W546–W550 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 6 June 2011*
doi:10.1093/nar/gkr438

The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences

Corin Yeats*, Jonathan Lees, Phil Carter, Ian Sillitoe and Christine Orengo

Research Department of Structural and Molecular Biology, Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower St, London. WC1E 6BT, UK

Received February 18, 2011; Revised May 3, 2011; Accepted May 13, 2011

ABSTRACT

The Gene3D structural domain database provides domain annotations for 7 million proteins, based on the manually curated structural domain superfamilies in CATH. These annotations are integrated with functional, genomic and molecular information from external resources, such as GO, EC, UniProt and the NCBI Taxonomy database. We have constructed a set of web services that provide programmatic access to this integrated database, as well as the Gene3D domain recognition tool (Gene3DScan) and protein sequence annotation pipeline for analysing novel protein sequences. Example queries include retrieving all curated GO terms for a domain superfamily or all the multi-domain architectures for the human genome. The services can be accessed using simple HTTP calls and are able to return results in a range of formats for quick downloading and easy parsing, graphical rendering and data storage. Hence, they provide a simple, but flexible means of integrating domain annotations and associated data sets into locally run pipelines and analysis software. The services can be found at http://gene3d.biochem.ucl.ac.uk/WebServices/.

INTRODUCTION

Most proteins consist of one or more independently folding compact globular structures known as protein domains (e.g. see Figure 1). Comparison of domain structures allows the identification of deeper homology relationships than can be detected with sequence information alone, and domain structures can provide a framework for interpreting sequence conservation patterns and the effects of polymorphisms. The CATH-Gene3D resources provide an evolutionary classification of known protein domains for both protein structures [CATH (1)] and protein sequences [Gene3D (2)].

Using a combination of manual curation and automatic boundary recognition algorithms, protein structures sourced from the Protein Databank [wwPDB (3)] are decomposed into their constituent domains and conservatively grouped into homologous superfamilies [for details see (1)]. Non-redundant representatives (at 35% sequence identity) are then selected and used to construct a profile-HMM library. In turn, this is used to search the major sequence databases UniProt (4), RefSeq (5) and Ensembl (6) to identify domain homologues [for details see Gene3DScan below and (2,7)]. These predicted domains are made available through the Gene3D website (http://gene3d.biochem.ucl.ac.uk), along with integrated functional annotations from GO (8), KEGG (9), the Enzyme Commission nomenclature, UniProt and other protein family resources including Superfamily (10) and Pfam (11). However, up until now no programmatic methods for accessing the data or using the domain recognition tools have been available. Here, we present a newly developed set of web services from the Gene3D protein domain discovery and analysis resource.

High-throughput sequencing has placed the computational analysis of sequences under two opposing pressures. First, the high volume of sequence data has allowed more complex and more subtle investigations to be carried out, benefitting from a wider set of tools in the analytical pipelines, such as those developed for metagenomics research. Second, the increasing rate of sequencing, combined with the number of laboratories now handling their own significant-sized sequence databases, means that it is increasingly difficult for centralised resources to provide up-to-date comprehensive annotation sets.

In order to better support the development of external resources that use Gene3D, we have constructed a web service platform that provides access to the annotation software and data used to generate the Gene3D

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

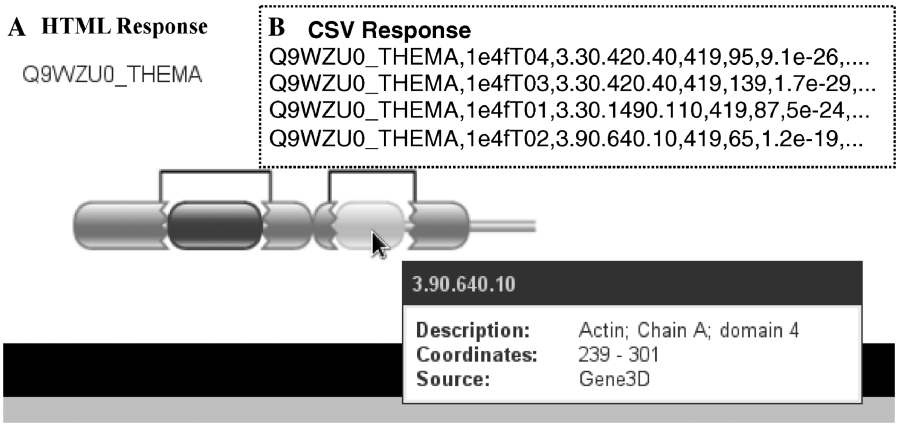

Figure 1. (**A**) Graphical output from the Gene3DScanSvc, generated using the Pfam domain drawing JavaScript library. In this example, the domain architecture of FtsA from Thermatoga maritima has been calculated to consist of four domains, including two discontinuous domains from the same superfamily as indicated by the colours. Discontinuous domains are identified by jagged internal edges and linking black lines. Information about displayed elements is shown in pop-up boxes activated by rolling the mouse over the domain of interest. (**B**) Part of the corresponding CSV format response. *E*-values and boundaries produced by HMMER and DomainFinder are reported.

database. These services have been built largely in accordance with RESTful principles, using simple URIs to identify resources and HTTP commands to specify the action and format. This approach makes the creation of client applications simple since libraries using HTTP are well supported in all major programming languages, including allowing direct browser access and basic *Nix command line tools such as wget or curl. HTTP itself has proven to be completely cross-platform compatible through its success as a fundamental component of the World Wide Web.

