D226–D229 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh039

SCOP database in 2004: refinements integrate structure and sequence family data

Antonina Andreeva, Dave Howorth, Steven E. Brenner[1], Tim J. P. Hubbard[2], Cyrus Chothia[3] and Alexey G. Murzin*

MRC Centre for Protein Engineering and [3]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK, [1]Department of Plant and Microbial Biology, 461A Koshland Hall 3102, University of California, Berkeley, CA 94720-3102, USA and [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Received September 17, 2003; Accepted September 18, 2003

ABSTRACT

The Structural Classification of Proteins (SCOP) database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships. Protein domains in SCOP are hierarchically classified into families, superfamilies, folds and classes. The continual accumulation of sequence and structural data allows more rigorous analysis and provides important information for understanding the protein world and its evolutionary repertoire. SCOP participates in a project that aims to rationalize and integrate the data on proteins held in several sequence and structure databases. As part of this project, starting with release 1.63, we have initiated a refinement of the SCOP classification, which introduces a number of changes mostly at the levels below superfamily. The pending SCOP reclassification will be carried out gradually through a number of future releases. In addition to the expanded set of static links to external resources, available at the level of domain entries, we have started modernization of the interface capabilities of SCOP allowing more dynamic links with other databases. SCOP can be accessed at http://scop.mrc-lmb.cam.ac.uk/scop.

BACKGROUND

The SCOP (Structural Classification of Proteins) database is developed as an evolutionary classification, in which the main focus is to place the proteins in a coherent evolutionary framework, based on their conserved structural features. The database aims to provide a comprehensive and detailed description of the relationships between all proteins whose 3D structures have been determined. A fundamental unit of classification in the SCOP database is the protein domain. A domain is defined as an evolutionary unit observed in nature either in isolation or in more than one context in multidomain proteins. The protein domains are classified hierarchically into families, superfamilies, folds and classes, whose meaning has been discussed before (1,2).

An advantage of the SCOP database is that it embeds a theory of protein evolution as defined by human experts rather than by empirical rules implemented in a variety of bioinformatics algorithms and tools. Computational support in SCOP is used to extend the human ability to analyse and interpret the data and to make the invaluable knowledge of protein evolutionary repertoire broadly available to scientific researchers.

The first official SCOP release 9 years ago comprised 3179 protein domains grouped into 498 families, 366 superfamilies and 279 folds (1). The seven main classes in the latest release (1.65) contain 40 452 domains organized into 2327 families, 1294 superfamilies and 800 folds. These domains correspond to 20 619 entries in the Protein Data Bank (PDB) (3,4) and one literature reference to a structure with unpublished coordinates. Statistics of the current and previous releases, summaries and full histories of changes and other information are available from the SCOP website (http://scop.mrc-lmb.cam.ac.uk/scop/) together with parsable files encoding all SCOP data (5). The sequences and structures of SCOP domains are available from the ASTRAL compendium (6), and hidden Markov models of SCOP domains are available from the SUPERFAMILY database (7).

Here we present further improvements and new features implemented in SCOP since the previous update (5). Starting with release 1.63, large parts of the SCOP classification are being reorganized to facilitate the integration of structural classification with the contemporary sequence and functional classification schemes. On the top levels of the SCOP hierarchy these changes will affect only a small number of entries (~20 folds and superfamilies in SCOP have been reclassified so far). The more substantial but not so apparent rearrangements are being carried out at the lower levels and are aimed at the refinement of relationships amongst proteins and protein families. Major changes introduced in SCOP 1.63 and 1.65 are described in more detail below.

RECLASSIFICATION

The dynamic nature of SCOP is one of its main features and needs to be taken into account in applications that use the

*To whom correspondence should be addressed. Tel: +44 1223 402132; Fax: +44 1223 402140; Email: agm@mrc-lmb.cam.ac.uk

SCOP database. The continual accumulation of sequence and structural data nowadays allows more rigorous analysis and provides important information for understanding the protein world and its evolutionary repertoire. If there is new evidence about protein relationships, then this may result in a redefinition of domain boundaries and/or rearrangements of nodes in the SCOP hierarchy. A typical example is when a part of a large novel protein first classified as a single multidomain entry is subsequently observed as a stand-alone protein or in a combination of different domain types and therefore it is reclassified as a separate domain. Frequently two separately classified proteins are shown to be related through an intermediate, the structure of which has been determined more recently. The appearance of such proteins in the structural databases can help to identify more distant relationships between protein domains and thus can lead to a rearrangement that unifies distinct protein superfamilies.

Another factor influencing reclassification is integration with other databases. A project has started during the past year that aims to rationalize and integrate the SCOP information with the data about protein families housed by prominent sequence and structural databases, including InterPro (8), Pfam (9), CATH (10) and MSD (11). A milestone in this ambitious goal is the provision of stricter and more precise definitions behind the different classification schemes used in these different databases. In response to these requirements, starting with release 1.63, we have initiated a refinement of the SCOP classification that introduces a number of changes mostly at the levels below superfamily.

