Nucleic Acids Research, 2004, Vol. 32, Database issue ***D235–D239***
DOI: 10.1093/nar/gkh117

The SUPERFAMILY database in 2004: additions and improvements

Martin Madera*, Christine Vogel, Sarah K. Kummerfeld, Cyrus Chothia and Julian Gough[1,2]

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK, [1]Department of Structural Biology, School of Medicine, Stanford University, Stanford, CA 94305-5125, USA and [2]RIKEN Genomic Sciences Centre, W121 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received September 5, 2003; Revised October 7, 2003; Accepted October 15, 2003

ABSTRACT

The SUPERFAMILY database provides structural assignments to protein sequences and a framework for analysis of the results. At the core of the database is a library of profile Hidden Markov Models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent an entire superfamily. We have applied the library to predicted proteins from all completely sequenced genomes (currently 154), the Swiss-Prot and TrEMBL databases and other sequence collections. Close to 60% of all proteins have at least one match, and one half of all residues are covered by assignments. All models and full results are available for download and online browsing at http://supfam.org. Users can study the distribution of their superfamily of interest across all completely sequenced genomes, investigate with which other superfamilies it combines and retrieve proteins in which it occurs. Alternatively, concentrating on a particular genome as a whole, it is possible first, to find out its superfamily composition, and secondly, to compare it with that of other genomes to detect superfamilies that are over- or under-represented. In addition, the webserver provides the following standard services: sequence search; keyword search for genomes, superfamilies and sequence identifiers; and multiple alignment of genomic, PDB and custom sequences.

INTRODUCTION

Here we give an up-to-date overview of the SUPERFAMILY database, and describe in detail a number of significant developments since the first publication of the method (1) and a subsequent database article (2). The first two sections provide background information for those who have no previous knowledge of the database; the remainder is devoted to new features.

The SUPERFAMILY database is based on the SCOP classification of protein domains (3). SCOP defines domains as independent evolutionary units of protein structure that either: occur on their own (an entire protein consisting of a single domain); combine with a group of domains that also occur on their own; or combine with at least two different domains in two separate proteins. SCOP then progressively groups domains of known 3D structure according to the nature of their similarity (sequence, evolutionary and structural). This process results in a hierarchical classification with several levels. Of particular importance to SUPERFAMILY users is the superfamily (or evolutionary) level: SCOP places two domains in the same superfamily if, and only if, they share distinctive features that suggest a common evolutionary ancestor.

The principal goal of SUPERFAMILY is to identify within protein sequences domains that belong to superfamilies of known structure. To achieve this, SUPERFAMILY uses expertly built profile Hidden Markov Models (HMMs) (4). Profile HMMs are able to detect more remote homologies (5,6) than more commonly used methods such as PSI-BLAST (7), yet their application is still feasible on a genomic scale. SUPERFAMILY assignments have been carried out on most publically available protein sequences, including all sequences in the Swiss-Prot and TrEMBL databases (8) and predicted proteins from all completely sequenced genomes. All SUPERFAMILY profile HMMs and results are available for download.

THE SUPERFAMILY DATABASE

The database consists of three main components: a library of profile HMMs that represent all proteins of known structure; a collection of assignments to predicted proteins from all completely sequenced genomes and several databases of protein sequences; and a suite of services and tools, available either online or for download from our webserver. This section describes in turn all three components.

Model library

The library of profile HMMs lies at the core of the database. Each model corresponds to a protein domain and aims to represent an entire SCOP superfamily. With each release, new

*To whom correspondence should be addressed. Tel: +44 1223 402479; Fax: +44 1223 213556; Email: mm238@mrc-lmb.cam.ac.uk

Table 1. Release 1.63 statistics

Release date	August 2003	
Number of superfamilies	1232	
Number of models	7924	
Number of completely sequenced genomes—total	128 (+26 strains)	
—eukaryotic	16 (+one strain)	
—archaebacterial	17 (no strains)	
—eubacterial	95 (+25 strains)	
Genomes/database	Proteins with at least one match (%)	Amino acid coverage (%)
Eukaryotes	57	44
Archaebacteria	61	53
Eubacteria	61	54
Swiss-Prot	74	61
TrEMBL	64	53

