SCOP: a Structural Classification of Proteins database

Tim J. P. Hubbard[1], Alexey G. Murzin[1], Steven E. Brenner and Cyrus Chothia

MRC Laboratory of Molecular Biology and [1]Cambridge Centre for Protein Engineering, Hills Road, Cambridge
CB2 2QH, UK

Received November 1, 1996; Accepted November 6, 1996

ABSTRACT

The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of all known proteins 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. SCOP also provides for each structure links to atomic co-ordinates, images of the structures, interactive viewers, sequence data, data on any conformational changes related to function and literature references. The database is freely accessible on the World Wide Web (WWW) with an entry point at URL http://scop.mrc-lmb.cam.ac.uk/scop/

INTRODUCTION

At present (October, 1996) the Brookhaven Protein Databank [PDB, (1)] contains 4870 entries and the number is increasing by about 100 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 (2). It includes not only all proteins in the current version of the PDB, but many proteins for which there are published descriptions but whose co-ordinates 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.

CLASSIFICATION

The classification of the proteins is on hierarchical levels:

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 similarities; second, proteins with whose functions and structures are extremely similar; for example, globins with sequence identities of 15%.

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.

Common fold

Superfamilies and families are defined as having a common fold if their proteins have the same major secondary structures in the same arrangement and with the same topological connections [for recent reviews see (3,4)]. 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.

Class

The different folds have been grouped into classes. Most of the folds are assigned to one of the five structural classes:
1. All-α, those whose structure is essentially formed by α-helices;
2. All -β, those whose structure is essentially formed by β-sheets;
3. α/β, those with α-helices and β-strands;
4. α+β, those in which α-helices and β-strands are largely segregated, and
5. Multi-domain, those with domains of different class 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 Figure 1.

Corresponding Email: scop@mrc-lmb.cam.ac.uk

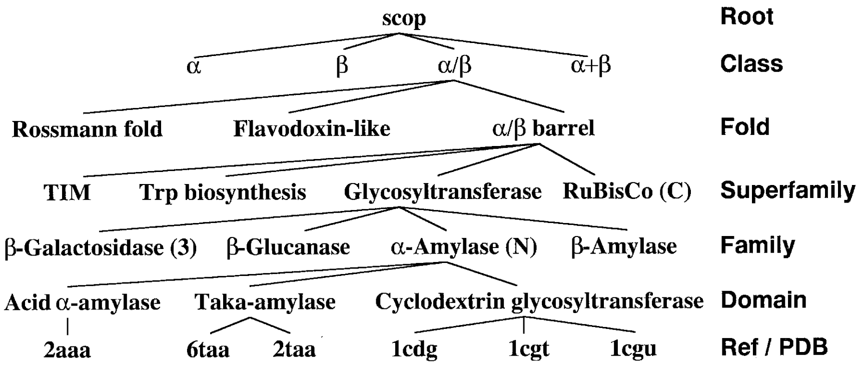


Figure 1. Region of SCOP hierarchy. All the major levels, including class, fold, superfamily and family are shown. Also shown are individual proteins and the lowest level either the PDB coordinate identifier or a literature reference. Copyright © 1994 Steven E. Brenner; reproduced with permission.

Table 1. Mirrors of SCOP

Location	Site	URL	Services
Europe	SCOP Home Server in Cambridge	http://scop.mrc-lmb.cam.ac.uk/scop/	P,K,B
East Coast USA	Protein Data Bank	http://www.pdb.bnl.gov/scop/	P,K
Japan	Biomolecular Engineering Research Institute	http://www.beri.co.jp/scop/	P,K
Israel	Weizmann Institute	http://pdb.weizmann.ac.il/scop/	P,K
Taiwan	National TsingHua University	http://life.nthu.edu.tw/scop/	P,K
China	Peking University	http://www.ipc.pku.edu.cn/scop/	P,K
Australia	Walter & Eliza Hall Institute	http://pdb.wehi.edu.au/scop/	P,K

P, Hypertext Pages with main SCOP hierarchy; K, Keyword searching; B, BLAST sequence similarity searching.
Each of the mirror sites carries out its own scientific role, and generously provides access to SCOP as a public service. We are grateful to the institutions and people that maintain these sites for enabling us to make SCOP more accessible.

Table 2. Facilities and databases to which SCOP has links

Link	Source	URL	Ref
Coordinates	PDB	http://www.pdb.bnl.gov/	(1)
Static Images	SP3D	http://expasy.hcuge.ch/ gopher://pdb.pdb.bnl.gov/	(8)
On-the-fly images	NIH molecular modelling group	http://www.nih.gov/www94/molrus	(9)
Sequences and MEDLINE entries	NCBI Entrez	http://www.ncbi.nlm.nih.gov/	(10)
Protein Motions Database	Mark Gerstein	http://hyper.stanford.edu/~mbg/ProtMotDB/	(11)
Nucleic Acids Database	Rutgers University	http://ndbdev.rutgers.edu/	(12)

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.

