Acta Cryst. (1998). D**54**, 1147–1154

SCOP, Structural Classification of Proteins Database: Applications to Evaluation of the Effectiveness of Sequence Alignment Methods and Statistics of Protein Structural Data

Tim J. P. Hubbard,[a]* Bart Ailey,[b] Steven E. Brenner,[c] Alexey G. Murzin[b] and Cyrus Chothia[d]

[a]*Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, England,* [b]*Centre for Protein Engineering, MRC Centre, Hills Road, Cambridge CB2 2QH, England,* [c]*Department of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA, and* [d]*Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH, England. E-mail: th@sanger.ac.uk*

(*Received 20 April 1998; accepted 6 July 1998*)

Abstract

The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of all known protein structures. The classification is on hierarchical levels: the first two levels, family and superfamily, describe near and far evolutionary relationships; the third, fold, describes geometrical relationships. The distinction between evolutionary relationships and those that arise from the physics and chemistry of proteins is a feature that is unique to this database, so far. The database can be used as a source of data to calibrate sequence search algorithms and for the generation of population statistics on protein structures. The database and its associated files are freely accessible from a number of WWW sites mirrored from URL http://scop.mrc-lmb.cam.ac.uk/scop/.

1. Introduction

At present (April 1998) the Brookhaven Protein Data Bank (PDB, Abola *et al.*, 1987) contains 7435 entries and the number is increasing by about 200 a month. These proteins have structural similarities with other proteins and, in many cases, share a common evolutionary origin. To facilitate access to this information, we have constructed the Structural Classification of Proteins (SCOP) database (Murzin *et al.*, 1995). It includes not only all proteins in the current version of the PDB, but many proteins for which there are published descriptions but whose coordinates are not yet available.

The classification of proteins in SCOP has been constructed by visual inspection and comparison of structures. Given the current limitations of purely automatic procedures, we believe this approach produces the most accurate and useful results. The unit of classification is usually the protein domain. Small proteins, and most of those of medium size, have a single domain and are, therefore, treated as a whole. The domains in large proteins are usually classified individually.

The classification of the proteins is on hierarchical levels.

1.1. *Family*

Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have significant sequence similarity; second, proteins with lower sequence identities; but whose functions and structures are very similar; for example, globins with sequence identities of 15%.

1.2. *Superfamily*

Families, whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, the variable and constant domains of immunoglobulins.

1.3. *Common fold*

Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in same arrangement and with the same topological connections (for recent reviews see Orengo, 1994; Murzin, 1994). The structural similarities of proteins in the same fold category, probably arise from the physics and chemistry of proteins favouring certain packing arrangements and chain topologies.

1.4. *Class*

The different folds have been grouped into classes. Most of the folds are assigned to one of the five structural classes.

(i) All-α, those whose structure is essentially formed by α-helices;

(ii) all-β, those whose structure is essentially formed by β-sheets;

(iii) α/β, those with α-helices and β-strands;

ISSN 0907-4449

(iv) $\alpha+\beta$, those in which α-helices and β-strands are largely segregated; and

(v) multi-domain, those with domains of different fold and for which no homologues are known at present.

Other classes have been assigned for peptides, small proteins, theoretical models, nucleic acids and carbohydrates. These hierarchical levels are illustrated in Fig. 1.

There are now a number of other databases which classify protein structures, such as CATH (Orengo *et al.*, 1993, 1997), FSSP (Holm & Sander, 1994, 1996), Entrez (Hogue *et al.*, 1996) and DDBASE (Sowdhamini *et al.*, 1996), however the distinction between evolutionary relationships and those that arise from the physics and chemistry of proteins is a feature that is unique to SCOP, so far. Because functional similarity is implied by an evolutionary relationship but not necessarily by a physical relationship, we believe that this classification level is of considerable value, for example as a way of linking very distant sequence families reliably.

2. Steps used to classify proteins in SCOP

The following description outlines the major steps in the classification of protein structures at the different levels listed above.

Computational methods are used to aid the classification process, however the information they provide is incomplete and so final decisions in all cases are the result of manual inspection. For example, sequence comparison is used to automatically detect relationships between parts of new structures and domains already classified, however it fails to identify many of the structural relationships in SCOP either because the sequence relationship has become too weak (for evolutionarily related proteins) or never existed (for evolutionarily unrelated proteins with similar folds). Structure–structure comparison programs can identify domains of similar structure, however manual inspection is required to verify the choice of fold as frequently several similar but distinct folds are identified. The assignment of proteins of known structure to evolutionarily related superfamilies is perhaps the single most powerful and important feature of the database, but is the one most reliant on the manual procedures described below as current computational methods are almost entirely unhelpful in this regard.

