J. Mol. Biol. (1995) 247, 536–540

JMB

COMMUNICATION

SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures

Alexey G. Murzin, Steven E. Brenner, Tim Hubbard and Cyrus Chothia*

MRC Laboratory of Molecular Biology and Cambridge Centre for Protein Engineering, Hills Road Cambridge CB2 2QH England

*Corresponding author

To facilitate understanding of, and access to, the information available for protein structures, we have constructed the Structural Classification of Proteins (scop) database. This database provides a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure. It also provides for each entry links to co-ordinates, images of the structure, interactive viewers, sequence data and literature references. Two search facilities are available. The homology search permits users to enter a sequence and obtain a list of any structures to which it has significant levels of sequence similarity. The key word search finds, for a word entered by the user, matches from both the text of the scop database and the headers of Brookhaven Protein Databank structure files. The database is freely accessible on World Wide Web (WWW) with an entry point to URL http://scop.mrc-lmb.cam.ac.uk/scop/

scop: an old English poet or minstrel (Oxford English Dictionary);

скоп: pile, accumulation (Russian Dictionary).

Keywords: protein families; superfamilies; folds; evolutionary relationships

Nearly all proteins have structural similarities with other proteins and, in many cases, share a common evolutionary origin. The knowledge of these relationships makes important contributions to molecular biology and to other related areas of science. It is central to our understanding of the structure and evolution of proteins. It will play an important role in the interpretation of the sequences produced by the genome projects and, therefore, in understanding the evolution of development.

The recent exponential growth in the number of proteins whose structures have been determined by X-ray crystallography and NMR spectroscopy means that there is now a large and rapidly growing corpus of information available. At present (January, 1995) the Brookhaven Protein Databank (PDB, (Abola *et al.*, 1987)) contains 3091 entries and the number is increasing by about 100 a month. To facilitate the understanding of, and access to, this information, we have constructed the Structural Classification of Proteins (scop) database. This database provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose three-dimensional structures have been determined. It includes all proteins in the current version of the PDB and almost all proteins for which structures have been published but whose co-ordinates are not available from the PDB.

The classification of protein structures in the database is based on evolutionary relationships and on the principles that govern their three-dimensional structure. Early work on protein structures showed that there are striking regularities in the ways in which secondary structures are assembled (Levitt & Chothia, 1976; Chothia *et al.*, 1977) and in the topologies of the polypeptide chains (Richardson, 1976, 1977; Sternberg & Thornton, 1976). These regularities arise from the intrinsic physical and chemical properties of proteins (Chothia, 1984; Finkelstein & Ptitsyn, 1987) and provide the basis for the classification of protein folds (Levitt & Chothia, 1976; Richardson, 1981). This early work has been taken further in more recent papers; see, for example, Holm & Sander (1993), Orengo *et al.* (1993), Overington *et al.* (1993) and Yee & Dill (1993). An extensive bibliography of papers on the classification and the determinants of protein folds is given in scop.

The method used to construct the protein classification in scop is essentially the visual inspection and comparison of structures though various automatic tools are used to make the task manageable and help provide generality. Given the

Abbreviations used: PDB, Protein Databank; scop, Structural Classification of Proteins.

0022–2836/95/140536–05 $08.00/0

© 1995 Academic Press Limited

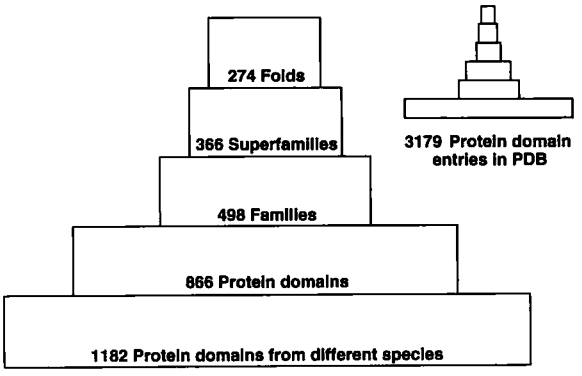

Figure 1. In scop, the unit of classification is usually the protein domain. Small proteins, and most of those of medium size, have a single domain and are, therefore, treated as a whole. The domains in large proteins are usually classified individually. The protein entries in the December 1994 of the Brookhaven Protein Databank (PDB) contain 3179 domains. Many of these become forms of the same protein whose differences are not significant in terms of the classification used here; for example they have different bound ligands or engineered mutations. To distinguish between these and structures of the same protein from different organisms, proteins listed within a family are subclassified by species. Classification of the 3179 domains show that they come from 498 families that can be clustered into 366 superfamilies and 279 different folds. In addition to these, scop contains entries for 195 proteins that do not have atomic co-ordinates available from the PDB at present but for which description of their structures have been published.

