D380–D386 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 26 November 2008*
doi:10.1093/nar/gkn762

SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny

Derek Wilson[1,*], Ralph Pethica[2], Yiduo Zhou[2], Charles Talbot[2], Christine Vogel[3], Martin Madera[2], Cyrus Chothia[1] and Julian Gough[2]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, [2]Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol BS8 1UB, UK and [3]Institute for Cellular and Molecular Biology University of Texas at Austin 2500 Speedway, MBB 3.210 Austin, TX 78712 USA

Received September 15, 2008; Revised October 5, 2008; Accepted October 6, 2008

ABSTRACT

SUPERFAMILY provides structural, functional and evolutionary information for proteins from all completely sequenced genomes, and large sequence collections such as UniProt. Protein domain assignments for over 900 genomes are included in the database, which can be accessed at http://supfam.org/. Hidden Markov models based on Structural Classification of Proteins (SCOP) domain definitions at the superfamily level are used to provide structural annotation. We recently produced a new model library based on SCOP 1.73. Family level assignments are also available. From the web site users can submit sequences for SCOP domain classification; search for keywords such as superfamilies, families, organism names, models and sequence identifiers; find over- and underrepresented families or superfamilies within a genome relative to other genomes or groups of genomes; compare domain architectures across selections of genomes and finally build multiple sequence alignments between Protein Data Bank (PDB), genomic and custom sequences. Recent extensions to the database include InterPro abstracts and Gene Ontology terms for superfamiles, taxonomic visualization of the distribution of families across the tree of life, searches for functionally similar domain architectures and phylogenetic trees. The database, models and associated scripts are available for download from the ftp site.

INTRODUCTION

All of the SUPERFAMILY protein assignments are available via the web site (http://supfam.org/) and for download in a number of different formats. A variety of navigation and analysis methods are provided on the web site. Recently we have added new content such as InterPro (1) abstracts and Gene Ontology (GO) (2) terms, and new functions such as visualization of the distribution of protein domains across the major kingdoms of life, suggested domain architectures with similar genomic distribution plus phylogenetic trees.

Below we will first describe what you can do with the SUPERFAMILY database and web site. This is followed by a summary of what other groups use SUPERFAMILY for. Then we will describe what is new in SUPERFAMILY. Finally, we will outline our future plans for the database. Note that there is a detailed discussion of the model building procedure and family assignment method in Supplementary Material 1.

WHAT CAN YOU USE SUPERFAMILY FOR?

Sequence search

One of the most commonly used features on the SUPERFAMILY web site is the Sequence search, which provides Structural Classification of Proteins (SCOP) (3) domain annotation for user submitted protein or DNA sequences. Users can submit up to 20 sequences. We are happy to run larger sequence sets by request. These sequences will be checked against the database of existing domain assignments. If there is no existing assignment, then the sequences will be compared with the SCOP domain sequences from ASTRAL (4) using BLAST (5). If no domains can be assigned at this stage, then domains are searched for using the hidden Markov model (HMM) library and the Sequencing and Alignment Modeling (SAM) (6) package.

The model library is at the core of the SUPERFAMILY procedure. On average, the models produce significant domain assignments for ~60% of sequences. Domain coverage can be increased through the use of profile–profile methods (7). These methods find remote homologues by collecting and aligning homologues of a query sequence,

*To whom correspondence should be addressed. Tel: +44 1223 402479; Fax: +44 1223 213556; Email: dw@mrc-1mb.cam.ac.uk

building a HMM (or profile) from the alignment and comparing this HMM with the SUPERFAMILY model library. PRC, the profile comparer (http://supfam.org/PRC/), is available as an option when the SUPERFAMILY procedure fails to find any significant domains.

Keyword search

The contents of the database can be searched using the simple Keyword search form, which allows searches for SCOP identifiers or superfamily/family names, protein sequence identifiers, Linnaean organism names (plus common names), PDB (8) identifiers and SUPERFAMILY model identifiers.

Accessing over 900 organisms

Newly sequenced genomes are continuously being added to SUPERFAMILY. New genomes can be added, or existing genomes updated by request. Two navigation methods are available for browsing all the genomes.