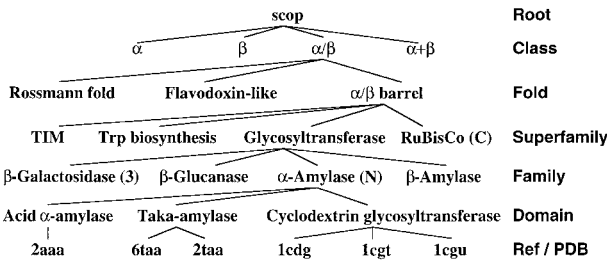

Fig. 1. Region of SCOP hierarchy. All the major levels, including class, fold, superfamily, and family are shown. Also shown are individual proteins and the lowest level either the PDB coordinate identifier or a literature reference. Copyright 1994, Steven E. Brenner; reproduced with permission.

2.1. *Domain and class*

The first step in the classification of a protein is to divide it, where necessary, into domains. The basic idea of a domain is a region of a protein which has its own hydrophobic core and has relatively little interaction with the rest, so that it is essentially structurally independent. Identification of domains is not trivial and can frequently be performed correctly only by using evolutionary information to see, for example, how domains have been 'shuffled' in different proteins. Typically domains are collinear in sequence, but occasionally one domain will have another 'inserted' into it, or two homologous domains will intertwine by swapping some topologically equivalent parts of their chains.

Where domains can be identified (which in many cases will be the entire protein chain) these are placed in classes based on whether their cores consist exclusively of α-helices, β-sheets, or some mixture. In some borderline cases a domain could be argued to fit equally well in more than one class, so for this reason class should be regarded as mainly for convenience of browsing and not always an unambiguous definition.

Because of the problem of identifying domains on the basis of a single protein structure there is a multi-domain class. Proteins here have multiple domains which have never been seen independently of each other, so accurate determination of their boundaries is not possible and perhaps not meaningful or useful. This is seen as a transient class, as proteins found here will be classified elsewhere as soon as evidence for their domain boundaries emerges.

There are also classes for proteins and domains which are not globular, soluble structures stabilized by the packing of α-helices and β-sheets. These are 'small proteins', for those proteins which structure stabilized by disulfide bridges or by metal ligands rather than by hydrophobic core; 'membrane proteins'; 'short peptides'; 'theoretical models'; and 'non-proteins', for entries in the Protein Data Bank such as nucleic acids.

2.2. *Folds*

Structural–structure similarity programs such as DALI, available *via* a World Wide Web (WWW) server (Holm & Sander, 1995), allow similarities to be identified in many cases, however interpreting the results is not always straightforward. There are now many proteins with similar, but distinct folds and topological similarity may not be sufficient. The approach used for SCOP to characterize a fold is to look first at the major

architectural features, and then identify more subtle characteristics. Where folds appear similar, but protein structures do not superimpose well, these proteins cannot be classified as having the same fold or superfamily. Topological similarities of this kind are on an intermediate level between class and fold, and, in the current version of SCOP, they are silently indicated by listing folds with similar topologies together on the class page. This approach is also used to segregate different architectural motifs, like two-sheet sandwiches and single-sheet barrels in the all-β class. Future versions of SCOP will include the necessary additional levels of classification to make such distinctions explicit.

2.3. *Superfamilies*

Protein structures classified in the same superfamily are probably related evolutionarily and, therefore, they must share a common fold and usually perform similar functions. If the functional relationship is sufficiently