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 is on hierarchical levels that embody the evolutionary and structural relationships.

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 residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very 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, actin, the ATPase domain of the heat-shock protein and hexokinase (Flaherty *et al.*, 1991).

COMMON FOLD. Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in same arrangement with the same topological connections. In scop we give for each fold short descriptions of its main structural features. Different proteins with the same fold usually have peripheral elements of secondary structure and turn regions that differ in size and conformation and, in the more divergent cases, these differing regions may form half or more of each structure. For proteins placed together in the same fold category, the structural similarities probably arise from the physics and chemistry of proteins favouring certain packing arrangements and chain topologies (see above). There may, however, be cases where a common evolutionary origin is obscured by the extent of the divergence in sequence, structure and function. In these cases, it is possible that the discovery of new structures, with folds between those of the previously known structures, will make clear their common evolutionary relationship.

CLASS. For convenience of users, the different folds have been grouped into classes. Most of the folds are assigned to one of the five structural classes on the basis of the secondary structures of which they composed: (1) all alpha (for proteins whose structure is essentially formed by α-helices), (2) all beta (for those whose structure is essentially formed by β-sheets), (3) alpha and beta (for proteins with α-helices and β-strands that are largely interspersed), (4) alpha plus beta (for those in which α-helices and β-strands are largely segregated) and (5) multi-domain (for those with domains of different fold and for which no homologues are known at present). Note that we do not use Greek characters in scop because they are not accessible to all world wide web viewers. More unusual proteins, peptides and the PDB entries for designed proteins,

Table 1

Facilities and databases to which SCOP has links

Link	Source	URL	Reference
Co-ordinates	PDB	http://www.pdb.bnl.gov/	(Abola *et al.*, 1987)
Static images	SP3D	http://expasy.hcuge.ch/ gopher://pdb.pdb.bnl.gov/	(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)

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 co-ordinates. The location and nature of the links will vary as databases evolve and relocate.