The first method presents the genomes in a human-centric taxonomically ordered hierarchy. The hierarchy is a simplified version of the complete lineage available from the National Center for Biotechnology Information (NCBI) taxonomy (9). There are three taxonomic categories in this hierarchy plus the final organism level. The three categories of the hierarchy are chosen to result in a reasonable number of genomes in the final category. For example, the *Homo sapiens* genome occurs in the Animals > Vertebrates > Mammals categories, along with 24 other mammals.

The second navigation method also uses a taxonomically ordered list of genomes by default, but presents a dozen statistical values which can be sorted on. For each genome, the statistical values provided are: number of proteins in the genome, number of proteins with at least one domain assignment, percentage of proteins with at least one assignment, protein sequence domain coverage as a percentage, total number of domains assigned, number of unique superfamilies assigned, number of unique families assigned, average superfamily size, percentage of domains produced by duplication, average sequence length, average length matched and number of unique domain pairs.

We put considerable effort into classifying new genomes as either model type or strain type. Our aim here is to reduce potential bias from bacteria such as *Staphylococcus aureus* and *Streptococcus pyogenes* for which numerous, mostly pathogenic, strains have been sequenced. For the higher eukaryotes from the Ensembl (10) resource, we provide two versions of each genome. The main version of the genome provides protein sequences for 'all transcripts'. The second version contains protein sequences for the 'longest transcripts' only. The 'longest transcript' versions cater to users who wish to eliminate bias from alternative splice forms.

Comparing genomes, domains and domain architectures

The highlight of the comparative genomics tools is the unusual superfamilies/families functionality. By default, these web pages list the over- and underrepresented families, or superfamilies, in each genome relative to the superkingdom the genome belongs to. For instance, the L domain-like superfamily from the leucine-rich repeat fold, is one of the most overrepresented superfamilies from the cellular slime mould *Dictyostelium discoideum* (relative to all other eukaryote genomes). Tandem repeats of trinucleotides are abundant in dictyostelids (11). The genome to taxonomic groups comparison is not limited to the three superkingdoms. Genomes can be compared with arbitrary taxonomic groups, so *D. discoideum* could be compared with the other Amoebozoa (currently comprising *Dictyostelium purpureum* and *Entamoeba histolytica*). Similarly, all three of the Amoebozoa genomes could be compared with all the genomes in the closest eukaryotic kingdoms, fungi and metazoa. There is a new dedicated page for the production of custom lists of genomes which can be used in these comparisons.

A domain architecture (N- to C-terminal arrangement of domains) is provided for every protein with significant matches to one, or more, of the models. The generation of these domain architectures is described here: http://supfam.org/SUPERFAMILY/comb.html. A number of tools are supplied for the comparison and analysis of domain architectures in and across genomes. They range from the simple, such as unique domains in genomes, to the complex, like co-occurrence networks of domain architectures across all genomes. Starting from the domain architecture of a particular protein of interest, one can get a list of the proteins with the same architecture and then compare these domain architectures visually.

Some domains occur next to each other on domain architectures more commonly than others (12,13). So for each genome, we list all domain pairs that occur next to each other. The resulting directed network of adjacent domain pairs can be visualized. Nodes represent domains, and edges link domains which form an adjacent pair. In addition, undirected domain architecture co-occurrence networks can be rendered for domain architectures containing a particular superfamily of interest. Nodes in these networks are genomes, and edges between nodes represent the presence of domain architectures which contain the superfamily of interest in both genomes.

Alignments

For each domain assignment, alignments between the sequence containing the domain and the seed sequence of the model providing the domain assignment are supplied. Additional sequences can be added to these alignments. Sources of additional sequences include genomic sequences also containing hits to the original model, sequences from the PDB and user supplied sequences. A page with detailed statistics on each alignment can also be viewed.

