 Nucleic Acids Research, 2000, Vol. 28, No. 1 257–259

SCOP: a Structural Classification of Proteins database

Loredana Lo Conte*, Bart Ailey, Tim J. P. Hubbard[2], Steven E. Brenner[3], Alexey G. Murzin[1] and Cyrus Chothia

MRC Laboratory of Molecular Biology and [1]Centre for Protein Engineering, Hills Road, Cambridge CB2 2QH, UK, [2]Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK and [3]Department of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA

Received October 12, 1999; Accepted October 13, 1999

ABSTRACT

The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of known protein structures. The classification is on hierarchical levels: the first two levels, family and superfamily, describe near and distant 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 sequences of proteins in SCOP provide the basis of the ASTRAL sequence libraries that can be used as a source of data to calibrate sequence search algorithms and for the generation of statistics on, or selections of, protein structures. Links can be made from SCOP to PDB-ISL: a library containing sequences homologous to proteins of known structure. Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL. 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/

INTRODUCTION

At present (October, 1999) the Brookhaven Protein Databank (PDB) (1) contains nearly 10 000 protein entries and the number is increasing by ~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 constructed the Structural Classification of Proteins (SCOP) database (2). It includes not only the 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 (3). 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 in SCOP is on hierarchical levels as follows:

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, 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. 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. $\alpha+\beta$, those in which α-helices and β-strands are largely segregated;
5. 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.

There are now a number of other databases which classify protein structures, such as CATH (4), FSSP (5), Entrez (6) and DDBASE (7), however, the distinction between evolutionary

*To whom correspondence should be addressed. Tel: +44 1223 402010; Fax: +44 1223 213556; Email: loredana@mrc-lmb.cam.ac.uk

relationships and those that arise from the physics and chemistry of proteins is a feature that is so far unique to SCOP. 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 reliably linking very distant sequence families.

ORGANISATION AND FACILITIES OF SCOP

The SCOP database is available as a set of tightly coupled hypertext pages on the WWW via the 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:

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.

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 using the algorithms BLAST (8), FASTA or SSEARCH (9). SCOP can then be entered from the list of PDB chains found to be similar and the similarity can be displayed visually.

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.

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

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 co-ordinates are available) has links to images of the structure, interactive molecular viewers, the atomic co-ordinates, data on functional conformational changes, sequence data and homologues and MEDLINE abstracts.

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 above URL. 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.

OTHER USES OF SCOP

Non-redundant sequence databases and the evaluation of sequence alignment methods

The clustering of sequences of protein chains of known structures at different levels of sequence similarity gives a series of non-redundant sequence databases known as PDB40, PDB90, PDB95 etc. (the number refers to maximum percentage sequence identity of any pair of sequences in the sequence databases) and these are available from SCOP. The current versions are produced by the ASTRAL procedure (10).

These databases contain large sets of sequence whose evolutionary relationships are known unambiguously and are, therefore, suitable test data in the calibration of sequence searching algorithms. They form the basis of a calibration of the pairwise sequence methods (11) and of methods that use multiple sequences (12). The particular databases used for these studies are available via the SCOP URL.

Assignment of protein structures to sequences using the intermediate sequence library PDB-ISL

Two homologous sequences, which have diverged beyond the point where their homology can be recognised by a simple direct comparison, can be related through one or more other sequences that are suitably intermediate between the two. A library containing potential intermediate sequences for proteins of known structure (PDB-ISL) has been constructed (13) and can be accessed directly or through SCOP. The sequences in the library were collected from a large sequence database using the sequences of the domains of proteins of known structure as the query sequences and the program PSI-BLAST (14). Sequences of proteins of unknown structure can be matched to distantly related proteins of known structure by using pairwise sequence comparison methods to find homologues in PDB-ISL. For a given error rate the number of correct matches found is the same as that found using PSI-BLAST and a large sequence database. The advantage of this library is that, because it uses pairwise sequence comparison methods such as FASTA or BLAST, it can be searched easily and, in most cases, much more quickly (13).

ACKNOWLEDGEMENT

AGM is grateful to the MRC for financial support.

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.
3. Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A. (1995) *Methods Enzymol.*, **266**, 635–653.
4. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
5. Holm,L. and Sander,C. (1994) *Nucleic Acids Res.*, **22**, 3600–3609.

6. Hogue,C., Ohkawa,H. and Bryant,S.H. (1996) *Trends Biochem. Sci.*, **21**, 226–229.
7. Sowdhamini,R., Rufino,S.D. and Blundell,T.L. (1996) *Folding Des.*, **1**, 209–220.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
9. Pearson,W.R. (1996) *Methods Enzymol.*, **266**, 227–258.
10. Brenner,S.E., Koehl,P. and Levitt,M. (2000) *Nucleic Acids Res.*, **28**, 254–256 (this issue).
11. Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
12. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) *J. Mol. Biol.*, **284**, 1201–1210.
13. Teichmann,S.A., Chothia,C., Church,G.M. and Park,J. (2000) *Bioinformatics*, in press.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.