264–267 *Nucleic Acids Research, 2002, Vol. 30, No. 1*

SCOP database in 2002: refinements accommodate structural genomics

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

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK, [1]Department of Plant and Microbial Biology, 111 Koshland Hall #3102, University of California, Berkeley, CA 94720-3102, USA, [2]Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK and [3]MRC Centre for Protein Engineering, Hills Road, Cambridge CB2 2QH, UK

Received September 17, 2001; Accepted October 30, 2001

ABSTRACT

The SCOP (Structural Classification of Proteins) database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships. Protein domains in SCOP are grouped into species and hierarchically classified into families, superfamilies, folds and classes. Recently, we introduced a new set of features with the aim of standardizing access to the database, and providing a solid basis to manage the increasing number of experimental structures expected from structural genomics projects. These features include: a new set of identifiers, which uniquely identify each entry in the hierarchy; a compact representation of protein domain classification; a new set of parseable files, which fully describe all domains in SCOP and the hierarchy itself. These new features are reflected in the ASTRAL compendium. The SCOP search engine has also been updated, and a set of links to external resources added at the level of domain entries. SCOP can be accessed at http://scop.mrc-lmb.cam.ac.uk/scop.

BACKGROUND

The SCOP (Structural Classification of Proteins) database (1) is a comprehensive ordering of all proteins of known structures, according to their evolutionary and structural relationships. The basic classification unit is a *protein domain*. A domain is defined as an evolutionary unit, in the sense that it is either observed in isolation in nature, or in more than one context in different multidomain proteins. Protein domains in SCOP are grouped according to *species* and hierarchically classified into *families*, *superfamilies*, *folds* and *classes*, whose meaning is described in the original articles (1,2).

A unique aspect of SCOP is that it embeds a theory of evolution as defined by a human expert, rather than the necessarily more limited set of rules implemented by a series of algorithms and automatic tools. Automation in SCOP is meant to support and extend human ability; to make expert knowledge broadly available to students and the scientific community; and, last but not least, to sustain the automation process itself, given increasing needs and constant number of persons.

The growth in content of the SCOP database closely tracks the corresponding growth in the number of deposited structures. With the exception of major software upgrades, like in the case of 1.55, in the last 2 years SCOP has been released every 4–6 months, and included the classification of all proteins whose coordinates were available in the PDB at the time we started working on the release (usually a few weeks before the release itself, a few months in the case of SCOP 1.55).

An early SCOP release in 1995 included 3179 domains clustered into 498 families, 366 superfamilies and 279 different folds (1). The first seven classes in the current release (1.55) comprise 30 403 domains, grouped into 605 different folds, 947 superfamilies and 1557 families. These domains correspond to 12 794 entries in the Protein Data Bank (3,4) and 39 references from the literature, for which the experimental coordinates are not available. The 10-fold increase in the number of domains since 1995 roughly produced twice as many folds, 2.5 times as many superfamilies and three times as many families.

Together with the ASTRAL Compendium (5), which provides sequences for all domains filtered according to different criteria, SCOP introduces a meaningful order in the huge amount of sequence and structural information coming out of the various genomics projects, and forms the basis upon which other services, like SUPERFAMILY (6), can be built.

This article briefly summarizes the new features introduced in SCOP starting with release 1.55. For a detailed description, the reader is referred to the online release notes (http://scop.mrc-lmb.cam.ac.uk/scop/release-notes.html).

NEW FEATURES IN SCOP

The original design of the SCOP database (7) proved to be flexible enough to accommodate in a relatively easy way not only the growth in number of experimentally determined protein structures in the last 7 years, but also deeper modifications of the database itself, like the recent introduction of unique identifiers.

In computational terms, although SCOP is essentially a hierarchy, a mechanism for cross-linking between nodes of the tree makes it a graph, which allows the representation of complicated

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

biological relationships (8), as well as the more restrictive parent–child relationships in a tree.

Since the original implementation of SCOP is based on a *description* of the underlying data structure, rather than on the data structure itself, it is easy to introduce new classification levels, whose need was clear since the very beginning (2), but which have not been actually included so far. To introduce a new level, all that is required is to modify the description accordingly, and the rest will fall into place automatically.

New SCOP identifiers

It is with these future extensions in mind that we designed a new set of identifiers, unique integers (*sunid*) associated to each node of the hierarchy, and a new set of concise classification strings (*sccs*). Both will be kept stable across SCOP releases in the sense defined below.