Currently the services divide into three sets: (i) Gene3DScan for annotating proteins with accurate structural domain assignments; (ii) sequence annotation services for identifying transmembrane regions, coiled coils and others; and (iii) data access services for extracting pre-calculated multi-domain architectures (MDA), phylogenetic profiles, GO and EC annotations and other information. All tools have a web page front-end providing descriptions, documentation and direct access, though it is not recommended for downloading large data sets (see Table 1 for complete list of services).

Why be RESTful?

At its simplest, REpresentational State Transfer (REST) is a set of principles for designing scalable services that are easy to use and navigate, and are independent of specific technologies. Services consist of three parts—the resources (e.g. a particular slice of data), the actions (e.g. 'GET' or 'POST') and the representations (e.g. machine-readable XML or human-readable HTML).

'Resources' are identified by a Unique Resource Identifier (URI), which is essentially the same as the URL ('web address') of a website. Typically, a call to a base URI returns a list of resources, e.g. http://myservice.org/proteins will return a list of proteins. To retrieve information on a specific protein ('resource') its identifier is added to the URI (e.g. http://myservice.org/proteins/{protein_id}.

'Actions' are issued to resources using HTTP, the standard Internet communication protocol, which has comprehensive built-in functionality for simple request management. The four basic commands of HTTP are GET, POST, PUT and DELETE. The requested resource is 'represented' according to the allowed formats in the HTTP header of the submitted request and a single resource may be represented in multiple ways (e.g. XML, JSON and HTML). Methods for identifying allowed formats are built into HTTP libraries. Error handling is also built-in with numerical codes indicating the issue while detailed error messages can also be embedded. For instance, if the requested format is not available a '406 Not Acceptable' will be returned while the more frequently seen '404 Not Found' is returned when a resource does not exist.

RESTful HTTP services allow for ease of developing computational interfaces while also making it easy to overlay web browser friendly interfaces. This simplifies testing for users who can explore the system without writing a line of code. In a similar manner to SOAP's WSDLs, REST also provides a means for documenting the services in a machine-readable way called WADL. WADL descriptors are provided for all services.

The data stored by Gene3D is particularly amenable to this architecture: by using the basic hierarchical nature of a URI each release can be given its own branch and each point of data fixed at a particular address. For instance, http://gene3d.biochem.ucl.ac.uk/Gene3DDataServices/rest/DomainArchitectures/v9.1.0/sequence-databases/uniprot/Q9XA16 will always refer to the domain description for UniProt protein Q9XA16 from Gene3D release v9.1.0. Also, the format selection greatly improves the usability and scaling of the web services. While XML is very useful as a data storage format, it is also highly verbose. By providing compressible CSV and JSON, responses

Table 1. Complete list of current Gene3D web services, their root URIs and a brief description of the services

Service name	URI	Description
Gene3DScan (http://gene3d.biochem.ucl.ac.uk/Gene3DScanSvc)		
Synchronous	/SuperfamilyScan	Scan FASTA for structural domains (<1000 sequences).
Asynchronous	/async	Scan large FASTA for structural domains (<2.5MB).
Computational Services (http://gene3d.biochem.ucl.ac.uk/Gene3DComputeServices)		
Coiled-coils	/*marcoils*	Simple interface to the Marcoils (15) predictor.
Transmembrane helices	/tmhmm	Simple interface to TMHMM v2 (14).
Disordered regions	/anchor	Simple interface to the IUPred (16) predictor.
MetaMotif	/metamotif	Unified front end to the computational services.
Data services (http://gene3d.biochem.ucl.ac.uk/Gene3DDataServices/rest)		
CATH Superfamily Descriptions	/CathFamilyDescriptions	Get descriptions for CATH superfamily codes.
CATH Superfamily Members	/CathFamilyMembers	Get the UniProt members for a CATH superfamily.
CATH structural domains mapped to UniProt	/CathToUniprotMap	Get the location of CATH (PDB) domains in protein sequences.
CATH-Gene3D phylogentic profiles	/GenomeProfiles	Get the distribution of superfamilies for approximately 2000 genomes.
Detailed domain assignments for complete genomes	/GenomeAssignments	Get detailed domain assignments for complete genomes in Gene3D.
Domain assignments and protein architectures	/DomainArchitectures	Get domain assignments for individual proteins and large-scale collections.
Enzyme Commission Code Assignments	/EnzymeCodes	Get EC codes associated with superfamilies.
Functional residues	/FunctionalResidues	Get functional residues (e.g. active sites) that overlap with domains.
GO functional annotations	/GoFunctions	Get GO function terms associated with superfamilies.
Pfam families with no structural representatives	/PfamNsr	Get the Pfam family annotations that do not overlap with a Gene3D domain.