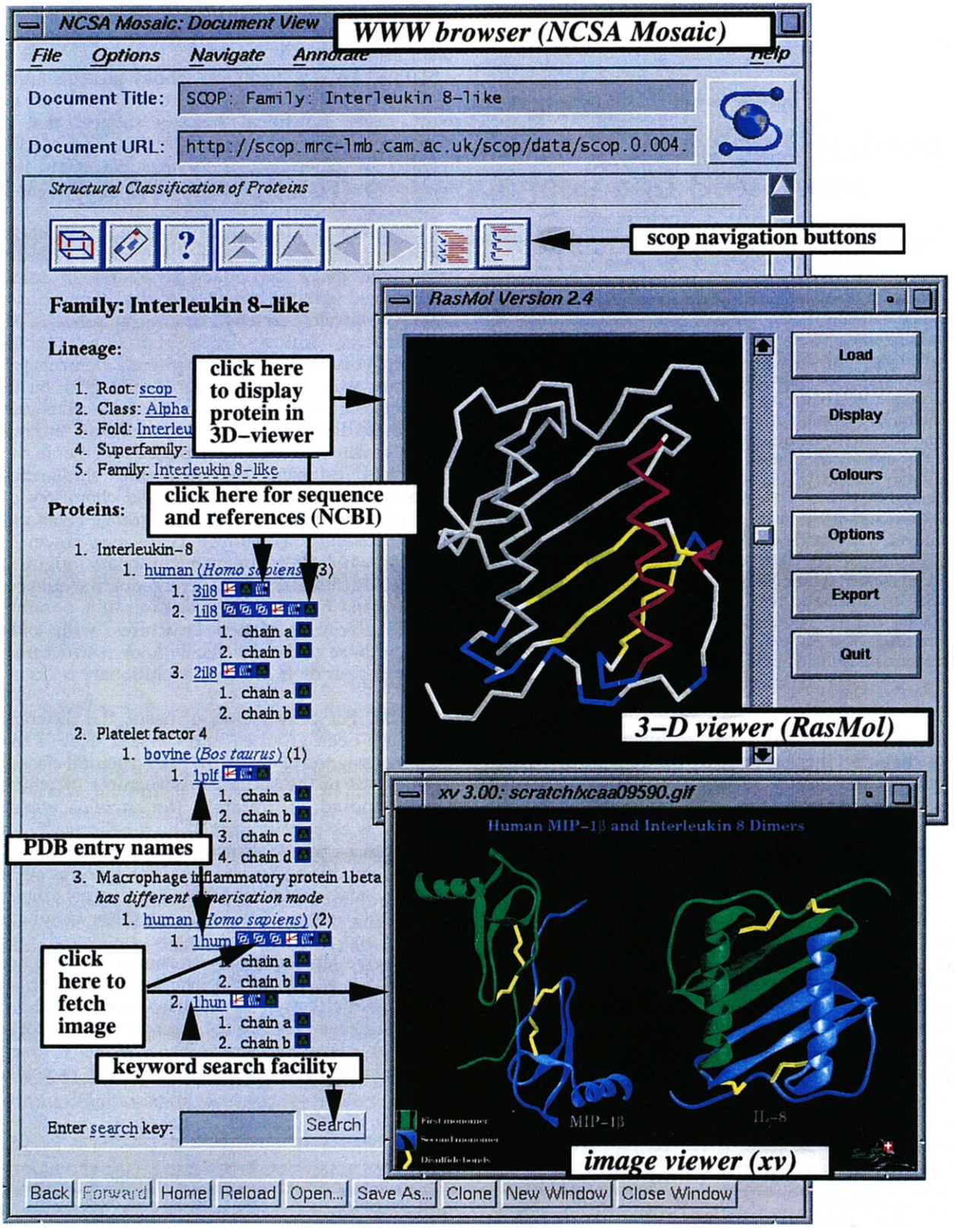

Figure 2. A typical scop session is shown on a unix workstation. A scop page, of the Interleukin 8-like family, is displayed by the *WWW browser program (NCSA Mosaic)* (Schatz & Hardin, 1994). Navigating through the tree structure is accomplished by selecting any underlined entry, by clicking on buttons (at the top of each page) and by keyword searching (at the bottom of each page). The static image comparing two proteins in this family was downloaded by clicking on the icon indicated and is displayed by image-viewer program *xv*. By clicking on one of the green icons, commands were sent to a molecular viewer program (*RasMol*) written by Roger Sayle (Sayle, 1994), instructing it to automatically display the relevant PDB file and colour the domain in question by secondary structure. Since sending large PDB files over the network can be slow, this feature of scop can be configured to use local copies of PDB files if they are available. Equivalent WWW browsers, image-display programs and molecular viewers are also available free for Windows-PC and Macintosh platforms.

theoretical models, nucleic acids and carbohydrates, have been assigned to other classes.

The number of entries, families, superfamilies and common folds in the current version of scop are shown in Figure 1. The exact position of boundaries between family, superfamily and fold are, to some degree, subjective. However, because all proteins that could conceivably belong to a family or superfamily are clustered together in the encompassing fold category, some users may wish to concentrate on this part of the database.

In addition to the information on structural and evolutionary relationships, each entry (for which co-ordinates are available) has links to images of the structure, interactive molecular viewers, the atomic co-ordinates, sequence data and homologues and MEDLINE abstracts (see Table 1).

Two search facilities are available in scop. The homology search permits users to enter a sequence and obtain a list of any structures to which it has significant levels of sequence similarity. The key word search finds, for a word entered by the user, matches from both the text of the scop database and the headers of Brookhaven Protein Databank structure files.

