rossmann fold motif of a parallel β sheet flanked on both sides by α helices. Although the Rossmann fold typically describes structures with six-stranded β sheets, CATH has used this definition in a broader sense. Nevertheless, for domains in the CATH Rossmann fold family, the most commonly found SCOP classification for these folds is also the Rossmann fold. All the domains in this fold level in SCOP are found in the Rossmann fold level in CATH, and so a very high degree of consistency is apparent between the two schemes, even though SCOP defines a number of small subfamilies for these folds.

Rather more surprisingly, the opposite case also occurs, where domains from a fold within SCOP are classified into more than one fold in CATH. The TIM-barrel fold in SCOP (3.1) is one example. Corresponding CATH folds include transaldolase B, chain A fold (3.20.25; 1ucwA); urease, subunit C, domain 2 fold (3.20.50; 2kauC); and chitobiase, domain 3 fold (3.20.60; 1qba0). The three examples share the same architecture, but vary in the number of strands and helices comprising the barrel. CATH has separated each fold for geometric reasons, due to low structural similarity scores, whereas SCOP considers all these barrels to be sufficiently similar to group together. The frequent occurrence of this kind of fold definition discrepancy between SCOP and CATH is of course due to the effect of independent fold definition, and reflects the difference between using geometry and evolution to classify structure.

Favourable packing arrangements and protein architectures limit the number of possible protein folds, and for this reason large numbers of protein structures might be expected to fall into relatively small numbers of protein-fold families. Several estimations have been made in response to the question of how many folds exist in nature. Chothia originally estimated a conservative value for the total number of protein families of no more than 1000; thus the total number of folds would be even less [31]. This was followed by estimates varying from around 1000 folds [32] to over 6000 [33,34]. The issue is unresolved, with other estimates varying between around 650 [25] and less than 5200 for human proteins alone [24]. Clearly the issue is a controversial one. Given the arbitrary definition of what constitutes a protein fold, the only point that seems to be in agreement is that a finite number of naturally occurring folds exists (at the very least there can be no more folds than protein sequences). In the unlikely event that a standardization of fold definition and fold nomenclature can be agreed, then perhaps more agreement might be possible between the different estimates for fold numbers.

Homology assignment

Homology discrepancies can be seen clearly in cases where one database classifies two domains within one structure as homologous, but another database does not. Interestingly, in these cases a database may not only miss the homology between two domains, but may consider them to have different folds or architectures. Over 150 domains disagree in this way within SCOP and CATH. The most commonly occurring problems arise in the two largest fold groups (which are also superfolds) in these databases: the Rossmann fold and the immunoglobulin-like fold.

Several cases exist where CATH recognizes a homologous relationship between two domains within one structure that SCOP classifies as different fold. Elongation factor

Tu (EF-Tu) structures are split into three domains, two of which CATH considers homologous (2.40.80.10; 1eft0, domains 2 and 3); presumably the belief here is that these domains are the results of a distant gene-duplication event. These same domains are classed in separate folds in SCOP: the reductase/isomerase/elongation factor common domain fold (2.29) and the EF-Tu C-terminal domain fold (2.30). Both are closed Greek-key barrels, with six β strands, but SCOP has chosen to separate these folds while CATH combines them into a single fold family. A similar situation occurs in the 1cgt domain 4 family of the immunoglobulin-like fold (mainly β, sandwich: 2.60.40.110). CATH considers domains 3 and 4 of these proteins to be homologous. SCOP considers the former an immunoglobulin-like β sandwich (2.1) and the latter a prealbumin-like fold (2.3). Both folds are described within SCOP as being Greek keys with seven strands in two sheets, with additional strands in some members. The reason for the separation of these obviously similar folds is unclear, although it is presumably due to a higher apparent degree of calculated similarity within one subgroup.