ORGANISATION AND FACILITIES OF SCOP

The SCOP database is available as a set of tightly coupled hypertext pages on the world wide web (WWW) and can be accessed by any machine on the internet (including Macintoshes, PCs, and unix workstations) from:
<URL: http://scop.mrc-lmb. cam.ac.uk/scop/ >.

To facilitate rapid and effective access to SCOP, a number of mirrors have been established (see Table 1). The facilities at various sites may differ; for example, sequence similarity searching is only available at the main, scop.mrc-lmb.cam.ac.uk, site.

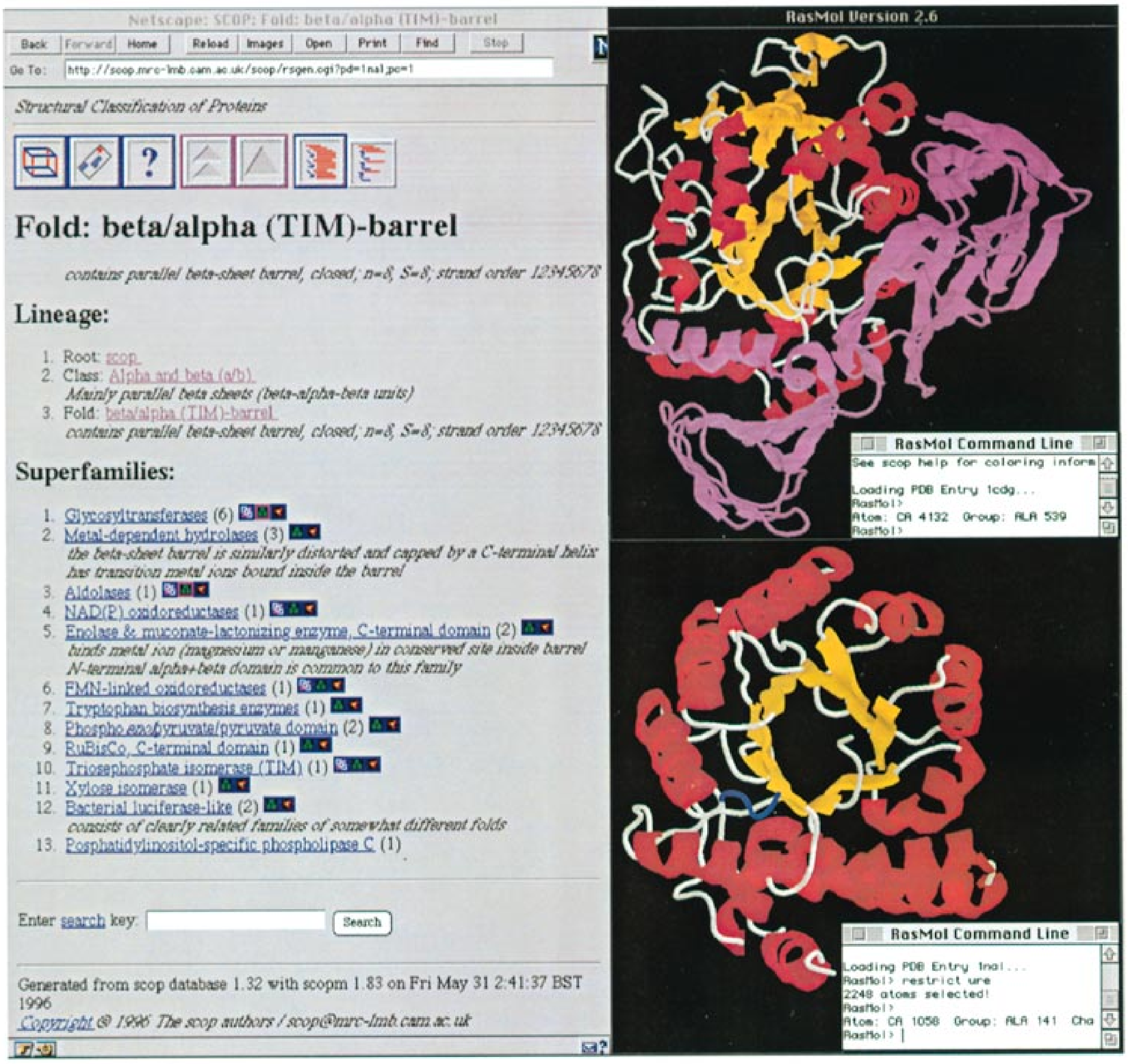
The WWW interface to SCOP has been designed to facilitate both detailed searching of particular families and browsing of the whole database. To this end, there are a variety of different techniques for navigation:

(i) Browsing through the scop hierarchy

SCOP is organised 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.