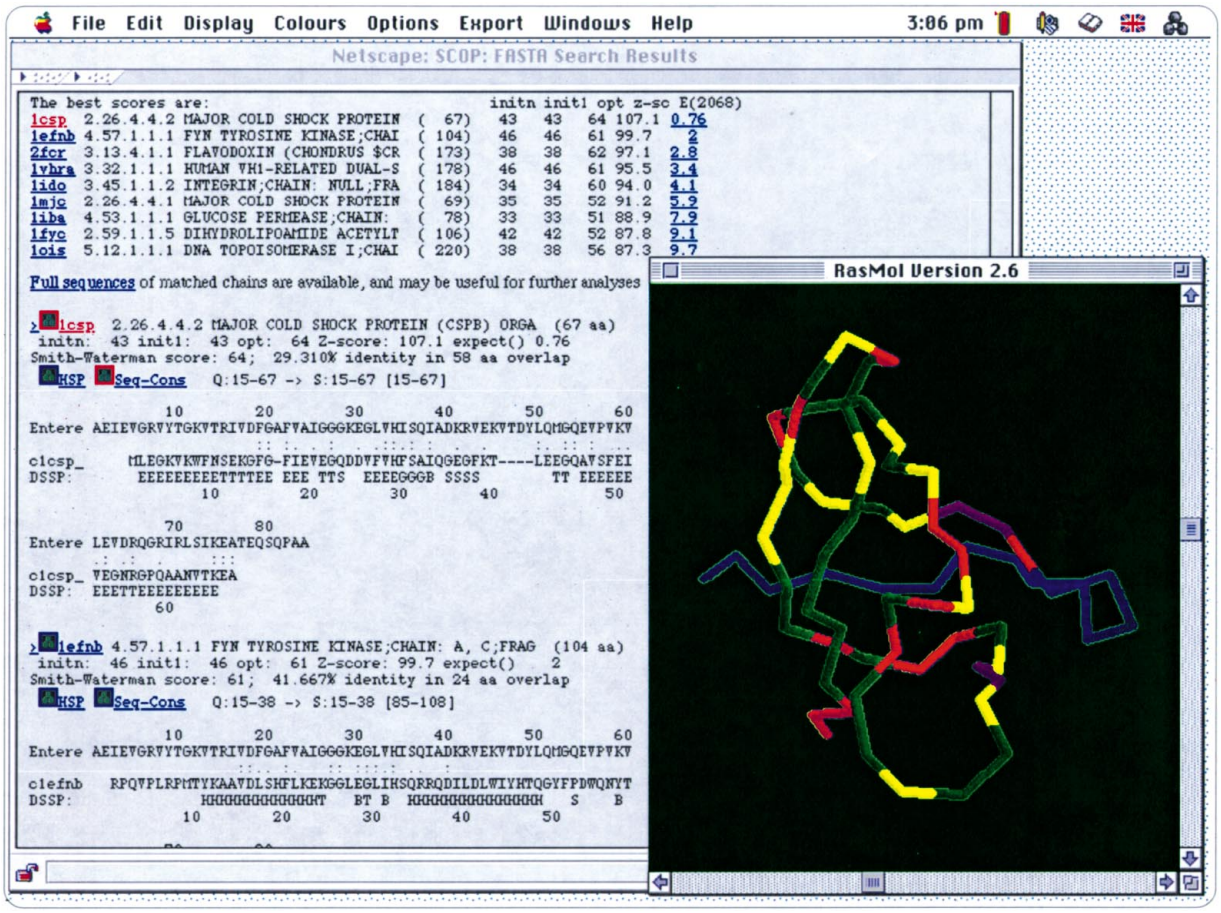

Fig. 2. A example of the use of the SCOP sequence similarity search facility is shown on a Macintosh workstation. The PDB90 database is searched using *FASTA* (Pearson, 1996) with the sequence of the PDB entry 1SRO (S1 RNA-binding domain of polyribonucleotide phosphorylase, PNP), which in 1996 was unpublished and target T0004 in the CASP2 structure prediction experiment (Moult *et al.*, 1997) and is used here to illustrate the utility of the search facility in SCOP in looking for distant relationships. Because the headers of the PDB90 file contain a SCOP classification code (a.b.c.d.e) it is immediately obvious when several sequences from the same superfamily or fold are in the list. In this list (the self hit to 1SRO has been removed) none of the matches have a significant score [the *E* value must be <0.01 for 99% confidence (Brenner *et al.*, 1998)], however a match to the superfamily 2.26.4 (2 = β class; 26 = OB -fold; 4 = nucleic acid-binding proteins superfamily) is found twice, and is the only one. This is indeed the correct fold for 1SRO and further investigation of this promising lead might well result in many users coming to this conclusion.As well as the page being linked to the SCOP classification, on a correctly configured workstation (see below) clicking on the green icons results in a structure that the sequence match is to being automatically loaded into the molecular viewer program *RasMol* [written by Roger Sayle (Sayle & Milner-White, 1995)] with the sequence of the unknown mapped onto it according to the alignment. The view shown is for one CSP when the button next to the 'Seq-Cons' was clicked. The colouring scheme is: red for identical residues; yellow for similar residues (+ in *BLAST* alignments); green for dissimilar residues; blue for non aligned parts of the chain. From this it can be seen that the majority of the structure is matched and that there are clusters of conserved residues in the core of the β-barrel. This 'instant' homology modelling can be a useful way to discriminate interactive between likely and unlikely matches. Throughout SCOP the green icons are used to display protein structures with classification features highlighted. Information and software to configure a workstation to enable this visualization facility are available for download from the SCOP URL.