Membrane all-α proteins

One of the major rearrangements in the SCOP 1.63 release was a revision of the so-called Membrane all-alpha fold. Created in SCOP when there were a handful of known membrane protein structures, this fold listed protein domains classified solely on the basis of their secondary structural content without explicit consideration of their fold topologies. Prompted by the rapid progress of membrane protein crystallography, a comprehensive analysis of these domains has been undertaken and the membrane all-α proteins have been reclassified from scratch into 24 new or already existing folds in the SCOP database. Currently these protein folds are encompassed under more precise fold definitions based on the number of helices that span the membrane. New structural and probable evolutionary relationships have been discovered during the reclassification. The discovery of a new haem-binding fold is arguably the most interesting. This protein fold comprises four transmembrane helices arranged in an up-and-down bundle with the haem groups bound in between the helices. The haem-binding four-helical fold is observed in the structures of the cytochrome b subunit of the bovine cytochrome bc1 complex (1be3:C) (12), the γ subunit of *Escherichia coli* formate dehydrogenase N (1kqf:C) (13) and in the transmembrane subunits of fumarate reductase respiratory complex (1qla:C) (14). Three of the four haem ligands are conserved between the cytochrome bc1 complex and formate dehydrogenase N subunits and occupy structurally equivalent sites with the haem-binding modes of both proteins being very similar. These features considered in conjunction with good overall structural similarity of the four-helical domains could be interpreted as evidence for their common evolutionary origin. Currently these protein domains constitute a superfamily of transmembrane di-haem cytochromes.

Viral capsid and coat proteins

The former SCOP classification of viral capsid and coat proteins was based on the assumption that viruses co-evolved with their hosts. The protein domains of this fold were classified into a number of families according to the infected host. However, the increasing amount of available data on virus structures and genome sequences has caused a reassessment of the old classification concept. Mammalian picornaviruses (positive-stranded ssRNA viruses) for instance are morphologically and genetically very similar to small so-called Cricket paralysis-like viruses and to a number of plant viruses (15,16). Their coat proteins form similar heterooligomeric assemblies and display several conserved characteristic features in their folds. These similarities between mammalian and insect viruses extend to the post-processing of structural polyproteins. In SCOP release 1.65 these protein domains are grouped together and classified as belonging to the superfamily of positive-stranded ssRNA viruses. The reclassification of the viral capsid and coat protein fold results in four new superfamilies and 11 new families. The new classification explicitly follows the naming convention and virus taxonomy established by the International Committee on Taxonomy of Viruses (ICTV) (17). In addition to the internal reorganization, this protein fold was been merged with the former nucleoplasmin/PNGase F-like fold.

Antibody domains

Antibodies and their fragments are the largest group of homologous proteins of known structure. In SCOP, there are more than 2000 antibody domains organized previously in 228 separate species of variable domain combinations and 185 species of constant domain combinations. In SCOP release 1.65 all variable and constant domains have been reclassified according to their chain and source organism. The constant domains have been additionally sorted by their chain order. Our main goal was to provide a more comprehensive and systematic characterization of the structural repertoire of variable domains—a task that is not easy, having in mind the number of engineered antibody structures deposited in the PDB. In our analysis we excluded the 51 hybrid and artificial variable domains from the domain set and classified them as a separate 'engineered' species. In order to identify different groups among antibody variable domains we performed a two-phase sequence clustering. First the sequences corresponding to the germline segments were clustered using a threshold of 85% identity for the inclusion of a protein sequence to the cluster set. Then the segments were sorted according to the size of the CDR1 and CDR2 regions. We anticipate that the resulting clusters might correspond to the putative germline families in the species genomes.

E-set domains

The E-set domains are presumed to be early domains of the immunoglobulin-like fold and may be the evolutionary link between the immunoglobulin and fibronectin type III domain superfamilies. In release 1.63 the former E-set domains family was taken out of the immunoglobulin superfamily in SCOP

and transformed into a superfamily. The constituent domains were reorganized into 15 new families. The C-terminal domain of mollusc haemocyanin sharing only a partial structural similarity with the immunoglobulin-like domain of arthropod haemocyanin was reclassified into a new fold.

Protein kinases

In SCOP release 1.65 the related catalytic domains of Tyr and Thr/Ser kinases were merged into a single family of protein kinases. In fact their close relationship was confirmed by the structure determination of type I TGF-β receptor R4, a Thr/Ser protein kinase that is more similar in sequence to Tyr kinases than to the other Thr/Ser kinases (18). Even though the catalytic domains of protein kinases are very similar, there are certain motifs that can be identified in their sequences and used to characterize the functional properties of each distinct kinase. We used these specific features to assign all protein kinase domains of known structure to the major groups, defined by the substrate specificity and/or mode of regulation, and then by functional subfamilies (19). For each protein kinase, SCOP now provides a detailed description in the annotation field. This field is searchable and allows users to extract a protein set of particular interest.

