scop: structural classification of proteins

Take world authorities on protein structure, add innovative programming, the Internet and hypertext, and the result is **scop**. The **scop** database differs from existing protein structure databases[1–3], by organizing proteins hierarchically, according to their family and fold, so that once the protein of interest has been identified, it is easy to find other proteins that 'look' similar. Since **scop** is maintained and updated by its authors, it represents the complete public-domain knowledge of protein structure at any given time. It is available via the World Wide Web (WWW) and Mosaic software (Box 1), so that anyone with a modest computer and an Internet connection can access the database.

Over the past few years, advances in the techniques to determine protein structures by X-ray crystallography and nuclear magnetic resonance have led to an exponential rise in the number of structures determined. Cyrus Chothia, one of **scop**'s authors, estimates that there will be one new protein structure determined for every day in 1995. Faced with this onslaught of new information, the effective and fast categorization of new structures is essential; **scop** is one answer to the data overload problem. The pre-release version of **scop** contains 2854 protein chains and 1027 proteins from different species. These are organized into 413 families, 299 superfamilies and 219 folds (Box 2). An important feature of **scop** is that it also includes 128 references to structures that, although published in the scientific literature, have yet to find their way into the public databases.

Box 1. Accessing scop

scop will work under any WWW browser. One of the best of these is Mosaic from the US National Center for Supercomputer Applications (NCSA). Mosaic is available (free) for computers that run X-windows (such as Unix workstations) or Microsoft Windows, and Macintoshes. All versions can be obtained by anonymous FTP (file-transfer protocol) from ftp.ncsa.uiuc.edu. It is well worth getting Mosaic, since it will provide you with access to many other biologically useful databases. If run on a Unix colour workstation, Mosaic can support the RasMol images built into **scop**. Currently, the RasMol option is not available for Microsoft Windows or Macintoshes, but this does not prevent you from browsing or searching the unique **scop** database or displaying images from the SWISSPROT 3D Image database.

On the WWW, documents are identified by URLs (uniform resource locators). To access **scop** you should use your WWW browser to open the URL http://scop.mrc-lmb.cam.ac.uk/scop/. In the US there is a copy of **scop** accessible at the URL ftp://ncbi.nlm.nih.gov/repository/scop/index.html, and other sites have their own local copies. Once you have the **scop** home page on your screen, you can get help on the various options in the database by pressing the '?' button.

Box 2. The scop hierarchy

scop is organized in a natural hierarchy of protein structure that reflects the evolutionary relationships between proteins. First there are *families*. Proteins that have a clear evolutionary relationship are clustered together. This is taken as 30% or greater sequence identity except where definitive evidence for common ancestory is known from other sources. For example, all globins are grouped into a single family, even though some members only show pairwise identities of 15%.

Superfamilies include proteins that have a probable common evolutionary origin owing to their similar three-dimensional structures and functions, but that show little or no sequence identity. For example, hexokinase, the ATPase domain of heat-shock proteins and actin form a superfamily.

