Nucleic Acids Research, 2006, Vol. 34, Database issue ***D281–D284***
doi:10.1093/nar/gkj057

Gene3D: modelling protein structure, function and evolution

Corin Yeats*, Michael Maibaum, Russell Marsden, Mark Dibley, David Lee, Sarah Addou and Christine A. Orengo

Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK

Received September 15, 2005; Accepted October 4, 2005

ABSTRACT

The Gene3D release 4 database and web portal (http://cathwww.biochem.ucl.ac.uk:8080/Gene3D) provide a combined structural, functional and evolutionary view of the protein world. It is focussed on providing structural annotation for protein sequences without structural representatives—including the complete proteome sets of over 240 different species. The protein sequences have also been clustered into whole-chain families so as to aid functional prediction. The structural annotation is generated using HMM models based on the CATH domain families; CATH is a repository for manually deduced protein domains. Amongst the changes from the last publication are: the addition of over 100 genomes and the UniProt sequence database, domain data from Pfam, metabolic pathway and functional data from COGs, KEGG and GO, and protein–protein interaction data from MINT and BIND. The website has been rebuilt to allow more sophisticated querying and the data returned is presented in a clearer format with greater functionality. Furthermore, all data can be downloaded in a simple XML format, allowing users to carry out complex investigations at their own computers.

INTRODUCTION

Detailed knowledge of the functional modules that a protein is composed of often allows a more accurate prediction of its function than simply transferring functional information from the most similar annotated sequences. Conversely, grouping protein sequences into families can aid in accurate information transfer when the domain architecture does not provide a specific function. In Gene3D we have attempted to combine both of these approaches in a synergistic manner. To further aid interpretation, we have begun including external sources of high quality functional data [i.e. GO (1)].

One principle function of Gene3D is to map CATH (2) domain families to protein sequences. This is a similar task as that carried out by Superfamily (3) for SCOP (4). It requires a different approach than that for identifying domains within structural data and the steps required to model structural domains and to correctly locate their boundaries within the large sequence databases are not trivial. This process is carried out by Gene3D and we continually look to improve—our recent progress is described in (5)—by exploiting hidden Markov model (HMM) technology. In this release we have extended our predictions to the entire UniProt sequence database (6).

To improve the reliability of functional data transfer between sequences, we have also clustered UniProt into protein families using Tribe-MCL (7). There are several databases supplying either domain family information (8,9) or whole protein family information (10), but Gene3D is the most comprehensive resource to combine both views of the protein world into a unified system. We also provide specific calculations for 240 genomes (as of September 1st 2005) derived from Integr8 (11).

Another major renovation includes the use of the BioMap warehouse (12) to supply other sources of structural data, protein–protein interaction data and various functional annotations, including GO and COGs (13). Finally, we have completely redesigned the website to provide a more intuitive and flexible interface so as to cope with the much richer data we are able to provide.

RECENT DEVELOPMENTS

Gene3D protein families

Information can be more easily transferred between sequences when they belong to the same protein family; i.e. they have a common evolutionary ancestor. We have clustered the

*To whom correspondence should be addressed. Tel: +44 20 7679 3890; Fax: +44 20 7679 7193; 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

~1.8 million sequences in Gene3D into families using Tribe-MCL. The process is described in detail in (14).

There are 203 982 non-singleton families including 56 221 of more than 5 sequences and 556 224 'orphan' sequences. Within the complete genome data only, which consists of 862 886 proteins, there are 80 291 non-singleton families and 212 567 orphan sequences. These numbers will change as we improve our family definitions and new genomes are added. Our estimates suggest that, while 50% of domains found in genomes belong to families common to all kingdoms (universal), only 10% of proteins belong to universal families (14). This suggests the importance of using protein families in conjunction with domain families to accurately predict functions.

The families have also been subclustered by sequence identity at ten different levels from 30% to 100%. This allows greater precision in information transfer and improved understanding of the evolution of the protein family. The families have been refined by ensuring an acceptable similarity in domain composition and sequence length; furthermore, manual examinations are carried out to further improve the accuracy and consistency of these families.

Functional data

In order to provide comprehensive functional annotation we now use the in-house BioMap database (12). BioMap is essentially a warehouse for diverse biological data and contains mappings between several resources and the UniProt sequence database. This provides us both with rich descriptions for each sequence but also provides a strong internal infrastructure, allowing regular updating of the website. These data are linked through representative sequences using the MD5 sequence digest value. Being able to combine data in this manner can be very powerful when analysing the evolution of protein function, as was demonstrated in (15).