The opposite case also occurs, where SCOP declares a homologous relationship with which CATH disagrees. The three domains in the A and B chains of the phosphoglucomutase (first three domains) family, superfamily and fold within SCOP have a mixed β sheet of four strands (e.g. rabbit phosphotransferase, 3pmgA/B). CATH classifies the three domains completely differently: all have the same three-layer (αβα) sandwich architecture, but each has a different fold. The first has the α-D-glucose-1,6-biphosphate, subunit A, domain 1 fold (3.40.460), the second has a Rossmann fold (nitrogenase molybdenum-iron protein, subunit A, domain 3) (3.40.50), and the third has the α-D-glucose-1,6-biphosphate subunit A domain 3 fold (3.40.120). The domain locations within each database are roughly the same. Oddly, although the description of the SCOP fold is a mixed β sheet of four strands, two of the domains have a different number of strands.

It is also common to find the databases agreeing on the fold of multiple domains, but disagreeing on homology. One example is the GMP synthetase subunit A domain 3 fold (αβ two-layer sandwich, 3.30.300) in CATH, which is subdivided into three homology levels. Most proteins with a domain from one homology level also contain domains from the other two homology levels in the GMP synthetase fold (1mxa domain 1 family, 3.30.300.20; 1mxa domain 2 family, 3.30.300.40). In SCOP, the three domains in these proteins are classified within the same family: the S-adenosylmethionine synthetase homologous family (fold of the same name, α+β class; 4.75.1). With proteins having one or more homologous domains, there are no cases of SCOP missing a homologous relationship that CATH identifies.

There are of course many more examples of missed or uncertain superfamily definitions in both SCOP and CATH, as the assignment of homology is often more subjective than the assignment of fold. Evolutionary relationships are often disputed or unclear, and different groups may make individual decisions as to domain relationships.

CATH architecture level

CATH defines another level of classification that SCOP does not consider, namely architecture. There are a small number of multidomain proteins that have an internal homology recognized by SCOP but are classed into different architectures (and thus folds) by CATH. These examples belong to the same SCOP homologous family: the actin-like ATPase domain (α/β, ribonuclease H-like motif fold; 3.41.1). The G chains of all five members of the glycerate kinase family have two domains in both SCOP and CATH, but SCOP classifies them as duplicated domains of three layers, whereas CATH classifies one as three-layer and the other as complex. A visual inspection of the domains using RASMOL [35] shows that the complex domain (1glaG:4–253) differs sufficiently from the standard three-layer (αβα) sandwich (see Figure 7a–c). CATH has the advantage of segregating such examples by using the architecture level of classification, whereas SCOP must incorporate these domains into fold groups that may include members with substantially different global structure.