In the 3 years since the original publication (1), the number of genomes has more than doubled (from 56 to 128), the number of superfamilies in SCOP has grown by almost 50% (from 859 in release 1.53 to 1232 in release 1.63) and the percentage of proteins matched and the amino acid coverage have both increased by 10% (from 49 to 59% and from 39 to 49%, respectively).

models are added using a previously described procedure (1) to make sure that all superfamilies in SCOP classes a–g are covered by the library. All models are also updated with hits to the latest version of the NCBI non-redundant database and our collection of predicted proteins from completely sequenced genomes. The library in a variety of formats is available for download from the webserver along with a program for carrying out the assignment procedure (see the next section for details).

Genome assignments

Using TimeLogic DeCypher hardware, the library has been used to carry out assignments to predicted proteins from all completely sequenced genomes, the Swiss-Prot and TrEMBL databases (8) and other sequence collections (see Table 1 for details). The assignments are kept up to date with additions and improvements in the model library, changes in protein predictions and new releases of sequence databases. We estimate that the error rate of our assignments is <1%. For the purpose of large-scale genome analysis this is an acceptable level, but when examining individual cases in detail the confidence score should be taken into account. The complete results including alignments are available from the webserver as either individual web pages for online browsing, flat files or MySQL dumps for bulk download, or via a Distrubed Annotation Server (DAS, see below).

Tools and services available from the webserver

In addition to the download facilities mentioned above, the webserver at http://supfam.org provides the following services: sequence search for both amino acid and nucleotide queries; a page for viewing multiple alignments of genomic, PDB (9) and custom sequences; keyword search for models, superfamilies, organisms and individual sequences; a collection of web pages for analysis of whole-genome results; and a number of other features described below.

NEW FEATURES

Domain architectures

We have parsed all SUPERFAMILY genome assignments into simple strings that for each protein give the N-to-C sequence of its domains. The strings, which we call domain architectures, are analogous to protein sequences but the alphabet consists of SCOP superfamilies rather than amino acids. The parsing algorithm is described in detail on our web page (http://supfam.org/SUPERFAMILY/comb.html) and will be published elsewhere (C. Vogel, manuscript in preparation). The resultant data, suitable for bioinformatics research (10–12), can be downloaded as part of the relational database.

Several new tools on the web interface make use of domain architectures. Starting from a superfamily of interest, users can find out in which architectures it occurs, and for each architecture then determine the proteins that exhibit it. It is believed that multi-domain proteins that share the same architecture have the same or related function (13).

We have also added a page that for each genome lists all pairs of superfamilies that occur next to each other in its domain architectures. An example of the resulting network of combinations is shown in Fig. 1. For each pair it is again possible to determine the architectures that contain it, and for each architecture all the proteins.

Users can also remove from the initial list those pairs that are already present in proteins of known structure or which also occur in other genomes, and thereby obtain combinations whose structure is not known or which are unique to a given genome. These proteins are likely to have novel functions which may be mediated by the domain–domain interfaces, and thus present suitable targets for structural genomics.

Unusualness and comparative genomics

It is now possible to compare the domain composition of a given genome with that of other genomes and thereby detect superfamilies that are over- or under-represented. The group for comparison can consist of several predefined choices, such as eukaryotes or archaebacteria, or any user-defined set of genomes (or a single genome), e.g. other strains of the same species.

Over-represented superfamilies have typically expanded as the organism specialized for its environmental niche; e.g. in *Shewanella oneidensis*, a Gram-negative bacterium with diverse respiratory strategies that are of potential use in bioremediation (14), the five most unusual superfamilies include multiheme cytochromes, porins and transferrins. Proteins in these superfamilies may provide interesting targets for investigation.

Experimental profile–profile search using PRC

The current SUPERFAMILY procedure relies on comparisons of a query sequence with profile HMMs in our library. Recent work (15,16) has suggested that significant improvements in detection of remote homologs (and presumably also in alignment quality) can be obtained by collecting homologs of the query sequence, constructing a profile (or profile HMM) from their alignment, and comparing this profile (rather than the initial sequence) with the library.

