D414–D418 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 21 November 2007*
doi:10.1093/nar/gkm1019

Gene3D: comprehensive structural and functional annotation of genomes

Corin Yeats[1,*], Jonathan Lees[1], Adam Reid[1], Paul Kellam[1], Nigel Martin[2], Xinhui Liu[1] and Christine Orengo[1]

[1]UCL, Department of Molecular Biology & Biochemistry, Darwin Building, Gower St, London WC1E 6BT and [2]School of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

Received September 14, 2007; Revised October 23, 2007; Accepted October 27, 2007

ABSTRACT

Gene3D provides comprehensive structural and functional annotation of most available protein sequences, including the UniProt, RefSeq and Integr8 resources. The main structural annotation is generated through scanning these sequences against the CATH structural domain database profile-HMM library. CATH is a database of manually derived PDB-based structural domains, placed within a hierarchy reflecting topology, homology and conservation and is able to infer more ancient and divergent homology relationships than sequence-based approaches. This data is supplemented with Pfam-A, other non-domain structural predictions (i.e. coiled coils) and experimental data from UniProt. In order to enhance the investigations possible with this data, we have also incorporated a variety of protein annotation resources, including protein–protein interaction data, GO functional assignments, KEGG pathways, FUNCAT functional descriptions and links to microarray expression data. All of this data can be accessed through a newly re-designed website that has a focus on flexibility and clarity, with searches that can be restricted to a single genome or across the entire sequence database. Currently Gene3D contains over 3.5 million domain assignments for nearly 5 million proteins including 527 completed genomes. This is available at: http://gene3d.biochem.ucl.ac.uk/

INTRODUCTION

The identification of structural domains and their homologous relationships, and hence the domain composition of protein sequences, allows both the practical application of powerful approaches for functional prediction and theoretical investigations into protein structure evolution. Several resources exist to support this field, each bringing a particular perspective: the most comprehensive is InterPro (1)—an amalgamation of resources that links well to the UniProt (2) sequence database. Pfam (3) is the single largest resource and is of interest since it primarily classifies domains through the creation of sequence families. This means that it can be of particular use in functional association studies, though it does miss many ancient evolutionary relationships. Also, and in some ways most similar to Gene3D, there is the Superfamily database (4), which provides the SCOP-derived domain assignments (5) for genomic sequences.

Gene3D is designed to extend the CATH (6) structural domain database from the wwPDB (7) to the protein sequence databases UniProt and RefSeq (8) and the completed genomes defined by Integr8 (9). CATH domains are manually classified following automated analyses of the PDB and assigned a place in the CATH structural hierarchy, reflecting their structural composition and evolutionary relationships. To predict sequence relatives for these CATH domains, sets of hidden Markov Models (HMMs) are generated to represent each CATH superfamily. By scanning these HMM models against the sequence resources and resolving the 'hits', Gene3D v6.0 provides >3.5 million CATH domain assignments for nearly 2.5 million distinct proteins, including 49% of UniProt and 47% of complete genomes. The protocol for doing this has been previously described in (10) and is also illustrated in Supplementary Figure 1.

To further enhance these investigations Gene3D also integrates other domain predictions (i.e. Pfam-A), PDB-based assignments directly from CATH, several function resources [i.e. GO (11)] and protein–protein interaction (PPI) data [i.e. IntAct (12)]. All protein sequences are also clustered into hierarchical protein families to facilitate functional grouping of sequences.