Functional annotation of domain superfamilies

Christine Vogel has manually annotated all SCOP superfamilies (13–16) with respect to their usual role in a protein, in a particular pathway or in the cell. The annotation scheme used classifies each of the superfamiles into one of 50 detailed functional categories, which map to seven general functional categories. The general categories

of function are: (i) information: storage, maintenance of the genetic code, DNA replication/repair, general transcription/translation; (ii) regulation: regulation of gene expression and protein activity, information processing in response to environmental input, signal transduction, general regulatory or receptor activity; (iii) metabolism: anabolic and catabolic processes, cell maintenance/homeostasis, secondary metabolism; (iv) intra-cellular processes, cell motility/division, cell death, intra-cellular transport, secretion; (v) extra-cellular processes: inter- and extra-cellular processes (e.g. cell adhesion), organismal processes (e.g. blood clotting), immune system; (vi) general: general and multiple functions, interactions with proteins/ions/lipids/small molecules; and (vii) other/unknown: unknown function, viral proteins/toxins. This functional annotation has been applied to overrepresented domain combinations (13), domain recombination (14) and protein family expansions in relation to biological complexity (13).

Data availability

All of the domain assignments, models and scripts are immediately available from the ftp site after registering for a SUPERFAMILY license, which is free for academic or commercial use. The domain assignments and associated data are provided in both flat and relational (MySQL) formats. The database schema can be reviewed in Supplementary Figure 1. The model library can be downloaded in SAM (6), HMMER (17) and PSI-BLAST (18) formats. Requests for custom datasets can be accommodated by contacting the authors.

WHAT DO OTHER GROUPS USE SUPERFAMILY FOR?

Some recent work by other groups which utilized SUPERFAMILY include: experimental verification of protein function (19), studies of individual families (20,21) and protein complexes (22). Wang *et al.* (23) used the SUPERFAMILY domain assignments to produce a chronology of the SCOP folds and superfamilies. The SUPERFAMILY models have been used in benchmarking various algorithms, including most recently remote homology techniques (24).

SUPERFAMILY has been integrated into several other biological databases. InterPro (1) have integrated the HMM library at the heart of SUPERFAMILY and annotated SCOP superfamilies. Likewise, the DBD database (25) uses a curated subset of the HMMs for the prediction of sequence-specific transcription factors. For every major update of The Arabidopsis Information Resource (TAIR) (26) we provide a new set of domain assignments for incorporation. Meta II (27) is a web server for protein structure and function prediction. It will poll numerous protein structure and function tools and databases, including SUPERFAMILY sequence classification, and return a single combined set of results.

The Ensembl (10) eukaryotic genome resource render our domain annotation on their 'protein view' pages using the lightweight Distributed Annotation Server (DAS) (28) protocol. The DAS is a communication protocol used to exchange biological sequence annotations. All SUPERFAMILY assignments are immediately made available through our DAS server, which offers the opportunity to stay up to date with the SUPERFAMILY assignments without downloading and installing database dumps. The SUPERFAMILY DAS server provides protein domain assignment details in XML format for use by DAS clients such as gbrowse [the GMOD genome browser (29)] and DASTY (http://www.ebi.ac.uk/dasty/).

WHAT IS NEW IN SUPERFAMILY?

The most important update is the recent release of a new model library based on SCOP 1.73. SCOP have added 238 superfamilies and 619 families since the previous model library release. The SUPERFAMILY database endeavours to provide domain assignments for all completely sequenced genomes. Since the previous article (15) appeared 650 genomes have been added or updated. We anticipate including over 1000 genomes before the end of 2008. New database content such as InterPro abstracts and GO terms have been incorporated, and three major new features integrated.

InterPro abstracts and GO terms

The InterPro consortium integrates databases of protein families, domains and functional sites, including: Gene3D (30), PANTHER (31), Pfam (32), PIRSF (33), PRINTS (34), ProDom (35), PROSITE (36), SMART (37), SUPERFAMILY and TIGRFAMs (38). The InterPro annotation of families includes short name, name, abstract, GO terms and cross-references to specialized databases and protein structural information. The InterPro abstract for a typical protein domain is a detailed description (complete with literature references) of its biological function, mode of operation and an account of its structure. InterPro have added abstracts for 1079 SCOP 1.69 superfamilies. We have integrated these abstracts into the SUPERFAMILY web site to provide additional information describing superfamilies.