Database updates

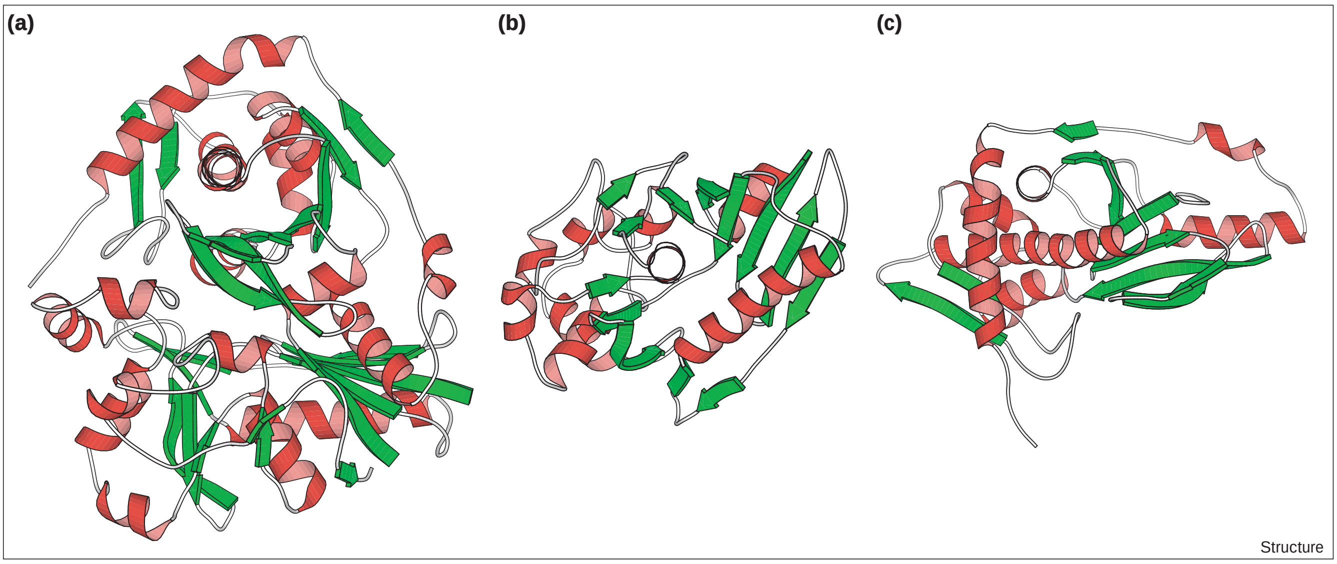
One possible drawback to using SCOP may become apparent when attempting to use the codes within this paper to access data from the current version of the database. The CATH authors anticipated the likely addition of data to each level of their hierarchy, and numbered each architecture, topology and homology level as a multiple of ten. As such, new entries to each level can either be added to the end of the database, or slotted in the middle of the current version by numbering between existing entries (i.e. the super-roll architecture is indexed as 3.15 to fit between the roll (3.10) and the barrel (3.20) in the αβ class). This ensures that current entries need not be changed. In contrast, the SCOP authors have apparently chosen to renumber entries upon the addition of new data to their database. So the flavodoxin-like fold, index number 3.13 in version 1.37 of the database, is now 3.14 in the current version of the database (v1.39); the α/β-hydrolase fold has gone from 3.50 to 3.56. This makes consistent use of the data more difficult, especially when considering the number of other resources that link to or cross-reference SCOP data.

Notable features of FSSP

Homology at low Z-score

Of the 21,637 pairwise matches within FSSP, 10,322 (almost 48%) have a corresponding Z-score below 4.0 (i.e. between 2.0 [the FSSP cut-off] and 3.9). A low Z-score

Figure 7



Escherichia coli glycerate kinase (1glaG), separated into two domains by both SCOP and CATH. SCOP considers the two domains to be homologous, classing them as members of the actin-like ATPase domain superfamily of the ribonuclease H-like motif fold. CATH assigns the domains different architectures. **(a)** Chain G in full. The interface between the two domains runs horizontally, across the middle of this diagram. **(b)** Chain G, domain 1: 4–253. This domain is classified as 'complex' within the αβ class in CATH. **(c)** Chain G, domain 2: 254–499. The domain is assigned the 'three-layer (αβα) sandwich' architecture within the αβ class in CATH.

(i.e. less than 4.0) does not necessarily rule out a structural similarity or even homology between two structures. A small percentage of FSSP pairwise matches with Z-scores less than 4.0 is found in both SCOP and CATH fold families or even within superfamilies. Understandably, a much smaller number of agreements is seen at the homology level than at the fold level: only 166 FSSP pairwise matches exist at the SCOP and CATH homology level, as compared to 673 at the fold level. At the fold level, the immunoglobulin-like β sandwich is the predominant fold, involving almost one-third of the matches. The TIM-barrel fold, Rossmann fold (three-layer αβα sandwich) and arc repressor mutant fold (DNA-binding three-helical bundle in SCOP) also recur frequently. Although the matches at this level are only a small percentage of the possible matches within each of these folds, this is still an indication that some folds are more easily matched than others. These folds would seem to present a particular challenge to the Dali comparison method.