A *sunid* is simply a number which uniquely identifies each entry in the SCOP hierarchy, from root to leaves, including entries corresponding to the protein level for which there was no explicit reference before. An *sccs* is a compact representation of a SCOP domain classification, including only the most relevant levels—for class, fold, superfamily and family. For example, the *sccs* for the ribosome anti-association factor domain (PDB entry 1g61, chain A) is d.126.1.1, where 'd' represents the class, '126' the fold, '1' the superfamily, and the last '1' the family. Also, the associated *sunid*s are 53931 for class, 55908 for fold, 55909 for superfamily, 55910 for family, 55911 for protein, 55912 for species and 41126 for domain. The old SCOP domain identifier, *sid* d1g61a_, is still valid.

Together, *sunid* and *sccs* replace the old *classification page numbers* (like 1.002.044.001.002.021). These classification page numbers changed with every release, because they reflected the order in which new entries appeared in SCOP, as well as any internal rearrangement of old entries. To avoid links that would randomly point to a completely different fold as soon as SCOP is updated, all pages in the SCOP 1.55 release have been renamed.

The new identifiers provide an unambigous way to link to a SCOP entry (see http://scop.mrc-lmb.cam.ac.uk/scop/release-notes.html for details) and to refer to SCOP in related research work and in the literature. The old (correct) way of linking to SCOP, using an *sid* which identifies a domain, together with the desired classification level, remains valid for backward compatibility, but it is not recommended for new releases.

New parseable files

All the information in SCOP, with the exclusion of comments, is available in three easy-to-parse files. Together, they replace and extend the now obsolete dir.dom.scop.txt and dir.lin.scop.txt. Each of these files has a header which includes release, version and copyright information. They fully describe all domains in SCOP and the hierarchy itself, and have been designed in such a way that the likely inclusion of new levels in the current SCOP hierarchy will not break code, provided they are properly parsed. These files are ideal for computer-based large scale analysis, comparison across releases and historical summaries.

One of the files, dir.hie.scop.txt, has no precursor in releases before 1.55. It represents the SCOP hierarchy in terms of *sunid*. Each entry corresponds to a node in the tree and has two additional fields: the *sunid* of the parent of that node (i.e. the node one step up in the tree), and the list of *sunid*s for the children of that node (i.e. the nodes one step down in the tree). A second file, dir.cla.scop.txt, contains a description of all domains, their definition and their classification, both in terms of *sunid* and *sccs*. The third file, dir.des.scop.txt, contains a description of each node in the hierarchy, including English names for proteins, families, superfamilies, folds and classes.

Since the order in which entries appear within the same level are meaningful in SCOP (it is not uncommon for comments to group together some of the superfamilies, as in the TIM β/α-barrel fold, for example) this order is preserved in both dir.hie.scop.txt and dir.cla.scop.txt.

If a SCOP domain includes portions from different PDB chains which come from a single chain precursor, these are listed in the order in which they appear in the original single protein sequence. A new set of SCOP sequences corresponding to these 'genetic' domains is now available as part of the ASTRAL compendium (5), together with a manually curated mapping between SEQRES and ATOM field in PDB, which uniquely define a SCOP domain in terms of PBD coordinates.

Stable identifiers and standard reference data sets should make comparison, linking and integration of SCOP-based or related results a trivial task. The purpose is to help develop a common language that can be used without ambiguities when talking about a domain or its classification, and to avoid duplication of efforts, so that energy can be applied to further progress, and build upon solid and well tested blocks.

New links and an improved search engine

Information in SCOP is interactively accessible as a set of HTML pages and through a search engine at http://scop.mrc-lmb.cam.ac.uk/scop or one of several mirrors scattered around the world. Previous SCOP releases, starting with SCOP 1.48, are also available online at the home SCOP site at MRC (http://scop.mrc-lmb.cam.ac.uk/scop-x.xx, where 'x.xx' is the release number).

The HTML pages reflect the underlying SCOP hierarchy in an easy-to-navigate way, and include pictures of SCOP domains as well as other useful links and information. The new identifiers are visible by positioning the mouse on links. All SCOP pages have been renamed in release 1.55.