Table 1. *Facilities and databases to which SCOP has links*

The SCOP database contains links to a number of other facilities and databases in the world. Several interactive viewers can be linked with SCOP using PDB coordinates. The location and nature of the links will vary as databases evolve and relocate.

Link	Source	URL	Reference
Coordinates	PDB	http://www.pdb.bnl.gov/	Abola *et al.* (1987)
Static Images	SP3D	http://expasy.hcuge.ch	Appel *et al.* (1994)
On-the-fly images	NIH molecular modelling group	http://www.nih.gov/www94/molrus	Fitzgerald (1994)
Sequences and MEDLINE entries	NCBI Entrez	http://www.ncbi.nlm.nih.gov/	Benson *et al.* (1993)
Protein Motions Database	Mark Gerstein	http://bioinfo.mbb.yale.edu/MolMovDB	Gerstein *et al.* (1994)
Nucleic Acids Database	Rutgers University	http://ndbdev.rutgers.edu/	Berman *et al.* (1992)

strong, for example, the conserved interaction with substrate or cofactor molecules, the shared fold can be relatively small, provided it includes the active site (for example, Bycroft *et al.*, 1997). It is in contrast with classification on the fold level, which ordinarily requires greater structural similarity.

3. Organization and facilities of SCOP

The SCOP database is available as a set of tightly coupled hypertext pages on the WWW *via* URL http://scop.mrc-lmb.cam.ac.uk/scop/. The interface to SCOP has been designed to facilitate both detailed searching of particular families and browsing of the whole database. To this end, there are a variety of different techniques for navigation.

3.1. *Browsing through the SCOP hierarchy*

SCOP is organized as a tree structure. Entering at the top of the hierarchy the user can navigate through the levels of class, fold, superfamily, family and species to the leaves of the tree which are structural domains of individual PDB entries. An alternative hierarchy of folds, superfamilies and families by the date of solution of the first representative structure is also provided.

3.2. *From an amino-acid sequence*

The sequence similarity search facility allows any sequence of interest to be searched against databases of protein sequences classified in SCOP (see below) using algorithms *BLAST* (Altschul *et al.*, 1990), *FASTA* or *SSEARCH* (Pearson, 1996). SCOP can then be entered from the list of PDB chains found to be similar and the similarity can be displayed visually (see Fig. 2).

3.3. *From a keyword*

The keyword search facility returns a list of SCOP pages containing the word entered or combinations of words separated by a series of Boolean operators.

3.4. *From a PDB identifier*

The PDB entry viewer links PDB entries to various graphical views, external databases and SCOP itself.

3.5. *By history*

Pages are provided that order folds, superfamilies and families by date of entry into PDB or publication. This is both for interest and to make it easier to keep up to date with the appearance of new folds or significant new members of existing folds. In addition to the information on structural and evolutionary relationships contained within SCOP, each entry (for which coordinates are available) has links to images of the structure, interactive molecular viewers (Fig. 2), the atomic coordinates, data on functional conformational changes, sequence data and homologues and MEDLINE abstracts (see Table 1).

To facilitate rapid and effective access to SCOP, a number of mirrors have been established, a full current list of which can be found *via* the URL above. The facilities provided by the various sites are always the same, so you will lose nothing by accessing your nearest mirror. The implementation does differ: for example currently sequence similarity searching is always carried out at the main scop.mrc-lmb.cam.ac.uk site, however this is transparent to the user who will always be returned a search results page marked up with links to pages on the mirror that they started from.