We are in the process of developing a program for comparison of two profile HMMs called PRC. An option to

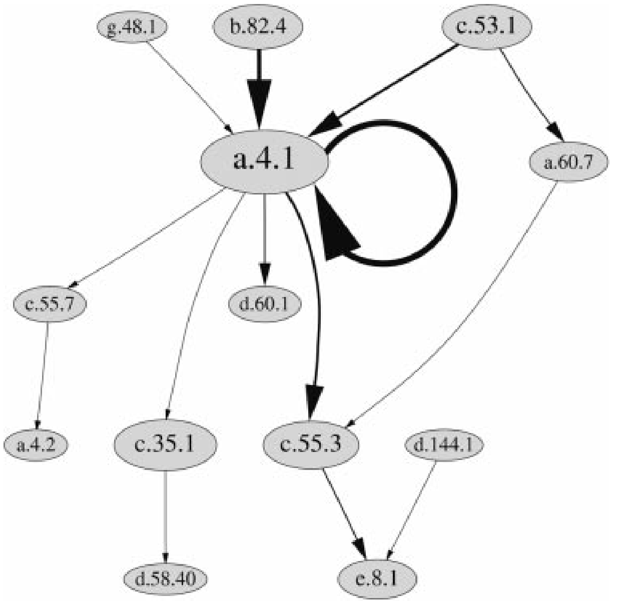

Figure 1. A section of the domain combination network in *E.coli* K-12. Nodes represent superfamilies labelled according to their SCOP (3) classification (see below for legend), edges indicate superfamilies that occur next to each other in domain architectures, and arrows show the N-to-C order. Node size and edge thickness are proportional to the logarithm of the number of proteins. All edges between the selected superfamilies are shown. The presence of edges in only one of the two possible directions illustrates the tendency of adjacent domains to appear in one N-to-C order (10,11). This and other visualizations are available online from the webserver. The superfamilies shown in the figure are: a.4.1, homeodomain-like; a.4.2, Methylated DNA–protein cysteine methyltransferase, C-terminal domain; a.60.7, 5′–3′ exonuclease, C-terminal subdomain; b.82.4, regulatory protein AraC; c.35.1, phosphosugar isomerase; c.53.1, resolvase-like; c.55.3, ribonuclease H-like; c.55.7, methylated DNA–protein cysteine methyltransferase domain; d.58.40, D-ribose-5-phosphate isomerase (RpiA), lid domain; d.60.1, probable bacterial effector binding domain; d.144.1, protein kinase-like (PK-like); e.8.1, DNA/RNA polymerases; and g.48.1, Ada DNA repair protein, N-terminal domain (N-Ada 10).

use this program appears on the results page in cases where the standard SUPERFAMILY search finds no significant hits. The program is used in conjunction with the SAM T99 procedure (17), which generates alignments of homologs from single-sequence inputs. PRC source code and binaries are available for download from http://supfam.org/PRC under the GNU General Public Licence.

Profile HMM diagrams

Each model now has a home page with a simple diagram that shows its principal features, such as amino acid composition, strongly conserved sites, hydrophobicity and regions in which insertions and deletions are common. A typical representation is shown in Fig. 2. Software used to create the diagrams is available for download from the webserver.

Model library available in HMMER and PSI-BLAST formats

The model library is now available for download in HMMER (4) and PSI-BLAST (7) formats in addition to the recommended SAM (17) format, along with a program for carrying out the assignment procedure using the SAM and HMMER packages. The PSI-BLAST binary format is architecture dependent and our library only works on x86 and Alpha machines. The coverage of SAM and HMMER versions of the library is comparable (6), but the PSI-BLAST version detects ~15% fewer remote homologs [in a SCOP all-against-all test (6), unpublished results]. The program used to convert between the formats is also available.

Distributed annotation server (DAS)

All SUPERFAMILY genome assignments are available via a protein DAS server (see http://biodas.org for more information). High-traffic genome servers and individual users alike are invited to use this interface as a preferred way of staying up to date with changes in SUPERFAMILY annotations.