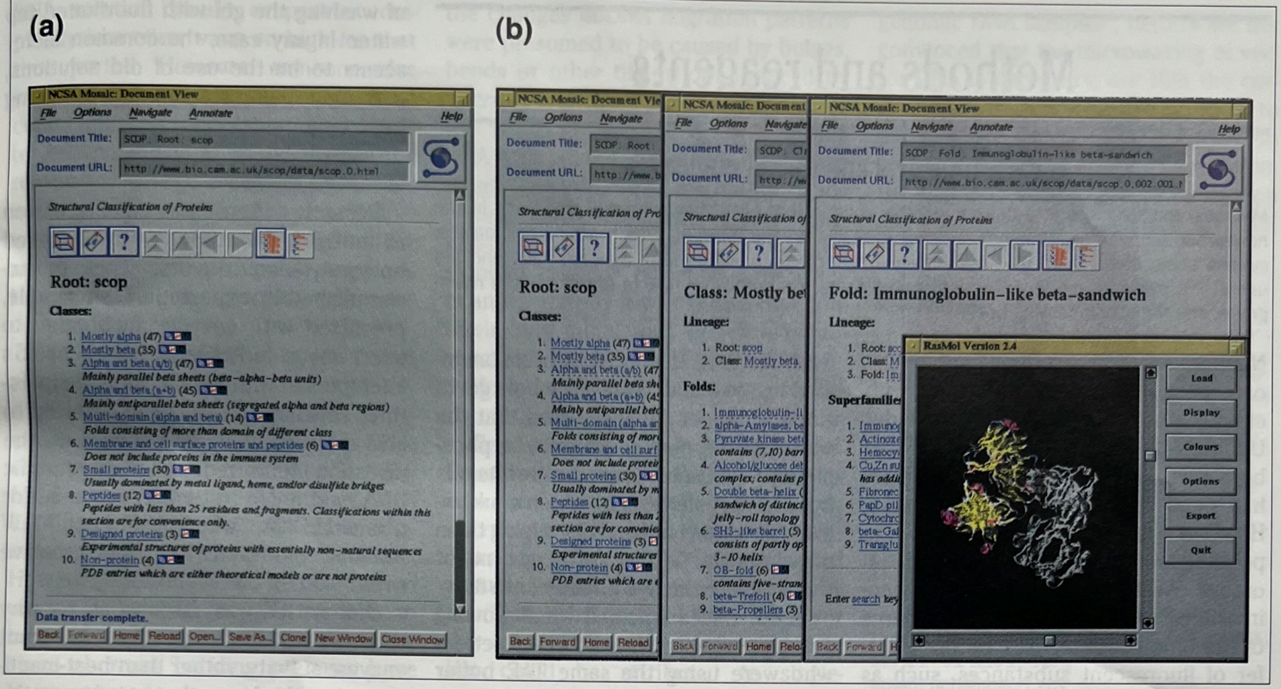
Fold families contain those proteins that share similar arrangements of secondary structures, with similar topologies, but no clear evidence of evolutionary links. The two proteins may have arrived at the same fold independently because it is thermodynamically or kinetically favoured, rather than by divergence from a common ancestor, but neither route can be proved. Proteins in the same fold family often have large insertions/deletions relative to each other that include several secondary structures or large differences in loop lengths. An example of this is the similarity between the SH2 domain and BirA domain II, where the two domains differ by three strands and a helix[4].

Although assigning proteins to families is usually straightforward, the distinction between superfamilies and fold families is often difficult. Some would argue with the details of the classification in **scop**, but it provides a good working definition based on the information currently available. As our knowledge of specific protein families increases, then proteins may shift from one family definition to another.

Using scop

Having made the connection to the **scop** database, you are presented with a page of information, known as the home page. As with all hypertext pages accessible on the WWW, clicking on blue text or blue-bordered images causes something to happen. For example, on the **scop** home page, clicking the '?' image gets a help page which explains that every page in **scop** has several components. At the top there is a set of buttons that is divided into general, navigation and display sections (Fig. 1a). General buttons allow you to return to the home **scop** page, send electronic mail to the **scop** authors or get a page of help. Navigation buttons allow you to move to the root of the classification tree, move up one level in the tree (for example, from a superfamily to a globin fold) and to move sideways in a level (for example, having selected one fold, you can browse all other folds by clicking a single button) (Fig. 1b). Display buttons allow the amount of detail available on a page to be expanded or collapsed. A few minutes of experimentation is all that is required to see the usefulness of each type of button. At the bottom of every page is a text window that allows search words to be entered; thus you can look for all proteins in the database that include the word 'kinase'. This would return a page of links that you can then click on to explore the structures further.

There are also links to other appropriate databases: SWISSPROT 3D Image, a database of images of protein structures maintained by Manuel C. Peitsch in Switzerland; Molecules 'R' Us picture generation tools maintained by the molecular modelling groups at the US National Institutes of Health; and NCBI Entrez Sequence Entries, which give the sequence information on a protein in a standard format together with Medline



Document Title: SCOP: Root: scop

Document URL: http://www.bio.cam.ac.uk/scop/data/scop.0.html

Structural Classification of Proteins

Root: scop

Classes:

1. Mostly alpha (47)
2. Mostly beta (35)
3. Alpha and beta (a/b) (47)
   Mainly parallel beta sheets (beta–alpha–beta units)
4. Alpha and beta (a+b) (45)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi–domain (alpha and beta) (14)
   Folds consisting of more than domain of different class
6. Membrane and cell surface proteins and peptides (6)
   Does not include proteins in the immune system
7. Small proteins (30)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Peptides (12)
   Peptides with less than 25 residues and fragments. Classifications within this section are for convenience only.
9. Designed proteins (3)
   Experimental structures of proteins with essentially non–natural sequences
10. Non–protein (4)
   PDB entries which are either theoretical models or are not proteins

Data transfer complete.