4. Evaluating the effectiveness of sequence-alignment methods

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Despite this the overall and relative capabilities of different search procedures have until recently been largely unknown. This is because it is difficult to verify algorithms on sample data as this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated (nearly all known homologs have been identified by sequence analysis, the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insuffi-

```
>c1dant 2.1.2.1.1 BLOOD COAGULATION FACTOR VIIA;CHAIN: L, H;EC: 3.4.21.21;ENGINEERE
TVAAYNLTWKSTNFKTILEWEPKPVNQVYTVQISTKSGDWKSKCFYTTDTECDLTDEIVK
DVKQTYLARVFSYPA
>c1danu 2.1.2.1.1 BLOOD COAGULATION FACTOR VIIA;CHAIN: L, H;EC: 3.4.21.21;ENGINEERE
EPLYENSPEFTPYLETNLGQPTIQSFEQVGTKVNVTVEDERTLVRRNNTFLSLRDVFGKD
LIYTLYYWSGKKTAKTNTNEFLIDVDKGENYCFSVQAVIPSRTVNRKSTDSPVECM
>c1danh 2.31.1.2.16 BLOOD COAGULATION FACTOR VIIA;CHAIN: L, H;EC: 3.4.21.21;ENGINEERE
IVGGKVCPKGECPWQVLLLVNGAQLCGGTLINTIWVVSAAHCFDKIKNWRNLIAVLGEHD
LSEHDGDEQSRRVAQVIIPSTYVPGTTNHDIALLRLHQPVVLTDHVVPLCLPERTFSERT
LAFVRFSLVSGWGQLLDRGATALELMVLNVPRLMTQDCLQQSRKVGDSPNITEYMFCAGY
SDGSKDSCKGDSGGPHATHYRGTWYLTGIVSWGQGCATVGHFGVYTRVSQYIEWLQKLMR
SEPRPGVLLRAPFP
>c1danl 7.24.1.1.1,7.3.9.1.3 BLOOD COAGULATION FACTOR VIIA;CHAIN: L, H;EC: 3.4.21.21;ENGINEERE
ANAFLLRPGSLRCKQCSFARIFKDARTKLFWISYSDGDQCASSPCQNGGSCKDQLQSYIC
FCLPAFEGRNCETHKDDQLICVNENGGCEQYCSDHTGTKRSCRCHEGYSLLADGVSCTPT
VEYPCGKIPILE
>c1cfi_ 7.24.1.1.3 COAGULATION FACTOR IX;CHAIN: NULL;FRAGMENT: THE GLA AND AROMATIC
YNSGKLFVQGNLRCMKCSFARVFNTRTTFWKQYVD
```

(*a*)

```
>d1dan.1 2.1.2.1.1 (t,u:1-16) Extracellular region of human tissue factor [human (Homo sapiens)
TVAAYNLTWKSTNFKTILEWEPKPVNQVYTVQISTKSGDWKSKCFYTTDTECDLTDEIVK
DVKQTYLARVFSYPAxEPLYENSPEFTPYLET
>d1danu1 2.1.2.1.1 (17-116) Extracellular region of human tissue factor [human (Homo sapiens)]
NLGQPTIQSFEQVGTKVNVTVEDERTLVRRNNTFLSLRDVFGKDLIYTLYYWSGKKTAKT
NTNEFLIDVDKGENYCFSVQAVIPSRTVNRKSTDSPVECM
>d1danh_ 2.31.1.2.16 Coagulation factor VIIa [human (Homo sapiens)]
IVGGKVCPKGECPWQVLLLVNGAQLCGGTLINTIWVVSAAHCFDKIKNWRNLIAVLGEHD
LSEHDGDEQSRRVAQVIIPSTYVPGTTNHDIALLRLHQPVVLTDHVVPLCLPERTFSERT
LAFVRFSLVSGWGQLLDRGATALELMVLNVPRLMTQDCLQQSRKVGDSPNITEYMFCAGY
SDGSKDSCKGDSGGPHATHYRGTWYLTGIVSWGQGCATVGHFGVYTRVSQYIEWLQKLMR
SEPRPGVLLRAPFP
>d1danl1 7.3.9.1.3 (37-76) Coagulation factor VIIa [human (Homo sapiens)]
GDQCASSPCQNGGSCKDQLQSYICFCLPAFEGRNCETHKD
>d1danl2 7.3.9.1.3 (77-132) Coagulation factor VIIa [human (Homo sapiens)]
DQLICVNENGGCEQYCSDHTGTKRSCRCHEGYSLLADGVSCTPTVEYPCGKIPILE
>d1danl3 7.24.1.1.1 (1-36) Coagulation factor VIIa [human (Homo sapiens)]
ANAFLLRPGSLRCKQCSFARIFKDARTKLFWISYSD
>d1cfi__ 7.24.1.1.3 Coagulation factor IX (IXa) [human (Homo sapiens)]
YNSGKLFVQGNLRCMKCSFARVFNTRTTFWKQYVD
```

(*b*)

```
d1dan.1     1dan     t:,u:91-106    1.002.001.002.001.001
d1danu1     1dan     u:107-210      1.002.001.002.001.001
d1danh_     1dan     h:             1.002.031.001.002.016
d1danl1     1dan     l:47-86        1.007.003.009.001.003
d1danl2     1dan     l:87-142       1.007.003.009.001.003
d1danl3     1dan     l:1-46         1.007.024.001.001.001
d1cfi__     1cfi     -              1.007.024.001.001.003
```

(*c*)