The GO consortium develop and apply three controlled biological vocabularies (ontologies): molecular function, cellular component and biological process. These ontologies are designed to avoid variations in terminology and describe gene and gene product attributes in a mostly species-independent manner. Thus, allowing uniform queries across different databases which have integrated the ontologies. Each GO entry, or term, is assigned to one ontology, and has a unique identifier plus a term name. For example, GO:0006352 and 'transcription initiation' from the biological process ontology has been assigned to the 'Sigma3 and sigma4 domains of RNA polymerase sigma factors' superfamily. The GO is widely used in genome annotation and functional association studies. Again, InterPro has assigned GO terms to superfamiles. A GO to SCOP superfamily mapping (covering 763 superfamilies) is available for download.

Figure 1. TaxViz displays the distribution of domains across the major taxonomic kingdoms, and organisms within each kingdom. Shown here is the distribution of the P-loop containing nucleoside triphosphate hydrolase domains. Each circle or node, represents the features of a single taxonomic group or individual organism. The nodes are arranged hierarchically in concentric rings. The higher taxonomic groups (superkingdoms: Eukaryota, Bacteria and Archaea), located in the centre, lead recursively outwards towards their children (the kingdoms or phyla within each superkingdom). For taxonomic groups, the size of the node increases logarithmically with the mean number of domains found per organism in the taxonomic group. The distribution of domains in individual species can be navigated to using the outer nodes. There are three specialized nodes which display the distribution of domains in (i) selected model organisms; (ii) organisms containing the maximum number of domains; and (iii) organisms containing the minimum number of domains.

TaxViz—Taxonomic Visualization of protein domains

TaxViz displays the distribution of domains across the major taxonomic kingdoms, and organisms within each kingdom, illustrated in Figure 1. The graphics produced by this tool could benefit researchers interested in gaining an insight into the taxonomic distribution of particular families or superfamilies. Each circle, or node, represents the features of a single taxonomic group or organism. The nodes are arranged hierarchically in concentric rings. The highest taxonomic groups are located in the centre, and lead recursively outwards towards their children. The name of the taxonomic group, and the mean number of domains per organism are displayed alongside the taxonomic nodes. For taxonomic groups, the size of the node increases logarithmically with the mean number of domains found per organism in the taxonomic group. For organisms, the size of the node is simply the number of domains in that organism.

There are five types of TaxViz nodes: Overall, Domain, Kingdom, Subkingdom and Species. The 'Overall' node displays the mean number of domains per organism across the entire tree of life. The Overall node is surrounded by the three 'Domain' (or superkingdom) nodes: Eukaryota, Bacteria and Archaea. The Domain nodes (which should not be confused with protein domains) are surrounded by the 'Kingdom' nodes, for example, metazoa from the Eukaryota Kingdom. The largest Kingdoms (in terms of fully sequenced organisms), contain 'Subkingdom' nodes. The current list of largest Kingdoms follows: Metazoa, Euryarchaeota, Proteobacteria,

Actinobacteria and Firmicutes. The remaining Kingdoms, and the Subkingdoms, contain 'Species' nodes. Note that the Species nodes do not include any of the organisms we have classified as strains. An example may help illustrate the hierarchy of taxonomic groups used in TaxViz. The *H. sapiens* (human) species occurs within the Subkingdom Mammalia, Kingdom Metazoa and Domain Eukaryota.

The organisms within a taxonomic group can be viewed by clicking on the outer (Kingdom or Subkingdom) taxonomic group node. The name of the organism, and the number of proteins in the organism which contain the domain are displayed alongside the organism nodes. From here, you can view the domain assignment details of each organism by clicking on an organism node. The colour of the node represents the superkingdom to which it belongs, and is detailed in the legend on each graphic. Finally, there are three specialized nodes which link to the distribution of domains in (i) selected model organisms, (ii) organisms containing the maximum number of domains and (iii) organisms containing the minimum number of domains.

Domain architectures with similar genomic distribution

We have added a tool to find functionally similar proteins. The functions of the majority of proteins have yet to be experimentally characterized. The most commonly used methods for functional annotation are sequence-based comparisons to proteins that are well characterized. The large average size of eukaryotic multi-gene families means they are susceptible to orthology mis-assignment (39).

Our approach compares two domain architectures: the domain architecture of interest and a 'query' domain architecture. The domain architecture of interest is in turn compared with all the other domain architectures in the SUPERFAMILY database. The 10 architectures which are most similar to the architecture of interest are selected for display.