A Z-score below 4.0 might be considered insignificant for assigning an evolutionary relationship between two protein structures, hence the advantage of taking sequence identity into account. However, most of these FSSP pairwise matches (10,233 of 10,322) have a sequence identity lower than 20%, a value commonly considered a threshold for assigning obvious homologous relationships. Thus, the majority of the pairwise comparisons within FSSP are presumably not indicative of definite evolutionary relationships. However, the inclusion of this information in FSSP is useful in that Z-score and sequence identity may be used to automatically identify very remote relationships between protein structures, and the relationships between structures in a neighbourhood (rather than a hierarchy) can be closely examined. Users are free to interpret this data as they wish, without any preformed decisions being made on the significance of the information.

Obviously, a clear picture cannot be derived from considering the Z-scores alone. The sequence identity between two structures presents additional information for assessing similarity of proteins, particularly regarding the possibility that the structures are evolutionarily related. Within the subset of FSSP pairwise matches with a Z-score below 4.0, sequence identity varies from below 10% to over 80%. As sequence identity relates only to the sequence region being aligned, this value may not necessarily reflect the global similarity between two proteins. For example, in the case of a pairwise match between two calmodulin chains with 88% identity, the alignment length is only 57 residues over two sequences of 148 residues (*Xenopus laevis* calmodulin [1dmo0] and *Paramecium tetraurelia* calmodulin [1osa0]). Although these two calcium-binding domains are obviously homologous, they superimpose with an rmsd of 11.3 Å, producing a Z-score of 3.6. This is presumably due to the calcium-induced conformational change in one of the chains. Without taking the degree of sequence identity into account, these low rmsd and Z-score values are not sufficient to indicate a homologous relationship. Few of the corresponding sequence identities are as high as the value in this example, but other values indicate that homologous structures may not necessarily superimpose well enough to produce a high Z-score.

High Z-score without homology

A high Z-score (i.e. greater than 6.0) does not necessarily indicate a structural similarity between two proteins that agrees with SCOP and CATH. At the fold level, there are 69 pairwise matches within FSSP above Z-score 6.0 that are not recognized in SCOP or CATH. The Z-scores range from 6.0 to 15.6, with corresponding sequence identities ranging from 4% to 24%. Grouping these examples gives an indication of the reasons that might have led SCOP and/or CATH to classify these structures as having no fold similarity or homology. A large proportion of these pairwise matches involves α and β domains (classed as α/β or α+β in SCOP) classified as different folds in SCOP, and different architectures in CATH. The folds vary in SCOP, but within CATH the architectures are largely limited to the three-layer (αβα) sandwich (3.40), the three-layer (ββα) sandwich (3.50) and the complex architecture (3.90). The complex architecture contains αβ proteins too elaborate to fit into any other CATH architecture, with some examples containing combinations of helices and strands that resemble portions of a typical three-layer sandwich. Small similarities like these explain the low sequence identity and high Z-score between some structural pairs; small subdomains or regions may superimpose well, despite the domains having no overall similarity or obvious evolutionary relationship.