A new set of links to external resources have been added at the level of SCOP domains. For each domain in the first seven classes, there are links to supplementary information related to that domain in Pfam (9), SUPERFAMILY (6), PartsList (10) and, in case there is one or more sequences predicted to have that fold, to PRESAGE (11). Links to Pfam provide alignments to homologs from sequence databases for most of the SCOP domains. SUPERFAMILY is a collection of Hidden Markov Models (HMMs) (12) for superfamilies in SCOP, and of HMM-based genome assignments to SCOP superfamilies. PartsList adds genomic, functional and structural information to most of the SCOP entries. PRESAGE is a collaborative resource for structural genomics with a collection of proteins' annotations reflecting current experimental status, structural assignment models and predicted folds.

Linking to SCOP from external sources is now straightforward. The same mechanism used to link can also be used to search SCOP (see http://scop.mrc-lmb.cam.ac.uk/scop/release-notes.html for details). Besides that, the standard keyword search now accepts *sunid*s, (possibly right-truncated)

*sccs*s and EC numbers, as well as words that appear in any of the SCOP pages, PDB identifiers and SCOP *sid*s. It also accepts ASTRAL identifiers, including those for the new genetic domain sequences (5).

The keyword search allows for right truncation ('+' at the right end of a word) and multiple keys, which can be combined with '+' (*and*) and '–' (*and not*) word-prefix operators. The simplest search form: 'casp4', will return all the pages with 'CASP4' appearing in the text; 'yeast' will return the list of pages containing the word 'yeast'; 'yeast –saccharom+' the list of pages in which the word 'yeast' appears, but not any completion of 'saccharom'; 'yeast +saccharom+', the list of yeast proteins restricted to *Saccharomyces cerevisiae*; 'yeast +saccharom+ +elongation' a list further restricted to elongation factor domains from *Saccharomyces cerevisiae*. Similarly, 'hypoth+' returns a list of hypothetical proteins, and 'fivefold' the list of pages in SCOP corresponding to the five-fold symmetry *pentein* fold.

HOW STABLE ARE STABLE IDENTIFIERS GOING TO BE?

Both SCOP unique identifiers (*sunid*) and SCOP concise classification strings (*sccs*) are expected to be stable across releases, with the caveat that a domain definition or its classification can change. A typical example is when a domain in a multidomain protein already classified in SCOP is observed for the first time either by itself, or in a different context, and therefore qualifies as a separate domain. Also, nodes in the hierarchy can merge or split as more evidence about evolutionary relationships becomes available from experimental data. In these cases, we will make corresponding identifiers obsolete, and introduce new ones. We will not reuse any identifier (in order to avoid identifiers with more than one meaning), and we will provide an easy-to-parse history tracking all changes.

A few examples can help explain what kind of changes should be expected to occur, and why.

It is not uncommon for only a fragment, or some of the domains of a protein to be crystallized. When the structure of the N-terminal domain of syntaxin 1-A was released (1br0) (13), it appeared to have a three-helix bundle fold with a topology similar to that of spectrin repeats, and therefore was classified in SCOP 1.50 as a new superfamily under the spectrin-like repeat fold. By the time of the SCOP 1.53 release, two new structures for the same protein became available (1ez3 and 1dn1) (14,15). While the three chains in 1ez3 are three-helix bundle fragments, the B chain in 1dn1 is a four-helix bundle with a peculiar arrangement which prompted the classification of all syntaxins as a separate superfamily under the STAT-like fold instead. A comment in SCOP indicates which of the structures are from a smaller fragment.

Perhaps more biologically interesting is the itinerary of the N-terminal domain of mukB, a protein implicated in ATP-dependent chromosome partitioning during cell division, and for which a putative nucleotide-binding region was predicted on the basis of a strong sequence signal for a P-loop motif. When the experimental structure of the N-terminal domain of mukB was determined (1qhl) (16), it came as a surprise. The predicted P-loop turned out to be located in between two helices on the protein's surface. There was no structural similarity to known P-loop NTPase proteins. Accordingly, mukB was classified in SCOP on its own, as a new fold.

By the time of the next SCOP release, two new structures were solved: Rad50 ATPase (1f2t) (17) and SMC head domain (1e69) (18). The two ATPase proteins are closely related to each other and to mukB, despite a low sequence similarity. They show all the characteristic motifs of the ABC transporter ATPase domain, and a convincing structural similarity to other members of this family (17,18). The N-terminal domain of mukB corresponds to the N-terminal part of SMC, and the structural similarity suggests a similar mode of dimer complementation, in which the helix preceding the P-loop in mukB is rearranged in a β-strand when both the N- and C-terminal domains of mukB are translated as a complete head domain (18).