To provide easy and broad access, we have made the scop database available as a set of tightly coupled hypertext pages on the world wide web (WWW). This allows it to be accessed by any machine on the internet (including Macintoshes, PCs and workstations) using free WWW reader programs, such as Mosaic (Schatz & Hardin, 1994). Once such a program has been started, it is necessary only to "open" URL:

http://scop.mrc-lmb.cam.ac.uk/scop/

to obtain the "home" page level of the database.

In Figure 2 we show a typical page from the database. Each page has buttons to go back to the top-level home page, to send electronic mail to the authors, and to retrieve a detailed help page. Navigating through the tree structure is simple; selecting any entry retrieves the appropriate page. In addition, buttons make it possible to move within the hierarchy in other manners, such as "upwards" to obtain broader levels of classification.

The scop database was originally created as a tool for understanding protein evolution through sequence-structure relationships and determining if new sequences and new structures are related to previously known protein structures. On a more general level, the highest levels of classification provide an overview of the diversity of protein structures now known and would be appropriate both for researchers and students. The specific lower levels should be helpful for comparing individual structures with their evolutionary and structurally related counterparts. In addition, we have also found that the search capabilities with easy access to data and images make scop a powerful general-purpose interface to the PDB.

As new structures are released by PDB and published, they will be entered in scop and revised versions of the database will be made available on WWW. Moreover, as our formal understanding of relationships between structure, sequence function and evolution grows, it will be embodied in additional facilities in the database.

Acknowledgements

We thank Sean Eddy, Graeme Mitchison and Erik Sonnhammer for discussions and useful suggestions and Roger Sayle, the author of rasmol, for suggesting the tcl/tk interface to rasmol. The University of Cambridge School of Biological Sciences is providing the principal database access point. S.E.B. is grateful to Herchel Smith and Harvard University, St. John's College, Cambridge Overseas Trust, American Friends of Cambridge University and CVCP/ORS for support. T.H. is grateful to ZENECA for support.

References

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

Appel, R. D., Bairoch, A. & Hochstrasser, D. F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.* **19**, 258–260.

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

Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**, 537–572.

Chothia, C., Levitt, M. & Richardson, D. (1977). Structure of proteins: packing of α-helices and β-sheets. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 4130–4134.

Finkelstein, A. V. & Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns. *Prog. Biophys. Mol. Biol.* **50**, 171–190.

FitzGerald, P. C. (1994). A WWW Forms interface to facilitate access (browsing, searching and viewing) of the molecular structure data contained within the Brookhaven Protein Data Bank (PDB). *Proceedings of WWW94 (First International Conference on the World Wide Web), Chemistry Workshop, CERN, Geneva*, Elsevier Science BV, Switzerland.

Flaherty, K. M., McKay, D. B., Kabsch, W. & Holmes, K. C. (1991). Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70 kDa heat shock cognate protein. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 5041–5045.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature (London)*, **261**, 552–558.

Orengo, C., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identifying and classifying protein fold families. *Protein Eng.* **6**, 485–500.

Overington, J. P., Zhu, Z. Y., Sali, A., Johnson, M. S., Sowdhamini, R., Louie, C. & Blundell, T. L. (1993). Molecular recognition in protein families: a database of three-dimensional structures of related proteins. *Biochem. Soc. Trans.* **21**, 597–604.

Richardson, J. S. (1976). Handedness of crossover connections in β-sheets. *Proc. Nat. Acad. Sci., U.S.A.* **73**, 2619–2623.

Richardson, J. S. (1977). β-Sheet topology and the relatedness of proteins. *Nature (London)*, **268**, 495–500.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167–339.

Sayle, R. (1994). Rasmol. *WWW, URL ftp://ftp.dcs.ed.ac.uk/rasmol.*

Schatz, B. R. & Hardin, J. B. (1994). NCSA Mosaic and the world wide web: global hypermedia protocols for the Internet. *Science*, **265**, 895–901.

Sternberg, M. J. E. & Thornton, J. M. (1976). On the conformation of proteins: the handedness of the β-strand–α-helix–β-strand unit. *J. Mol. Biol.* **105**, 367–382.

Yee, D. P. & Dill, K. A. (1993). Families and the structural relatedness among globular proteins. *Protein Sci.* **2**, 884–899.

Edited by F. E. Cohen

(Received 1 November 1994; accepted 11 January 1995)