Not surprisingly, at higher Z-score cut-offs, the proportion of FSSP pairwise matches absent at both the fold and homology levels in SCOP and CATH decreases steadily. The number of homology mismatches is always larger than the number of fold mismatches, and the relationship between the two values decreases by almost 50% with each unit increase in Z-score. At a Z-score of 9.0, only one mismatch remains at the fold level, where FSSP has paired a G-protein transducin (1tbgA) and a methanol dehydrogenase (4aahA) with a sequence identity of 8%, rmsd of 4.0 Å and a Z-score of 15.6. SCOP and CATH both classify the methanol dehydrogenase structure as an eight-bladed propeller fold (CATH indicates this at the architecture level), but disagree on the transducin: CATH classifies it as a six-bladed propeller architecture, whereas SCOP considers it a seven-bladed fold. A closer look at the structure reveals that the transducin does have seven blades, each consisting of four short β strands. This discrepancy does not affect the pairwise match in question: the two domains are propellers with a different number of blades; FSSP has apparently superimposed portions of them with a reasonable rmsd and Z-score, although no actual relationship is evident. The six-bladed propeller architecture in CATH has only one topology level, the neuraminidase fold (2.120.10). This fold encompasses two homologous groups, one of whose members has structures consisting of six blades (1mwe; 2.120.10.10) and the other whose members contain seven blades (1gotB; 2.120.10.20). As a result, any pairwise matches between these two homologous families (i.e. when matching at the fold level) will be inaccurate. In addition, CATH misses the relationships between this group and other seven-bladed propeller structures, such as domains within the methylamine dehydrogenase chain H (2.130.10) and galactose oxidase domain 2 (2.130.20) folds. After investigation of these errors within CATH, it transpires that they are the results of a single typographical error during the initial manual definition of the architecture (CA Orengo, personal communication). This example demonstrates the impact that one small human error can make on a hierarchical database of protein structures.

It is well known that low sequence identity between two proteins does not necessarily indicate the absence of an evolutionary relationship. There are examples of pairwise matches with low sequence identity in FSSP that are classified in the same fold or homologous family in SCOP and CATH. Sequence identity should therefore not be taken alone as a criterion of homology between two structures: the same fold can often be shared by a variety of different proteins that share virtually no sequence similarity. Sometimes a fold may be the ideal structure for several proteins with a range of function and ancestry. Some of the most interesting cases are pairs of protein domains with low sequence identity, high rmsd (indicating a suboptimal superposition of structures) and low Z-score (indicating the match may not be significant). At the homology level, 977 pairwise matches below 20% identity are shared between the three databases. (At the fold level this value jumps to 2716 pairwise matches.) Although this subset of domains represents a mere fraction of the 14,807 FSSP pairwise matches below 20% identity, they illustrate the dangers in depending on sequence identity to provide an accurate picture of structural relationships. A high sequence similarity between two proteins (assuming it covers a reasonable length of the sequences in question) usually indicates a homologous (and therefore structural) relationship, whereas a low sequence identity cannot be used to rule out the opposite.

Database updates

Under certain circumstances, one of the key advantages of FSSP, namely the automatic update procedure, may sometimes cause difficulties. Unlike CATH and SCOP, FSSP is updated continuously, with data derived from the Dali alignment program generating new and revised FSSP files automatically. Because data is taken directly from the PDB, which generally releases new structural information according to a weekly schedule, FSSP is constantly changing. Each version of CATH or SCOP is guaranteed to be relatively unchanged until the next major update, and archive copies of previous releases are available from the maintainers. Occasionally, new PDB codes are chosen to supersede original codes; any pairwise matches with obsolete codes would obviously be missed. Because the organization of FSSP depends on a representative set of domains and a set of sequence homologues, the addition of new structures might trigger a reorganization of the domain groups, with the possibility of a new representative being chosen. This would in turn affect the structural alignments, which would then influence the sequence identity, rmsd, and Z-score of each pairwise match. Thus, both the constituent protein domains, and the information generated by their alignments, are changing constantly in FSSP.

In a comparison of FSSP files available in November 1997 and July 1998, approximately 12% of the FSSP pairwise matches used in this analysis were absent. In the remaining pairwise matches, less than 0.05% of Z-scores, percentage identities and rmsd scores had changed, but changes included Z-score values that unfortunately crossed the arbitrary threshold of 4.0 commonly used to assign structural relationships. Measures were taken to minimize the impact these problems would have on the data presented here, but it is likely that some of the pairwise matches missed in FSSP are due in part to the continual updating of data.