(ii) From an amino acid sequence

The sequence similarity search facility allows SCOP to be entered from the list of PDB chains found to be similar to an entered sequence and for the similarity to be displayed visually.



http://scop.mrc-lmb.cam.ac.uk/scop/rsgen.cgi?pd=1nal,pc=1

Structural Classification of Proteins

Fold: beta/alpha (TIM)-barrel

contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678

Lineage:

1. Root: scop.
2. Class: Alpha and beta (a/b).
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: beta/alpha (TIM)-barrel.
   contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678

Superfamilies:

1. Glycosyltransferases (6)
2. Metal-dependent hydrolases (3)
   the beta-sheet barrel is similarly distorted and capped by a C-terminal helix
   has transition metal ions bound inside the barrel
3. Aldolases (1)
4. NAD(P) oxidoreductases (1)
5. Enolase & muconate-lactonizing enzyme, C-terminal domain (2)
   binds metal ion (magnesium or manganese) in conserved site inside barrel
   N-terminal alpha+beta domain is common to this family
6. FMN-linked oxidoreductases (1)
7. Tryptophan biosynthesis enzymes (1)
8. Phospho enolpyruvate/pyruvate domain (2)
9. RuBisCo, C-terminal domain (1)
10. Triosephosphate isomerase (TIM) (1)
11. Xylose isomerase (1)
12. Bacterial luciferase-like (2)
    consists of clearly related families of somewhat different folds
13. Posphatidylinositol-specific phospholipase C (1)

Enter search key: Search

Generated from scop database 1.32 with scopm 1.83 on Fri May 31 2:41:37 BST 1996
Copyright © 1996 The scop authors / scop@mrc-lmb.cam.ac.uk

RasMol Command Line

```
See scop help for coloring inform

Loading PDB Entry 1cdg...
RasMol>
Atom: CA 4132  Group: ALA 539
RasMol>
```

RasMol Command Line

```
Loading PDB Entry 1nal...
RasMol> restrict ure
2248 atoms selected!
RasMol>
Atom: CA 1058  Group: ALA 141  Cha
RasMol> |
```

Figure 2. A example of the use of scop is shown on a Macintosh workstation. The page of SCOP displayed using the WWW browser program Netscape Navigator is the list of superfamilies of proteins containing a β/α TIM barrel domain. On a correctly configured workstation clicking on the green icons results in a structure being automatically loaded into the molecular viewer program RasMol [written by Roger Sayle (7)] with the secondary structures of the domain coloured as: magenta for α-helix; yellow for β-sheet and white for coil. Regions of structure outside of the domain in question are coloured purple. Running more than one copy of RasMol as shown here allows different structures to be compared: the top shows the structure of Glycosyltransferase, which has a single chain consisting of a TIM barrel and three all-β domains. The bottom is an Aldolase structure where the TIM barrel domain makes up the entire chain. Since sending large PDB files over the network can be very slow, this feature of SCOP can be configured to use local copies of PDB files if they are available On a Macintosh this is done using the helper application rmscript which also caches PDB structures that have previously been downloaded. Equivalent functionality is possible on UNIX platforms.

(iii) From a keyword

The keyword search facility returns a list of SCOP pages containing the word entered.

(iv) From a PDB identifier

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

In addition to the information on structural and evolutionary relationships contained within SCOP, each entry (for which co-ordinates are available) has links to images of the structure, interactive molecular viewers (Fig. 2), the atomic co-ordinates, data on functional conformational changes, sequence data and homologues and MEDLINE abstracts (see Table 2).

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 (5). 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.

ACKNOWLEDGEMENTS

TJPH is grateful to the MRC/DTI/ZENECA LINK programme and AGM is grateful to the MRC for financial support. SEB is principally supported by a St. John’s College (Cambridge) Benefactors’ Scholarship.

REFERENCES

1 Abola, E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J. (1987) In Allen, F. H., Bergerhoff, G., and Sievers, R. (eds), *Crystallographic Databases—Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.

2 Murzin, A., Brenner, S. E., Hubbard, T. J. P. and Chothia, C. (1995) *J. Mol. Biol.*, **247,** 536–540 (see also: http://scop.mrc-lmb.cam.ac.uk/scop).

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

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

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

6 Pearson, W. R. (1995) *Protein Sci.*, **4**, 1145–1160.

7 Sayle, R. A. and Milnerwhite, E. J. (1995) *Trends Biochem. Sci.*, **20,** 374–376.

8 Appel, R. D., Bairoch, A. and Hochstrasser, D. F. (1994) *Trends Biochem. Sci.*, **19,** 258–260.

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

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

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

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