Fig. 3. Entries are shown for PDB files 1DAN and 1CFI in the SCOP *FASTA* format files for (*a*) PDB chains and (*b*) SCOP domains and (*c*) in the SCOP domain definition flat file. The format of (*a*) and (*b*) is: `>scopid scopcode [,scopcode] (region) Description SEQUENCE`. `scopid` is six characters for chains (cXXXXY) and seven characters for domains (dXXXXYZ), where the prefix c or d indicates chain or domain; XXXX is the PDB code; Y is the PDB chain and Z is an arbitrary number indicating the domain (*i.e.*, the first part of the sequence is not necessarily labelled dXXXXY1). For entries with an unlabelled chain, '`_`' is used for Y. For domains composed of multiple chains Y becomes '`.`' and the chain information is embedded in the `region` element. For entries with only a single domain, '`_`' is used for Z. `scopcode` is a domain classification identifier and is of the format `a.b.c.d.e.f` where `a` is class; `b` is fold; `c` is superfamily; `d` is family; `e` is species and `f` is protein. Thus, entries with `a.b.c` in common are from the same superfamily *etc.* If the `scopid` is for a PDB chain which contains more than one *type* of domain then a series of `scopcodes` are listed separated by '`,`'. Note that `scopcodes` *change with each release* of SCOP, where as `scopids` change only if the domain organization of that PDB entry is revised. `region` is found only in entries where `scopid` is a domain which is part of a PDB chain and specifies a range with respect to the *sequence* in the corresponding `scopid` chain entry. This does not necessarily correspond to the range of residue numbers in the PDB entry. `Description` is a description of the entry, in the case of chains extracted directly from the PDB header and in the case of domains extracted from SCOP. The format of (*c*) is similar: `scopid<TAB>pdbid<TAB>pdbregion<TAB>fullscopcode`. Differences are: `scopid` is always a domain code `pdbid` is the PDB id (XXXX from scopid). `pdbregion` is similar to `region` but is of format chain:start-end where start and end are PDB residue numbers (from ATOM records) and do not relate to the index of the corresponding sequence in the *FASTA* format file. `fullscopcode` is equivalent to `scopcode` expect for the leading zeros and the initial number (which is currently unused). These values map to the corresponding page in scop for the domain of that line, such that the page for `d1cfi_` is http://scop.mrc-lmb.cam.ac.uk/scop/data/1.007.024.001.001.003.html *in this release* of SCOP. However, these page numbers (and the associated scopcodes) change with each release. The correct way to refer to `d1cfi_` is: http://scop.mrc-lmb.cam.ac.uk/scop/search.cgi?sid = d1cfi_. 1CFI (bottom) is an example of the simplest type of entry: it has a single chain (unlabelled) which is also a single domain. 1DAN is one of the most complex examples. It has four chains, T, U, H and L. H is a single-chain domain (`d1danh_`). L is a chain which contains three domains (`d1danl1`, `d1danl2`, `d1danl3`). There are two more domains: one is the second part of chain U (`d1danu1`); the other is composed of all of chain T and first part of chain U (`d1dan.1`). Note that the sequence of this last domain is composed of fragments from two chains concatenated with a lower case 'x'. The same is performed where a domain is composed of two parts of the same chain, interupted by an insertion domain. Note also the differences between the `region` (in *b*) and `pdbregion` (in *c*) records, which show how different sequence indices and PDB residue numbers can be.

cient, imperfectly characterized, or artificial test data (see Brenner *et al.*, 1998).

As part of the maintenance of SCOP, new structures are automatically processed. One of the initial steps is to cluster the sequences of protein chains of known structures at different levels of sequence similarity. This has resulted in a series of non-redundant sequence databases, referred to as PDB40, PDB90, PDB95 (Fig. 3*a*), where the number refers to percentage sequence identity as modified by the HSSP equation (Sander & Schneider, 1991) and where the chain chosen as the representative is that with the best structural 'quality' defined from an equation combining resolution, *R* factor and *PROCHECK* values (Laskowski *et al.*, 1993). The final SCOP classification is used to annotate the headers of these *FASTA* format files and to split them into domains. The result is a set of domain sequence databases, PDB40D, PDB90D *etc.* where the full set of true and false pairwise relationships between the sequences can be inferred from the scopcode in the headers (Fig. 3*b*). These databases are used within SCOP for the sequence search facility (see above and Fig. 2), however they are also ideally suited as test data for the calibration of sequence searching algorithms.