Biological implications

The reliability and accuracy of structure classification methods are important to structural and non-structural biologists alike. Protein structure data is used in various aspects of biology such as benchmarking, protein modelling, evolutionary studies and drug design. This systematic comparison of SCOP, CATH and FSSP represents the first attempt at estimating the degree of consistency between these databases, and facilitates a comparison between fully and partially automated, and primarily manual, classification methods.

To a large extent, the three databases agree on classifications; certainly no one method is distinctly superior. Most of the differences and discrepancies that exist result from the unique guidelines by which structures are classified within each database. Biologists should note that there are no fixed principles of protein structure classification, and each method relies on independently devised rules.

Understanding these rules is crucial to making the most of each resource, as is the database structure (i.e. as a hierarchy or structural neighbourhood) and the way separate families are treated (i.e. whether small secondary structure elements are included or disregarded when assigning class, etc.). SCOP is a valuable resource for detailed evolutionary information, but its purely manual derivation influences update frequency and means some families or folds within the database may not be as exhaustively detailed as others. CATH provides useful geometric information, and the addition of ‘architecture’ can reveal broad features of protein-fold shape, but partial automation means examples near fixed thresholds may be assigned inaccurately. FSSP is continually updated and presents data for the user’s own assessment; however without sufficient knowledge, a user may not assess this data appropriately.

By presenting such a large amount of structural data with detailed geometric and evolutionary information, these databases are a valuable resource for benchmarking of methods, and structural studies. At present, using these databases in conjunction with human judgement and biological knowledge should be sufficient for providing accurate and reliable structural information to all biologists. Whether a consensus database, devised by extracting undisputed protein classifications from SCOP, CATH and FSSP, would improve the development of accurate threading templates is currently being assessed.

Materials and methods

Generating a set of shared structures

In order to compare the three databases, a standard list of common structural identifiers was first generated. Unlike FSSP, both SCOP and CATH append the four-character PDB code with chain and domain identifiers (e.g. 1pdbC1 where ‘C’ identifies the chain, and ‘1’ indicates the first domain). If no chains or domains exist, an ‘0’ is used. In SCOP (March 1998; v1.37), only classes 1–4 were considered (mainly α, mainly β, α/β, $\alpha+\beta$); in CATH (April 1998, v1.4), only classes 1–3 (mainly α, mainly β, α/β) were included. In comparing the codes found in each database, only chains (e.g. 1pdbA, 1pdbB and 1pdbC) were considered, as domains are not consistently allocated across different classification schemes. An ‘0’ was added to all FSSP four-character codes, as FSSP contains some structures of only one chain. An additional ‘0’ was added to all five-character codes, as FSSP does not separate any protein chains into domains. Once a list of shared five-character codes (i.e. protein chains) was created from the three databases, the corresponding domains for each protein structure chain were then included for SCOP and CATH. Each full protein structure code in SCOP and CATH corresponds to a classification index number used to define its class, fold, superfamily and family.

Structure comparisons within each database

An all-against-all comparison of the classification numbers of the six-character codes (i.e. protein domains) within the master list (pset3) produced a set of pairwise matches within both SCOP and CATH. FSSP files consist of pairwise matches with Z-scores above 2.0 between protein structures in the representative set and the sequence-homologue set, and these matches were extracted along with the corresponding Z-score and percentage sequence identity. Additionally, for CATH and SCOP, matches were determined at both the homology or fold levels. In CATH, homology (i.e. homologous superfamily) is found at the fourth place in the numbering system, with fold, or topology, at the third position, but in SCOP, fold holds the second place in the index, with homology (or superfamily) in the third place. No more than one match per pair of codes (with chain identifiers) was recorded.

Structure comparisons between databases

Upon generating a set of pairwise matches for each database, comparisons were made between the three databases. Each pairwise match was noted as being present in either both, neither, or one of the other remaining two databases. As FSSP is structured differently from both SCOP and CATH, an additional category of ‘incompatible’ was created to include SCOP or CATH matches comprising two codes found in either the representative FSSP set or the sequence homology set, but not both. A large number of these incompatible pairwise matches could be converted into compatible matches by substituting each PDB code in the pair with its representative code within FSSP.