The website

The Gene3D website has been completely redeveloped to provide more sophisticated querying capabilities and to be able to easily incorporate new functionality and new data types. No javascript is used, improving browser compatability. It is now possible to query by CATH code, Pfam ID or accession, UniProt ID or accession, COG identifier and NCBI taxonomy code. These terms are also tagged within the results pages to facilitate querying of results. We have also included a BLAST (16) search facility which will identify the likely family that the query sequence belongs to.

Two main types of data return pages have been developed—the detailed view and the summary view (see below). The detailed views return any CATH, Gene3D protein family, Pfam, GO, KEGG (17), COGs/KOGs, BIND (18) and MINT (19) data associated with the protein or set of proteins in the query. Domain information, and other structural information (low complexity regions, coiled coils and signal peptides) are displayed using the Pfam domain drawing service so as to provide a depiction that is familiar to many. We also expect to provide transmembrane helix predictions using the SPLIT 4.0 (20) software shortly.

The summary view provides a simple aggregate description of the data set. This view includes all GO terms and all distinct domain architectures found for the proteins under investigation.

We have also developed an XML format output so that users can easily download all the data returned in a machine and human readable format. This is to aid both automated queries and obtaining very large datasets without attempting to display them as HTML.

Other notable new features include: an on-the-fly sequence alignment facility [using MUSCLE (21)] to aid users in interpreting and validating structural and functional assignments, mouse-over activated summaries for structural and functional terms and direct links from terms to all the source databases.

Web services (DAS)

In addition to the website we provide comprehensive web-services [including XML-RPC and DAS (22)] for programmatic access to the resource. Web-services are crucial to provide remote users straightforward tools to integrate our resource in their applications. The web-services API is documented at <http://bsmmac1.biochem.ucl.ac.uk:8080/Gene3D/Info/Webservices>.

The Gene3D DAS server offers 2 services provided by ProServer (http://www.sanger.ac.uk/Software/analysis/proserver/). The gene3d_uniprot DAS server (http://128.40.46.20:9000/das/gene3d_uniprot/features?segment=<UniProt Accession>) returns a list of Gene3D features for a query UniProt sequence. For each feature the following information is supplied: the Gene3D ID, the feature source method (CATH, Pfam etc.), the feature start/stop coordinates and a note consisting of the method identifier (Cath ID, Pfam accession etc.). The g3dtribe_uniprot DAS server (http://128.40.46.20:9000/das/g3dtribe_uniprot/features?segment=<UniProtAccession>) returns the Gene3D family id for a query UniProt sequence. This annotation applies to the whole sequence and therefore has a range of 0 to 0 (DAS shorthand for 'the whole sequence'). Information on using the DAS servers can be found at http://cathwww.biochem.ucl.ac.uk:8080/Info/Webservices/.

USING GENE3D

The nature and arrangement of the data in Gene3D allows researchers to easily ask questions about the general rules of protein evolution and functional distribution and also to investigate individual proteins. Below we describe a few simple investigations.

Genome domain content (multi-domainicity)

With the Gene3D structural data it is simple to approximate the proportion of proteins in any genome that have more than one domain.

We initially used the PFscape (14) protocol and ProteinMiner—a locally developed data-mining tool to determine the domain composition for each protein in the complete genome set. The proteins were then split into two sets—those that had at least one known domain ('annotated') and those that didn't ('unannotated'). In the first set, if there was a gap of more than 50 residues then we considered that this indicated the presence of an unidentified domain, allowing us to split the

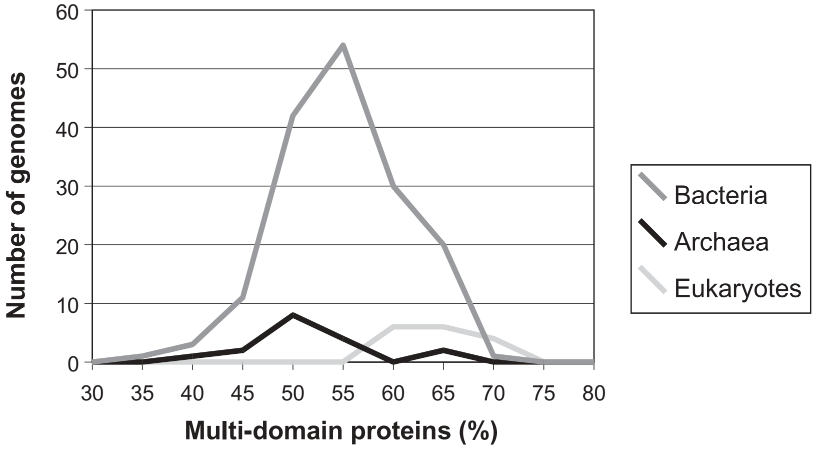

Figure 1. Distribution of multi-domain proteins in 240 genomes. For each genome the approximate percentage of multi-domain proteins was calculated. The likely domain content for those protein which have no known domains was approximated on the basis of length (for details see text). The length threshold was calculated for each genome individually. Of note, the multi-domain percentage for Eukaryotes was within the range displayed by Eubacteria, but the mean for Eukaryotes is substantially higher than for Prokaryotes.

annotated set into single domain proteins and multi-domain proteins.