The complete scoring function to assess similarity between domain architecture A and domain architecture B, is

$$Score(A,B) = \frac{1}{N_L}\sum_{i}^{N_s}\left(\frac{\min(A_i,B_i)}{\max(A_i,B_i)} \times I_i \times D_i\right),$$

where N_L is the larger number of genomes in the phylogenetic profiles of domain architecture A or B, and N_s is the number of shared genomes between the two domain architectures, I_i is the information content of genome i and D_i is a distance weighting factor based on the phylogenetic diversity of the genomes sharing both domain architectures. Supplementary Material 2 discusses the similarity function in detail.

Phylogenetic trees

Yang *et al.* (40) use a simple neighbour-joining approach with presence–absence data for 174 organisms and 1294 superfamiles (from a previous SUPERFAMILY database release). They find quite accurate phylogenies for the kingdoms within Archaea, Bacteria and Eukaryota, and 50 common superfamilies across all 174 organisms. These common superfamilies could be present in the last common ancestor of all life.

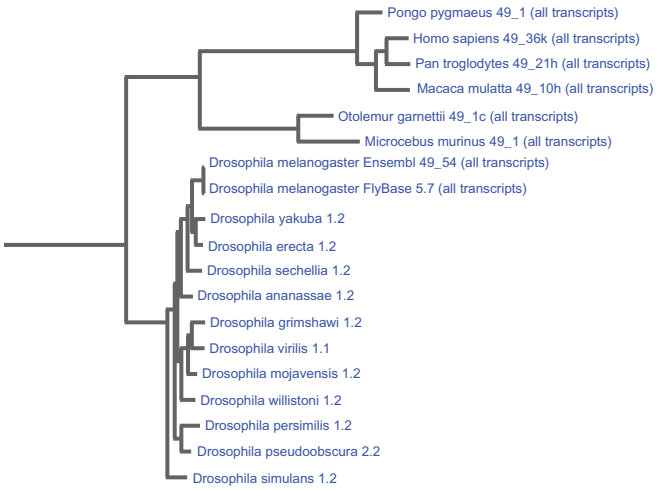

Figure 2. Phylogenetic tree example. Shows all sequenced *Drosophila* species, and Primates plotted on the same tree, using the relationships calculated from all genomes.

A presence/absence matrix can be generated using protein domain architecture data for all genomes in SUPERFAMILY. The PAUP (http://paup.csit.fsu.edu/) software is used to produce a single, large tree topology using both neighbour joining and maximum parsimony methods. Genome combinations or specific clades can be displayed as if individual trees had been produced. However, this data is extracted from the single large tree. This produces a higher quality topology than if the trees had been produced on their own, and allows the trees to be displayed instantly. Figure 2 shows all sequenced Drosophila species, and Primates plotted on the same tree, using the relationships calculated from all genomes. Trees can be plotted as scalable vector graphics (SVG), as well as a number of other formats. The advantages of SVG trees include: can be zoomed and panned, straightforward integration of hyperlinks, simple to search and index. All trees can be downloaded.

Due to time and processing power constraints, it is not possible to use the most exact phylogeny methods to produce the large tree, however, the trees gain accuracy due to the large number of genomes and quality of the data used. The massive rate of genome sequencing means that the tree gains further precision as more genomes are sequenced and added to it, as gaps in phylogenetic space are filled.

FUTURE DEVELOPMENTS

A new model library built against SCOP release 1.73 has recently become available. This library contains 13 920 models representing 1776 superfamilies. We expect this will increase domain assignment coverage. Regenerating assignments for all genomes is a priority. New genomes will continue to be added. Impending full release

eukaryotic genomes include: orangutan, marmoset, wallaby, lamprey, shark, *Amphimedon queenslandica* (a sponge), soybean and maize. InterPro are producing abstracts and GO term annotation for SCOP families. We plan to integrate this information as it becomes available. A new tool for the simultaneous visualization of superfamily and family domains within individual genomes is currently under development. This tool will indicate where families/superfamilies are expanding or contracting relative to higher taxonomic groups.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

European Union Framework Program 7 Impact grant (grant number 213037) and Medical Research Council for open access fees.