Accessing the databases

The URL for the SCOP database is http://scop.mrc-lmb.cam.ac.uk, CATH is http://www.biochem.ucl.ac.uk/bsm/cath/, and FSSP is http://www2.embl-ebi.ac.uk/dali/fssp. An interactive website with data from this analysis can be found at http://globin.bio.warwick.ac.uk/~hadley/db.

Acknowledgements

This work was supported by Zeneca Pharmaceuticals and the Royal Society. We thank Dr David Timms and the authors of SCOP, CATH and FSSP for valuable comments during the course of this work.

References

1. 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.
2. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 34, 167-339.
3. Crippen, G.M. (1978). The tree structural organization of proteins. *J. Mol. Biol.* 126, 315-332.
4. Rose, G.D. (1985). Automatic recognition of domains in globular proteins. *Methods Enzymol.* 115, 430-440.
5. Wodak, S.J. & Janin, J. (1981). Location of structural domains in proteins. *Biochemistry* 20, 6544-6552.
6. Holm, L. & Sander, C. (1994). Parser for protein folding units. *Proteins* 19, 256-268.
7. Swindells, M.B. (1995). A procedure for detecting structural domains in proteins. *Protein Sci.* 4, 103-112.
8. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552-558.
9. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
10. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure* 5, 1093-1108.
11. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* 1, 1691-1698.
12. Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C.A. & Thornton, J.M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.* 7, 233-242.
13. 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.
14. Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol.* 208, 1–22.
15. Orengo, C.A., Brown, N.P. & Taylor, W.R. (1992). Fast structure alignment for protein databank searching. *Proteins* 14, 139-167.
16. Holm, L. & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* 22, 3600-3609.
17. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123-138.
18. Martin, A.C., et al., & Thornton, J. M. (1998). Protein folds and functions. *Structure* 6, 875-884.
19. Abagyan, R.A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* 273, 355-368.
20. Russell, R.B., Saqi, M.A., Bates, P.A., Sayle, R.A. & Sternberg, M.J.E. (1998). Recognition of analogous and homologous protein folds–assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng.* 11, 1-9.
21. Murzin, A.G. & Bateman, A. (1997). Distant homology recognition using structural classification of proteins. *Proteins* 1, 105-112.
22. Chou, K.C. & Maggiora, G.M. (1998). Domain structural class prediction. *Protein Eng.* 11, 523-538.
23. Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* 7, 445-456.
24. Zhang, C.T. (1997). Relations of the numbers of protein sequences, families and folds. *Protein Eng.* 10, 757-761.
25. Wang, Z.X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* 11, 621-626.
26. Madej, T., Gibrat, J-F. & Bryant, S.H. (1995). Threading a database of protein cores. *Proteins* 23, 356-369.
27. Gibrat, J-F., Madej, T. & Bryant, S.H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* 6, 377-385.
28. Mizuguchi, K., Deane, C.M., Blundell, T.L. & Overington, J.P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* 7, 2469-2471.
29. 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.
30. 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.
31. Chothia, C. (1992). Proteins: One thousand families for the molecular biologist. *Nature* 357, 543-544.
32. Blundell, T.L. & Johnson, M.S. (1993). Catching a common fold. *Protein Sci.* 2, 877-883.
33. Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994). Protein superfamilies and domain superfolds. *Nature* 372, 631-634.
34. Alexandrov, N.N. & Go, N. (1994). Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci.* 3, 866-875.
35. Sayle, R.A. & Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*. 20, 374.
36. Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24, 946-950.

Because *Structure with Folding & Design* operates a ‘Continuous Publication System’ for Research Papers, this paper has been published on the internet before being printed (accessed from http://biomednet.com/cbiology/str). For further information, see the explanation on the contents page.