*To whom correspondence should be addressed. Tel: +02076793890; Fax: +02076797193; Email: yeats@biochem.ucl.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

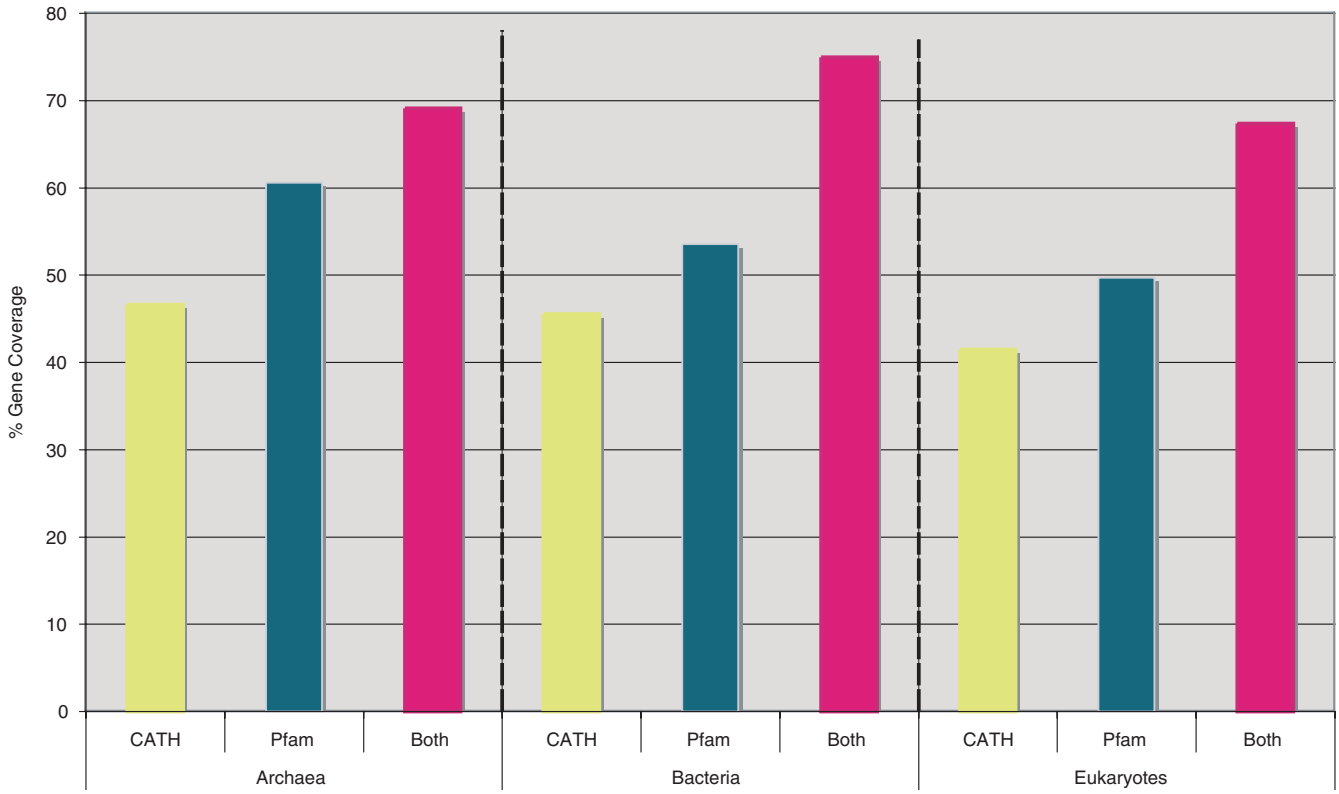

Figure 1. Gene coverage of completed genomes in Gene3D. Shown in this figure are the percentages of genes in bacteria, archaea and eukaryotes that have at least one domain assigned by either (**A**) CATH, (**B**) Pfam or (**C**) both. It should be noted that not all the genomes have been completely scanned with Pfam—hence the coverage is lower than would be expected.

By creating a simple interface between these resources it is now possible to examine in detail functional and evolutionary changes in relation to structures and to enhance functional and structural annotation of genomes. Gene3D has been a significant aid in selecting targets for the Structural Genomics Initiative (13). Over the last two years several new resources have been added and new methods of accessing the data made available. Foremost are the significant improvements made to the usability and functionality of the website, expansion of the pre-made sets available for download and the implementation of a suite of DAS servers using ProServer (14). These developments and their application are described in detail below.