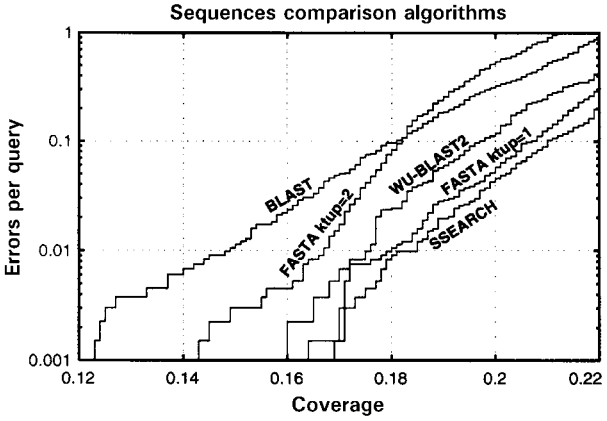

Fig. 4. Coverage *versus* error plot of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (*E* or *P* values) on the PDB40D-B database (Brenner *et al.*, 1998). In this analysis, the best method is *SSEARCH*, which finds 18% of relationships at 1% errors per query (EPQ). *FASTA* ktup = 1 and *WU-BLAST*2 are almost as good. In the coverage *versus* error plot, the × axis indicates the fraction of all homologs in the database (known from structure) which have been detected, *i.e.* the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D contains a total of 4522 homologs, so a score of 10% indicates identification of 452 relationships. The y axis reports the number of EPQ. Because there are 1323 queries made in the PDB40D all-*versus*-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Copyright National Academy of Sciences USA, Brenner *et al.* (1998); used with permission.

The databases are used for calibration in the following way. Using the algorithm to be tested, an all-against-all search is performed, *i.e.* each sequence in the database is searched against every other sequence. The entire set of results from all the database searches are then ranked together using the scoring scheme to be evaluated. For a database of 1323 sequences (*e.g.* PDB40D-B) the ranked list could contain as many as 874 503 distinct pairwise comparisons, however only 4522 represent true relationships (Brenner *et al.*, 1998). Two cumulative scores are generated moving a threshold down the list from the best score to the worst: the fraction of the total number of 'true' pairwise relationships that lie above the threshold (the coverage) and the fraction of the relationships in the list that are false (the accuracy). Plotting these two values as a coverage/accuracy plot, it is possible to compare the performance of different algorithms and establish the score threshold that relates to a given accuracy (Fig. 4).

Calibration of the commonly used algorithms *BLAST* (Altschul *et al.*, 1990), *WU-BLAST*2 (Altschul & Gish, 1996), *FASTA* and *SSEARCH* (Pearson, 1996) revealed three key conclusions that are of practical use for those carrying out sequence database searches (Brenner *et al.*, 1998).

4.1. *Algorithm*

Given an error limit of 1% *SSEARCH* detected the most distant relationships, with *FASTA* ktup = 1 and *WU-BLAST*2 being almost as good (Fig. 4). *FASTA* ktup = 1 is more computationally expensive than *BLAST* (~4 times slower) and *SSEARCH* is even more so (~25 times slower than *BLAST*).

4.2. *Scoring*

Statistical scoring schemes (*P* values and *E* values) produced the best results. Sequence identity was found to be a very poor measure of similarity, with examples of long alignments between unrelated protein structures having high percentage identity (*e.g.* 39% over 64 residues, 36% over 74 residues and 34% over 85 residues). However, whereas the empirical implementation of *E* values in *FASTA*/*SSEARCH* fairly accurately reflected the true error rate the analytical implementation of *P* values in *BLAST* (Karlin & Altschul, 1990, 1993) overestimated the likelihood of a match being correct by several orders of magnitude. Both *E* values and *P* values are based on extreme value distributions, the difference between them being that *P* values can be thought of as the probability that an alignment is incorrect (*i.e.* are corrected for database size), whereas *E* values represent raw expected errors per query (*i.e.* not corrected for database size).

4.3. *Coverage*

The coverage of even the best algorithm was remarkably low: only 18% of relationships in the PDB40D database are identified when applying the 1% error-rate threshold with the most sensitive algorithm tested (*SSEARCH*) and the most discriminating scoring function (*E* values). Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.

Knowing the meaning of the score of an alignment has become even more critical in the current era of genome analysis, where there are too many sequence comparisons to evaluate each manually. Applying the results of this calibration it has been possible to evaluate the distribution of families of proteins in whole genomes with confidence (Brenner *et al.*, 1995).

This calibration scheme has also been used to evaluate more sophisticated approaches to sequence searching. It has been anecdotal that 'intermediate' sequences can be used to link more distantly related proteins, *i.e.* first carry out a database search against the sequence of interest and then carry out database searches with each sequence returned from the first search. Calibration against PDB40D showed that using the same algorithm (*FASTA*) this approach increases the coverage by ~70% when applying the 1% error-rate threshold (Park *et al.*, 1997). Work to evaluate sequence search methods relying on multiple sequence alignments such as Hidden Markov Models (Eddy, 1996; Krogh *et al.*, 1994) and the recently developed iterative version of *BLAST*2 (Altschul *et al.*, 1997) (referred to as psi-*BLAST*) have shown significantly better performance by the same criteria (Park *et al.*, unpublished results; Brenner *et al.*, in preparation).