Because of this additional experimental evidence, the N-terminal domain of mukB is now classified under the ABC ATPase domain-like family, which also includes Rad50 and SMC.

This family is also a good example to illustrate another point, mentioned before, and source of possible changes in the near future. The ABC transporter ATPase family belongs to P-loop NTP hydrolases, a highly divergent superfamily that, should we decide to introduce new levels in SCOP, would classify as a *hyperfamily*. This would cause a rearrangement of the underlaying substructure, with the current superfamily becoming a hyperfamily, families within the superfamily turning into superfamilies, and proteins being regrouped into newly defined families.

In summary, stable identifiers will be as stable as the information assembled in SCOP is, and vary as a function of new evidence and new findings, or in order to accommodate new levels in SCOP. All variations will be recorded in an easy-to-track way and will constitute the history of the evolution not of the proteins themselves, but of how we come to know them, and we organize and structure this knowledge.

CONCLUDING REMARKS

Biological databases in all their diversity are an indispensable contribution to the scientific enterprise for interpreting and composing the huge amount of experimental data into a coherent understanding of life. The role played by these databases can only increase as the volume and intrinsic complexity of biological data rapidly expand.

The added value of a database is in the accuracy of its content, and in the structured knowledge that it embeds and makes accessible to a wider community. The SCOP database has been available on the web since 1994, and continously updated to include all newly determined protein structures since then, with the purpose of providing a coherent theory of how proteins have evolved. In this article, we have described some recent enhancements that will make its growth, usage and integration in the larger context of structured biological information available on the web a relatively easy and affordable task, even with the daunting perspective of 10 000 new proteins coming out of the various structural genomics projects in the next 10 years (19).

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., Bourne,P.E. and Berman,H.M. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
5. Chandonia,J.-M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.*, **30**, 260–263.
6. 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**, 903–909.
7. Brenner,S.E. (1996) *Molecular Propinquity*. PhD dissertation, Cambridge University, Cambridge, UK.
8. Fitch,W.M. (2000) Homology: a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
9. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 354–360.
10. Qian,J., Stenger,B. Wilson,C.A., Lin,J., Jansen,R., Teichmann,S.A., Park,J., Krebs,W.G., Yu,H., Alexandrov,V., Echols,N. and Gerstein,M. (2001) PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.*, **29**, 1750–1764.
11. Brenner,S.E., Barken,D. and Levitt,M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Res.*, **27**, 251–253.
12. Karplus,K., Barret,C. and Hughey,R. (1998) Hidden Markov Models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
13. Fernandez,I., Ubach,J., Dulubova,I., Zhang,X., Sudhof,T.C. and Rizo,J. (1998) Three-dimensional structure of an evolutionary conserved N-terminal domain of Syntaxin-1A. *Cell*, **94**, 841–849.
14. Lerman,J.C., Robblee,J., Fairman,R. and Hughson,F.M. (2000) Structural analysis of the neuronal snare protein Syntaxin-1A. *Biochemistry*, **39**, 8470–8479.
15. Misura,K.M.S., Scheller,R.H. and Weis,W.I. (2000) Three-dimensional structure of the neuronal-sec1-Syntaxin-1A complex. *Nature*, **404**, 355–362.
16. van den Ent,F., Lockhart,A., Kendrick-Jones,J. and Löwe,J. (1999) Crystal structure of the N-terminal domain of mukB: a protein involved in chromosome partitioning. *Structure*, **7**, 1181–1187.
17. Hopfner,K.P., Karcher,A., Shin,D.S., Craig,L., Arthur,L.M., Carney,J.P. and Tainer,J.A. (2000) Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell*, **101**, 789–800.
18. Löwe,J., Cordell,S.C. and van den Ent,F. (2001) Crystal structure of the SMC head domain: an ABC ATPase with 900 residues antiparallel coiled-coil inserted. *J. Mol. Biol.*, **306**, 25–35.
19. Brenner,S.E. (2001) A tour of structural genomics. *Nat. Rev. Genet.*, **1**, 801–809.