GENE3D V6.0

The September 2007 version of Gene3D contains ~4.5 million distinct proteins, grouped into 190 000 protein families with more than 5 members (method described below)—around 600 000 proteins remain as 'singletons'. Included in this are also 527 species (676 strains)—50 eukaryotes, 437 eubacteria and 39 archaea—totalling ~1.9 million distinct proteins. See Figure 1 for the coverage of these genomes with CATH and Pfam domains. All the HMM-identified domains assigned to the 2046 CATH v3.1.0 superfamilies are sub-clustered at ten discrete sequence identity levels, ranging from 30–95% (files available for download), so as to aid accurate function transfer. For further details on additional annotation, including Pfam, low complexity regions, coiled coils, transmembrane helices, see Supplementary Table 2.

Functional data is represented through the inclusion of the GO, KEGG (15), COGs (16) and FunCat (17) datasets. Gene3D also has protein–protein interaction (PPI) data sourced from IntAct, MINT (18) and manually curated high-quality interactions from MPact (19) PPI datasets and where possible proteins are linked to expression data at ArrayExpress (20). For a complete list of imported resources, see Supplementary Table 3. We also aim to import and enable the use of as many different types of identifier as possible: currently the website can be queried with more than 35 million distinct identifiers sourced from UniProt, CATH, Pfam, the wwPDB, RefSeq, COGs, OMIM (21), BioThesaurus (22) and more (see Supplementary Table 1).

Changes to structural data

We have incorporated the manually curated PDB-based CATH v3.1.0 domain assignments (88 774 out of 93 885) by mapping them to UniProt using the procedure described by Andrew Martin (23). Multi-Domain Architectures (MDAs) were fully resolved for 7591 proteins out of 8646 possible. The resolution is carried out very conservatively and if any mapping problems between PDB and UniProt are identified the MDA is not calculated.

Hence, this set can be considered a gold standard for structural annotation of UniProt.

CATH have also added 'unassigned' domains to their structural library. These are domains that have been identified within newly determined multidomain structures, but not yet classified in the CATH hierarchy. We also scan HMMs based on these to extend the possible structural coverage. In Gene3D v6.0, these add ~250 000 domain assignments to ~160 000 proteins.

The UniProt protein files are also a rich source of experimentally determined structural information and we now directly import various features including: signal peptides, active sites, metal-binding sites, splice sites and disulphide bonds. A collaboration with the BIOSAPIENS/ENCODE consortia exploiting this data revealed that 5–20% of human genes produced transcripts that exhibit some form of domain insertion, deletion or substitution whilst still remaining potentially functional (24).

Changes to interaction data

Protein–protein interaction (PPI) data is now sourced from the comprehensive MINT and IntAct resources, as well as the yeast-specific manually curated subset of MPact.

New whole chain families

One of the primary focuses of our research is to extend experimentally derived molecular studies to the vast number of experimentally uncharacterized proteins through bioinformatic methods. One of the most powerful means of carrying this out is through reliable functional inheritance between similar proteins. Various studies have shown that knowledge of domain architecture and sequence similarity can enable reliable transference of functional annotation (25,26). To enhance these approaches every protein in the database is assigned to a family based on sequence similarity. A novel approach has been employed that takes advantage of the fast affinity propagation clustering (APC) algorithm (27) and the comprehensive protein sequence similarity database SIMAP (28).

The Gene3D clustering protocol consists of several steps that aim to break down the problem of clustering 4.6 million sequences. The sequence database is clustered repeatedly, currently with fairly conservative thresholds (*E*-value 0.001, overlap length 80%), using the cd-hit (29) program and a mixture of single-linkage, multi-linkage and APC clustering. Ultimately, each derived cluster is subclustered at 10 levels of sequence identity. This quicker process should allow for improved benchmarking and analysis. For full details of the process, see Supplementary Data.

New identifier mappings

As mentioned above, we import and map identifiers from many new resources. This allows improved querying of and linking to Gene3D, as well as correlating disparate datasets. The new identifiers include SGD (Saccharomyces Genome Database) and most of those in the BioThesaurus database (i.e. OMIM and Ensembl identifiers). If an identifier maps to multiple proteins—either because several genes have been given the same name or because the identifier corresponds to a family—then all proteins are returned allowing the user to choose the one(s) of interest. However, as described below there are ways of refining the query or specifying particular (i.e. functional) subsets.