INTERPRO CONSORTIUM

SUPERFAMILY became a member database of the InterPro Consortium (18) in July 2003 (InterPro release 7.0). Starting with this release, users can run the SUPERFAMILY assignment procedure as part of InterProScan (19). SUPERFAMILY assignments to Swiss-Prot and TrEMBL (8) are also available from the InterPro website (http://www.ebi.ac.uk/InterPro) along with annotation from other member databases. However, only 468 out of the 1232 superfamilies in SCOP 1.63 were integrated into InterPro as of the 7.0 release; both InterProScan and the InterPro website are restricted to these superfamilies. Work is underway to incorporate the rest.

Annotation of SCOP superfamilies

As part of the integration process the InterPro team are annotating SCOP superfamilies. Each superfamily is described in a short abstract that includes references to relevant literature and, wherever possible, an outline of its function. In a separate but related project, Gene Ontology (20) terms are being assigned to an increasing number of superfamilies.

To our knowledge this is the first attempt to provide such information for SCOP superfamilies and should be of benefit to all SCOP users. The annotations can be accessed from SUPERFAMILY via the InterPro link on our web pages for individual superfamilies.

CHANGE IN LICENSING

Most licensing restrictions including the fee for commercial users have been abolished, making use of the database free for all. Access to the download site is granted immediately upon completion of a registration form.

FUTURE WORK

We are planning two major improvements to SUPERFAMILY. The first, already alluded to above, is a change in the underlying method from profile–sequence to profile–profile. Once PRC (our program for comparison of profile HMMs, see above) has reached a stable release, we intend to apply the method to all completely sequenced genomes. We are hoping that this will bring the coverage of our genomic assignments to a level comparable to the best fold recognition servers, while retaining the ability to handle multi-domain proteins.