For each genome we then calculated a length threshold based on the average length of the single domain proteins plus two thirds off the standard deviation. The unnatotated proteins were then divided in single domain and multi-domain proteins. The results of this calculation are shown in Figure 1. The numbers obtained are roughly in concordance with the results obtained by Eckman *et al.* (23) when they used a gap size of 50 residues to be equal to a domain and also within a couple of percent of that manually calculated by S. Teichmann *et al.* (24) for *Mycoplasma genitalium.*

Annotating hypothetical proteins

Gene3D can also be used to effectively predict functions for 'hypothetical proteins'. As an example we took the first three non-viral proteins with the name 'hypothetical protein' returned by UniProt-O43716 (human), Q9SMZ9 (*Arabidopsis thaliana*) and O30176 (*Archaeoglobus fulgidis*). By examining the functional terms and structural predictions associated with these proteins and the protein families they belong to, we were able to assign some annotation to all of them. Use of the sequence alignment tool allowed us to justify this transfer of information. O43716 is a Glu-tRNAGln amidotransferase C subunit (Pfam:PF02686). Q9SMZ9 is possibly involved in mitochondrial distribution and morphology (GO process:0007005). O30176 contains a MazG nucleotide pyrophosphohydrolase domain (Pfam:PF03819).

Identifying structural targets

Another current task for Gene3D is in identifying good targets for the second phase of the NIH-funded protein structure initiative (PSI2). Using our data we are able to identify those domain sequences as determined using the Pfam hits that do not have a close (>30% sequence identity) homologue with a solved structure as determined using the CATH classification.

DISCUSSION AND DEVELOPMENT

Gene3D has been redesigned to provide a feature rich workbench for both the laboratory and the computational biologist. It is possible to investigate individual proteins in detail, with most major sources of functional information presented, and also it is easy to download large datasets for global functional or evolutionary analyses. The new internal infrastructure allows novel data sources to be easily included and so we anticipate significant expansion in the data we present. We also wish to expand the website tools, particularly in regards to genome comparison, to aid researchers in understanding the structural evolution of proteomes.

ACKNOWLEDGEMENTS

Gene3D is a member of the BioSapiens Network of Excellence. We would also like to thank the previous creators of Gene3D—Dan Buchan and Al Grant—for their contributions. Funding to pay the Open Access publication charges for this article was provided by the MRC.

Conflict of interest statement. None declared.

REFERENCES

1. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C., Richter,J. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
2. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D., Akpor,A. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
3. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
4. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

5. Sillitoe,I., Dibley,M., Bray,J., Addou,S. and Orengo,C. (2005) Assessing strategies for improved superfamily recognition. *Protein Sci*., **14**, 1800–1810.
6. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
7. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.
8. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
9. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*., **32**, D142–D144.
10. Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001) 'CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins'. *Nucleic Acids Res*., **29**, 33–36.
11. Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K., Phan,I. *et al*. (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*., **33**, D297–D302.
12. Maibaum,M., Rimon,G., Orengo,C.A., Martin,N.J. and Poulovassilis,A. (2004) BioMap: Gene Family based Integration of Heterogeneous Biological Databases Using AutoMed Metadata. *15th International Workshop on Database and Expert Systems Applications (DEXA 2004)*, Vol. 1, IEEE Computer Society, Los Alamitos, pp. 384–388.
13. 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., Nikolskaya,A.N. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
14. 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.
15. Ranea,J.A., Buchan,D.W., Thornton,J.M. and Orengo,C.A. (2004) Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol*., **336**, 871–887.
16. Altschul,S.F., Madden,T.L., Schaffer,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.
17. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
18. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.
19. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
20. Juretic,D., Zoranic,L. and Zucic,D. (2002) Basic charge clusters and predictions of membrane protein topology. *J. Chem. Inf. Comput. Sci*., **42**, 620–632.
21. Edgar,R. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
22. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, e7.
23. Ekman,D., Bjorklund,A.K., Frey-Skott,J. and Elofsson,A. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol*., **348**, 231–243.
24. Teichmann,S.A., Park,J. and Chothia,C. (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.