Hierarchical phylogenetic domain profiles—PhyloTuner

As mentioned above, all CATH superfamilies are subclustered at ten levels of sequence similarity. From these clusters phylogenetic profiles at each level of similarity are generated for the complete genomes from Integr8. By using the actual copy number of occurrences of each domain at each identity level it is possible to identify co-evolving domain families or subfamilies, as exemplified by the PhyloTuner approach developed by Ranea *et al.* (30). These profiles are provided on the FTP site, linked to at the top of the website pages. The advantage of this type of approach for functional prediction is that it can detect entirely novel associations that have no previous experimental evidence and hence guide the discovery of new knowledge rather than extending what is already known.

SIMPLER, MORE POWERFUL WEBSITE

Improved interface

Considerable effort has been put into improving the usability of the website and query results are now presented more swiftly in a clearer format. The primary focus of the site is on searching the database with an identifier term (i.e. CATH superfamily code) and returning the proteins associated with that term (i.e. members of that superfamily). The results are returned as a single page containing a set of selectable tabs for the functional, structural and taxonomic information associated with the query. The main view is common to every query, while the tab content can vary depending on whether the query returns a single protein or multiple proteins. Within the tabs, individual identifiers that can be searched in Gene3D are marked with a twisting arrow tag; clicking on this tag will submit the query.

Sophisticated aided querying

One of the key enhancements has been the construction of a much more sophisticated query tool (Figure 2). It takes the form of a bar across the top of the page containing a series of boxes for entering different options. The first two boxes are for specifying the identifier (i.e. the query term); whilst the identifier type box is defaulted to 'Any' it can be used to pre-allocate the source of the identifier. This allows the removal of ambiguity where a single identifier string is used in different resources—for instance both the CATH-PDB and CATH-HMM assignments use the same codes.

The second two boxes provide a filtering stage, limiting the results to a particular subset. As an example entering 'genomes' and 'human' will limit any returned results to the human genome; entering 'genomes' and 'Mammalia'

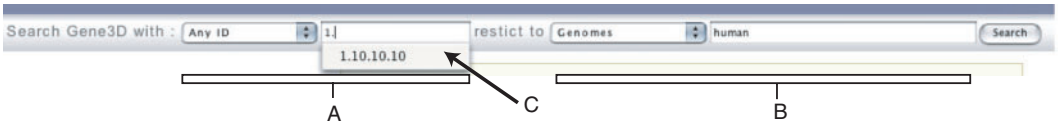

Figure 2. The Gene3D search bar. This bar can be found at the top of all the Gene3D pages and is used to navigate the site. It consists of two main components—the query (**A**) and the filter (**B**)—that allow sophisticated data retrieval. Both components also consist of two inputs. (A) The first box describes the identifier type, with the default being any. Different resources often use identical identifier types to represent different proteins or protein families. As a result, the returned data can be ambiguous; users can restrict the identifier to a certain resource to remove ambiguity. The second box accepts the search term. (B) The filter allows the results to be restricted to particular subsets of the database. The first input is the filter type: at the moment 'Genomes', 'GO Term', 'FunCat Category' and 'Affymetrix platform'. The second box accepts the filter term—for instance, 'human', '9606' or 'Mammalia'. (**C**) Possible terms for the query and the filter are shown as a drop-down list while the user types.

will limit the results to completed mammalian genomes. To aid the user, the possible filter terms (i.e. human, man, 9606) are shown as a drop-down list that will refresh itself as the user types. Finally, 'wildcard' matches have also been added, allowing the retrieval of partially matched terms. So for instance, it is now simple to examine the annotation of members of the Ig Fold (CATH code 3.40.50) in humans and then to compare that with mammals in general. Some further example queries are detailed in Supplementary Data.