Conflict of interest statement. None declared.

REFERENCES

1. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al*. (2008) New developments in the InterPro database. *Nucleic Acids Res*., **35**, D224–D228.
2. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.
3. Andreeva,A., Howorth,D., Chandonia,J.-M., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acid Res*., **36**, D419–D425.
4. 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.
5. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
6. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
7. Madera,M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, **24**, 2630–2631.
8. 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.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res*., **36**, D25–D30.
10. Hubbard,T.J.P, Aken,B.L., Beal,K., Ballester,B. and Caccamo,M. (2007) Ensembl 2007. *Nucleic Acids Res*., **35**, D610–D617.
11. Eichinger,L., Pachebat,J.A., Glöckner,G., Rajandream,M.-A., Sucgang,R., Berriman,M., Song,J., Olsen,R., Szafranski,K., Xu,Q. *et al*. (2005) The genome of the social amoeba Dictyostelium discoideum. *Nature*, **435**, 43–57.
12. Chothia,C., Gough,J., Vogel,C. and Teichmann,S.A (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.
13. Vogel,C., Berzuini,C., Bashton,M., Gough,J. and Teichmann,S.A. (2004) Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol*., **336**, 809–823.
14. Vogel,C., Teichmann,S.A. and Pereira-Leal,J. (2005) The relationship between domain duplication and recombination. *J. Mol. Biol*., **346**, 355–365.
15. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res*., **35**, D308–D313.
16. Vogel,C. and Chothia,C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol*., **2**, e48.
17. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
18. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
19. Brinkrolf,K., Ploger,S., Solle,S., Brune,I., Nentwich,S.S., Huser,A.T., Kalinowski,K., Puhler,A. and Tauch,A. (2008) The LacI/GalR family transcriptional regulator UriR negatively controls uridine utilization of Corynebacterium glutamicum by binding to catabolite-responsive element (cre)-like sequences. *Microbiology*, **154**, 1068–1081.
20. Pereira-Leal,J.B. (2008) The Ypt/Rab family and the evolution of trafficking in fungi. *Traffic*, 9, 27–38.
21. Virel,A. and Backman,L. (2007) A comparative and phylogenetic analysis of the {alpha}-Actinin Rod Domain. *Mol. Biol. Evol*., **24**, 2254–2265.
22. Rasteiro,R. and Pereira-Leal,J.B. (2007) Multiple domain insertions and losses in the evolution of the Rab prenylation complex. *BMC Evol. Biol*., **7**, 140.
23. Wang,M.L., Yafremava,L.S., Caetano-Anollés,D., Mittenthal,J.E. and Caetano-Anollés,G. (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Gen. Res*., **17**, 1572–1585.
24. Loewenstein,Y. and Linial,M. (2008) Connect the dots: exposing hidden protein family connections from the entire sequence tree. *Bioinformatics*, **24**, i193–i199.
25. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD – taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*., **36**, D88–D92.
26. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*., **36**, D1009–D1014.
27. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein Server. *Nucleic Acids Res*., **32**, W321–W326.
28. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinfoormatics*, **2**, 7.
29. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–5610.
30. Yeats,C., Maibaum,M., Marsden,R., Dibley,M., Lee,D., Addou,S. and Orengo,C.A. (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res*., **34**, D281–D284.
31. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O., Campbell,M.J. *et al*. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*., **33**, D284–D288.
32. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.
33. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al*. (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res*, **32**, D112–D114.
34. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al*. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*., **31**, 400–402.

35. Bru,C., Courcelle,E., Carrere,S., Beausse,Y., Dalmar,S. and Kahn,D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.
36. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
37. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
38. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
39. Ranea,J.A.G., Yeats,C., Grant,A. and Orengo,C.A. (2007) Predicting protein function with hierarchical phylogenetic profiles: the gene3D phylo-tuner method applied to eukaryotic genomes. *PLoS Comput. Biol.*, **3**, e237.
40. Yang,S., Doolittle,R.F. and Bourne,P.E. (2005) Phylogeny determined through protein domain content. *Proc. Natl Acad. Sci. USA*, **102**, 373–378.