Document Title: SCOP: Cl

Document URL: http://ww

Class: Mostly be

Lineage:

1. Root: scop
2. Class: Mostly beta

Folds:

1. Immunoglobulin–li
2. alpha–Amylases, be
3. Pyruvate kinase bet
   contains (7,10) barr
4. Alcohol/glucose deh
   complex; contains p
5. Double beta–helix (
   sandwich of distinct jelly–roll topology
6. SH3–like barrel (5)
   consists of partly op 3–10 helix
7. OB–fold (6)
   contains five–stran
8. beta–Trefoil (4)
9. beta–Propellers (3)

Document Title: SCOP: Fold: Immunoglobulin–like beta–sandwich

Document URL: http://www.bio.cam.ac.uk/scop/data/scop.0.002.001.h

Fold: Immunoglobulin–like beta–sandwich

Lineage:

1. Root: sco
2. Class: M
3. Fold: Im

Superfamilie

1. Immuno
2. Actinoxe
3. Hemocy
4. Cu,Zn su
   has addi
5. Fibronec
6. PapD pil
7. Cytochro
8. beta–Gal
9. Transglu

Enter search key

Figure 1

(a) The root of the **scop** tree, showing the navigation buttons. **(b)** Clicking on any hypertext link (in this case 'mostly beta' followed by 'Immunoglobulin-like beta-sandwich') allows you to move through the classification tree and view a three-dimensional image of any protein in the database.

abstracts for the protein. The protein structures can be viewed and rotated on-screen using the RasMol program written by Roger Sayle. RasMol is automatically started by clicking a button on a **scop** page and can use a local copy of the protein structure files if they are available. If not, then the file can be downloaded from the NIH Molecules 'R' Us server and RasMol is automatically invoked.

Who should use **scop**? It is so user friendly that, within a few minutes, even the greatest computerphobe can explore the structural relationships between all his or her favourite proteins. With modest computer hardware, the three-dimensional structures of the protein, automatically coloured to highlight the domain of interest and its secondary structure, can also be viewed (Fig. 1b). For the protein structure specialist, **scop** is an invaluable research aid since it provides the first up-to-date and accessible systematic classification of protein structure. In teaching, **scop** is useful to extend the images in standard textbooks of protein structure. For example, students could look up the enzymes of glycolysis to discover which ones have known three-dimensional structures and what their evolutionary relationships are. Alternatively, a survey of the structures of proteins known to bind DNA would be more comprehensive by reference to **scop** than by reading any current textbook or review.

Future improvements to scop

The **scop** database is an extremely valuable tool in its present form, but a number of important enhancements are planned. A more comprehensive text-searching facility will be incorporated that includes more of the text available in PDB files. An interface to the BLAST sequence-searching software will allow you to search the database for similarities to your sequence, then display the common regions highlighted on the protein structure. Hopefully the authors will put suitable warnings about the risks of over-interpreting sequence similarities on this option! Alternative hierarchies, for example a cofactor hierarchy and a functional tree, are also under consideration and these will greatly extend the usefulness of the database. It is currently difficult to search for all DNA-binding proteins, but the proposed functional hierarchy should overcome this problem. It would be nice if the authors could include references to the definitions of specific fold families, or where a structural similarity has been described in detail in the literature. These papers often provide insights that the structures alone cannot reveal to the casual observer. In the longer term, it would also be good to allow structural superpositions of the similarities to be displayed. This would really help to draw attention to the degrees of similarity possible within protein structure.

The **scop** database shows the sort of publication that is possible with current Internet and WWW tools. The simplicity of the interface hides some very innovative programming from the **scop** team, which draws on many other recent developments in computer software. It has arrived at just the right time to save us from being overwhelmed by the exponential growth in knowledge of protein structure.

References

[1] Islam S. A. and Sternberg M. J. E. (1989) *Protein Engineering* 2, 431–442

[2] Huysmans, M., Richelle, J. and Wodak, S. J. (1991) *Proteins Struct. Func. Genet.* 11, 59–76

[3] Bryant, S. H. (1989) *Proteins Struct. Func. Genet.* 5, 233–247

[4] Russell, R. B. and Barton, G. J. (1993) *Nature* 364, 765

GEOFFREY J. BARTON

Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford, UK OX1 3QU.
email: gjb@bioch.ox.ac.uk
WWW: http://geoff.biop.ox.ac.uk/