HMM and BLAST facilities

Whilst Gene3D contains a fairly comprehensive set of protein sequences we have also provided a facility for scanning user-provided sequences against the CATH HMM library with HMMER (31) and a BLAST (32) facility for identifying the most similar protein in the database. These facilities are designed for single sequence submissions, but we can also carry out genome-scale scans if requested.

DAS servers

The DAS servers have been recently re-implemented in ProServer 2 and we now provide four distinct services—for full paths and information on the servers please see the Supplementary Data. In summary, there are two Gene3D-specific servers, one providing the CATH HMM-assigned domain and one Gene3D protein cluster assignment for UniProt proteins. A third server provides the mapping of the CATH structural domains to UniProt, whilst a fourth provides the SPLIT 4.0 transmembrane region predictions. These servers are registered at the DAS registry and as a member of the BioSapiens network we will be actively working to create standards for improving the richness and display of the DAS content, as well as adding new servers.

FUTURE CHANGES

We are constantly improving Gene3D and a comprehensive plan has been initiated to completely redesign the underlying hardware/software architecture. The result of these changes will be manifold. First, it will become much easier and quicker to make partial updates to the site and keep it up-to-date. Second, it will allow more sophisticated retrieval, analysis and display of data in the results tabs. For example, more dynamic family analysis pages will be developed and powerful predictive tools like PhyloTuner incorporated. It will also allow Gene3D to tightly bind itself to CATH releases, ensuring a minimum of delay before CATH PDB domains can be linked to genomic and functional data.

A second major change is that we will be receiving frequent updates for sequence similarity data and CATH HMM scan data from SIMAP allowing us to rapidly expand Gene3D and keep pace with the explosion of protein sequences coming out of sequencing projects. Furthermore, we hope to use the SIMAP Pfam HMM results to expand the Pfam assignments in Gene3D to cover all the sequences, rather than being restricted to those provided in Pfam-A.

Whilst there is already a large set of flat files available for download, it is only a fraction of the possible datasets that can be generated. We are always happy to provide these and to help other teams in utilizing the data for functional prediction, experimental targeting and evolutionary analyses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank EU Biosapiens Network and the Wellcome Trust for providing funding to create the Gene3D resource. We would also like to thank the CATH teams for their help, Juan Antonio Ranea for leading the development of Gene3D-based prediction methods and Thomas Rattei at SIMAP for generously providing the protein similarity data and hosting the CATH HMMs on the SIMAP BOINC system. Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

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. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
2. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.

3. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
4. 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.
5. 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.
6. Greene,L.H., Lewis,T.E., Addou,S., Cuff,A., Dallman,T., Dibley,M., Redfern,O., Pearl,F., Nambudiry,R. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.
7. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
8. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
9. Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K. *et al.* (2005) Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, D297–D302.
10. Lee,D., Grant,A., Marsden,R.L. and Orengo,C. (2005) Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins*, **59**, 603–615.
11. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
12. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
13. Marsden,R.L., Lewis,T.A. and Orengo,C.A. (2007) Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics*, **8**, 86.
14. Finn,R.D., Stalker,J.W., Jackson,D.K., Kulesha,E., Clements,J. and Pettett,R. (2007) ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, **23**, 1568–1570.
15. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
16. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
17. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Guldener,U., Mannhaupt,G. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
18. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
19. Güldener,U., Münsterkötter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stümpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
20. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnikov,N., Lilja,P. *et al.* (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
21. McKusick,V.A. (1998) Mendelian inheritance in man. *A Catalog of Human Genes and Genetic Disorders, 12th edn.* John Hopkins University Press, Baltimore, Maryland.
22. Liu,H., Hu,Z.Z., Zhang,J. and Wu,C. (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.
23. Martin,A.C. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297–4301.
24. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
25. Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.
26. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
27. Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **16**, 972–976.
28. Rattei,T., Arnold,R., Tischler,P., Lindner,D., Stumpflen,V. and Mewes,H.W. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Res.*, **34**, D252–D256.
29. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
30. Ranea,J.A., 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.*, e237.
31. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
32. 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.