The databases used for these studies are now freely available *via* the SCOP URL and the format of their headers is shown in Fig. 3.

5. Statistics of protein structural data

With structural data conveniently organized into domains, it is straightforward to investigate the population statistics of the protein structures we currently know. A recent survey of the classification in SCOP (Brenner *et al.*, 1997) clearly shows that even after the high degree of redundancy in PDB has been taken into account, the frequency of occurrence of certain folds is much greater than would be expected by chance, as has been pointed out previously (Orengo *et al.*, 1994). Recalculation of the tables shown there for the most recent version of SCOP (1.37), which contains 20% more domains but only 11% more folds, shows an essentially similar picture.

The raw data to explore the classification in this way can of course be extracted from the SCOP WWW pages (if one likes writing HTML parsers) however there is an easier way in the form of the flat file shown in Fig. 3(*c*). This lists all domains classified in SCOP, not just the subset of protein chains which are defined in the headers of the *FASTA* format files listed above, and can again be accessed from the SCOP URL.

6. Conclusions

We have found that the easy access to data and images provided by SCOP make it a powerful general-purpose interface to the PDB (Brenner *et al.*, 1995). The specific lower levels should be helpful for comparing individual structures with their evolutionary and structurally related counterparts. On a more general level, the highest levels of classification provide an excellent overview of the diversity of protein structures now known and would be appropriate both for researchers and students. Having created the classification we have found that it has many other uses, some of which have been listed here. We hope that other researchers will find yet more uses for the raw data files that are now provided with each release.

TJPH is grateful to the MRC/DTI/ZENECA LINK programme and AGM is grateful to the MRC for financial support. SEB is grateful for support from a Sloan/DOE fellowship in computational molecular biology.

References

Abola, E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *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. & Gish, W. (1996). *Methods Enzymol.* **266**, 460–480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & D. J. Lipman. (1990). *J. Mol. Biol.* **215**, 403–410.

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

Appel, R. D., Bairoch, A. & Hochstrasser, D. F. (1994). *Trends Biol. Sci.* **19**, 258–260.

Benson, D., Lipman, D. J. & Ostell, J. (1993). *Nucleic Acids Res.* **21**, 2963–2965.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997). *Curr. Opin. Struct. Biol.* **7**, 369–376.

Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. (1995). In *Computer Methods for Macromolecular Sequence Analysis*, edited by R. F. Doolittle. Orlando: Academic Press.

Brenner, S. E., Hubbard, T. J. P., Murzin, A. & Chothia, C. (1995). *Nature (London)*, **378**, 140.

Bycroft, M., Hubbard, T. J. P., Proctor, M., Freund, S. M. V. & Murzin, A. G. (1997). *Cell*, **88**, 235–242.

Eddy, S. R. (1996). *Curr. Opin. Struct. Biol.* **6**, 361–365.

Fitzgerald, P. C. (1994). In *WWW94, First International Conference on the World Wide Web*, Chemistry Workshop, Elsevier Science BV, CERN, Geneva, Switzerland.

Gerstein, M., Lesk, A.M. & Chothia, C. (1994). *Biochemistry*, **33**, 6739–6749.

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

Holm, L. & Sander, C. (1994). *Nucleic Acids Res.* **22**, 3600–3609.

Holm, L. & Sander, C. (1995). *Trends Biochem. Sci.* **20**, 478–480.

Holm, L. & Sander, C. (1996). *Science*, **273**, 595–602.

Karlin, S. & Altschul, S. F. (1990). *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Karlin, S. & Altschul, S. F. (1993). *Proc. Natl Acad. Sci. USA* **90**, 5873–5877.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. J. (1994). *J. Mol. Biol.* **235**, 1501–1531.

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

Moult, J., Bryant, S. H., Fidelis, K., Hubbard, T. J. P. & Pedersen, J. T. (1997). *Proteins*, S**1**, 3–6.

Murzin, A. G. (1994). *Curr. Opin. Struct. Biol.* **4**, 441–449.

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

Orengo, C. (1994). *Curr. Opin. Struct. Biol.* **4**, 429–440.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). *Protein Eng.* **6**, 485–500.

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.

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

Pearson, W. R. (1996). *Methods Enzymol.* **266**, 227–258.

Sander, C. & Schneider, R. (1991). *Proteins*, **9**, 56–68.

Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.

Sowdhamini, R., Rufino, S. D. & Blundell, T. L. (1996). *Folding Des.* **1**, 209–220.