Non-coordinate entries

Early SCOP releases provided classification of dozens of protein structures published in the literature but not available at the time from PDB. Classified as literature references, these structures were sole representatives of their protein families at the time. After adoption of the policy of linking the publication of structure with the obligatory release of coordinates by most of the scientific journals, classification of new non-coordinate entries in SCOP was discontinued, and their number gradually decreased. Twenty-seven of the 28 remaining proteins in SCOP 1.63 were found by a recent inspection to have closely related representative structures in PDB and were made obsolete. In the latest release, there is just one literature reference (20) representing a unique superfamily.

TECHNICAL DEVELOPMENTS

As part of the database integration project we have started to modernize the interface capabilities of SCOP and to link the databases dynamically. An initial step suggested by MSD was to implement an on-demand server of SCOP domain definitions. This is intended to avoid synchronization problems arising from the different release schedules of the various databases. It is based on Simple Object Access Protocol (SOAP) technology and is currently used by the Pfam team to display comparisons of domains in the CATH, Pfam and SCOP databases. Further developments are expected and will be made available to other interested parties.

CONCLUDING REMARKS

From the beginning, the main focus of SCOP was on the probable evolutionary relationships between proteins that were undetectable by sequence comparison methods. The hierarchically organized structural data played a major part in the development of contemporary sequence-based methods with improved sensitivity. These methods allowed clustering of the multitude of known and hypothetical proteins in the sequence databases in a relatively small number of protein sequence families. The availability of complete genome sequences allowed the exploration of evolutionary and structural repertoires of different organisms and the refinement of their phylogeny. A large fraction of the protein families of unknown structure can be assigned with confidence into the existing SCOP superfamilies. The continual accumulation of structure, sequence and genome data will allow SCOP and related databases to play an increasingly effective role in the integration of these data.

ACKNOWLEDGEMENTS

We acknowledge Dr Loredana Lo Conte's contribution to maintenance and development of the SCOP database. This work was supported by the MRC strategic grant G0100305.

REFERENCES

1. Murzin,A., Brenner,S.E., Hubbard,T.J.P. and Chothia,C. (1995) SCOP: a Structural Classification of Proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
2. Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A. (1996) Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.*, **266**, 635–643.
3. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
4. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
5. Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
6. Chandonia,J.-M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
7. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffith-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
10. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH: A hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
11. Golovin,A., Oldfield,T.J., Tate,J.G., Velankar,S., Barton,G.J., Boutselakis,H., Dimitropoulos,D., Fillon,J., Hussain,A., Ionides,J.M.C. *et al.* (2004) E-MSD: an integrated resourse for bioinformatics. *Nucleic Acids Res.*, **32**, D211–D214.
12. Iwata,S., Lee,J.W., Okada,K., Lee,J.K., Iwata,M., Rasmussen,B., Link,T.A., Ramaswamy,S. and Jap,B.K. (1998) Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science*, **281**, 64–71.
13. Jormakka,M., Tornroth,S., Byrne,B. and Iwata,S. (2002) Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. *Science*, **295**, 1863–1868.
14. Lancaster,C.R., Kroger,A., Auer,M. and Michel,H. (1999) Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 Å resolution. *Nature*, **402**, 377–385.

15. Liljas,L., Tate,J., Lin,T., Christian,P. and Johnson,J.E. (2002) Evolutionary and taxonomic implications of conserved structural motifs between picornaviruses and insect picorna-like viruses. *Arch. Virol.*, **147**, 59–84.
16. Chandrasekar,V. and Johnson,J.E. (1998) The structure of tobacco ringspot virus: a link in the evolution of icosahedral capsids in the picornavirus superfamily. *Structure*, **6**, 157–171.
17. van Regenmortel,M.H.V., Fauquet,C.M., Bishop,D.H.L., Carstens,E.B., Estes,M.K., Lemon,S.M., Maniloff,J., Mayo,M.A., McGeoch,D.J., Pringle,C.R. *et al.* (Eds) (2000). *Virus Taxonomy: the Seventh Report of the International Committee on Taxonomy of Viruses*. Academic Press, San Diego, CA.
18. Huse,M., Chen,Y.G., Massague,J. and Kuriyan,J. (1999) Crystal structure of the cytoplasmic domain of the type I TGF β receptor in complex with FKBP12. *Cell*, **96**, 425–436.
19. Hanks,S.K. (2003) Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol.*, **4**, 111.
20. Hoess,A., Watson,S., Siber,G.R. and Liddington,R. (1993) Crystal structure of an endotoxin-neutralizing protein from the horseshoe crab, *Limulus* anti-LPS factor, at 1.5 Å resolution. *EMBO J.*, **12**, 3351–3356.