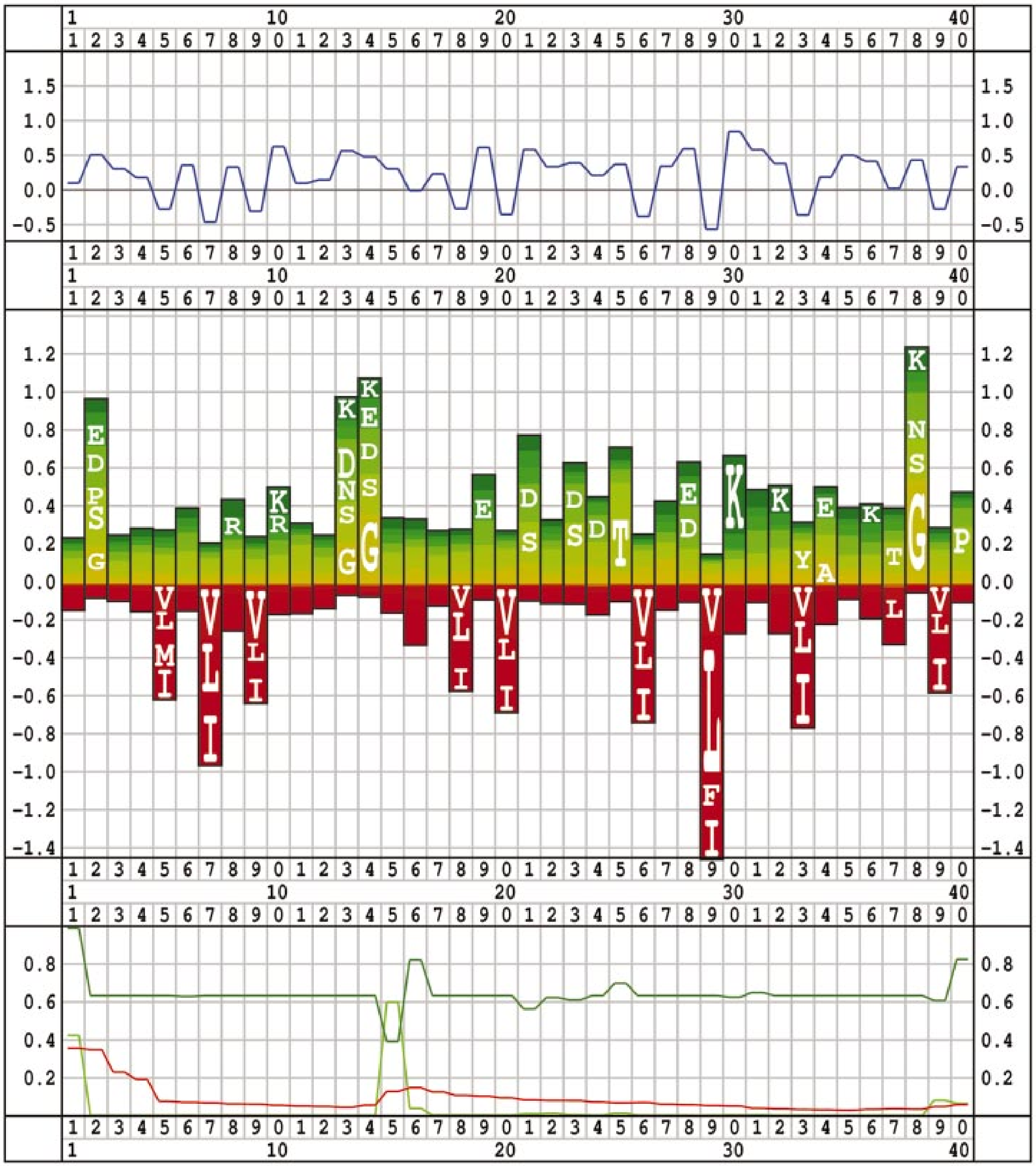

Figure 2. An example of a model diagram, for model 0013580 from the ubiquitin-like superfamily. The top plot (blue line) is the average hydrophobicity, calculated as the sum over all amino acids of match emission probability times $\Delta G_{\text{surface-buried}}$ (in kcal/mol). The middle plot shows match emission probabilities. The amino acids in each column are ordered from most hydrophilic (top) to most hydrophobic (bottom). The size of each column is proportional to the difference between the match emission distribution and the generic background distribution. The columns are partitioned between amino acids according to the ratio of their probabilities; only letters larger than a threshold size are shown. The columns are aligned at the bottom of A (alanine). The bottom plot gives the probability that there is an insertion (light green) or a deletion (red) at each position in the HMM. The dark green curve gives the probability P of an insert–insert transition; assuming there is an insertion at that node, $1/(1-P)$ gives its expected length. The secondary structure of the fragment is readily apparent from the graph: two β sheets (periodicity two) followed by a helix (periodicity three and four).

Secondly, we are developing a procedure that will allow us to identify the SCOP family of a query domain in addition to its superfamily. Because many superfamilies are very divergent functionally, identification of the precise function of a particular domain is often difficult based on its superfamily assignment alone. We believe that family-level assignments should provide a much more fine-grained picture.

ACKNOWLEDGEMENTS

We are grateful to Sarah Teichmann for comments on the manuscript. Martin Madera is supported by a Trinity College Internal Graduate Studentship, Christine Vogel has a predoctoral Fellowship from the Boehringer Ingelheim Fonds and Sarah Kummerfeld has a Studentship from the Laboratory of

Molecular Biology combined with a University of Sydney Travelling Scholarship.

REFERENCES

1. 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.
2. Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.
3. Murzin,A.G., Brenner,S. E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
4. Eddy,S.R. (1998) Profile Hidden Markov Models. *Bioinformatics*, **14**, 755–763.
5. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
6. Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.
7. Schaefer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
8. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
10. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
11. Bashton,M. and Chothia,C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.*, **315**, 927–939.
12. Chothia,C., Gough,J., Vogel,C. and Teichmann,S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.
13. Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.
14. Heidelberg,J.F., Paulsen,I.T., Nelson,K.E., Gaidos,E.J., Nelson,W.C., Read,T.D., Eisen,J.A., Seshadri,R., Ward,N., Methe,B. *et al.* (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.*, **20**, 1118–1123.
15. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
16. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
17. Karplus,K., Barrett,C. and Hughey,R. (1999) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
18. The InterPro Consortium (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
19. Zdobnov,E. M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
20. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.