The services divide into three sets: Gene3DScan, external computational services and data access services. Examples can be found in the on-line documentation for all services, along with a complete list of paths for the data services at http://gene3d.biochem.ucl.ac.uk/Gene3DDataServices/rest/service_paths.html.

download times can be dramatically reduced. This is of particular consideration when downloading complete annotation sets for the large-scale sequence databases. The computational services use very simple inputs and provide a single response if successful, and so can be easily accessed and managed just by using a combination of POST (to submit a sequence) and GET (to retrieve the results), while the data services just use GET. HTML responses are provided for the computational services so as to simplify generating graphical images of the results.

The CATH-Gene3D resources

The primary goal of CATH-Gene3D is to identify domains and catalogue their evolutionary relationships. Gene3D contains two sources of domain annotations for protein sequence. The primary, gold-standard, annotation is derived by directly mapping the manually identified structural domains to protein sequences. A second set of annotations uses an automated process to identify all the sequence homologues of these structural domains within a combination of public sequence databases.

CATH v3.4.0 contains 152 920 structural domains identified from proteins deposited in the wwPDB and homologous domains clustered into 2549 superfamilies. To create the gold-standard set of domain sequences, the assignments are mapped directly to protein sequences using the PDB-to-UniProt mapping from the Structure Integration with Function, Taxonomy and Sequence initiative (SIFTS) provided by UniProt and the PDBe (12) and then resolved into a single multi-domain architecture—multi-domain architecture being the order and position of the domains within the protein. The resolution method takes into account uncertainties in the mapping and overlaps between domains to map 142 774 (93%) of the domains in CATH to 20 673 domains in 13 232 proteins. The large reduction in the number of observed domains is principally due to the high level of redundancy in the PDB. This set provides a useful tool for benchmarking domain prediction methods, including assessing the performance of the Gene3D superfamily models used in the automated process below.

The second set of sequence domain annotations was created by searching the structural domains in CATH against the millions of proteins in public sequence repositories. First, HMMER 3.0 (http://hmmer.janelia.org/software) was used to generate a library of HMM models based on non-redundant (at 35% sequence identity) representative sequences from each CATH superfamily, resulting in a set of 11 330 HMM models. The HMMs were then used to scan a sequence database built from UniProt, RefSeq and Ensembl and the resulting matches resolved into multi-domain architectures using the DomainFinder protocol (7). DomainFinder is a graph-based algorithm developed by Gene3D for finding the best representative set of domain assignments amongst a collection of overlapping and conflicting domain assignments for a query sequence. The method has been shown to outperform simply assigning the closest match to a domain superfamily as regards accuracy of domain boundary assignments and minimising false negatives and positives. In Gene3D v10.0.0, 12.5 million domains

were identified in 7 million of 11.4 million scanned sequences (61%).

The Gene3D database also integrates molecular function annotations from GO, and other enzyme classification resources (e.g. EC); also pathway annotations, family assignments and taxonomic data from multiple resources, including UniProt descriptions, KEGG, InterPro (13) and the NCBI.

Service principles

General principles. The services are divided into two groups—computational tools and data access. The computational tools include the in-house Gene3DScan pipeline and external tools for feature prediction (e.g. coiled-coils from MARCOILS). Sequences are submitted to the computational tools for scanning and results are either returned instantly ('synchronous') or a job ID is returned, which can be used to access results at a later stage ('asynchronous'). All the computational services have a web page that describes them and provides an example POST form that can be used directly.

The data services are simply accessed by specifying the format and URI of the data; for very large data sets, there may be a lag of some minutes while the query is cached for the first time. For details on how to request specific formats and compression see http://gene3d.biochem.ucl.ac.uk/WebServices/formats.html. Example clients in Python and Perl can be found linked from http://gene3d.biochem.ucl.ac.uk/WebServices/. Table 1 shows the complete set of current services along with a brief description.

About the Gene3DScan service. The Gene3DScan service provides web-based access to the Gene3D domain architecture prediction pipeline for small to moderate sized sequence sets (i.e. 1–1000). Under low loads, a single sequence will return within a second, while a larger set may take a few minutes. The service may be accessed synchronously (limited at 100 sequences) or asynchronously with a simple ticket-based interface (limited to a 2.5 MB uploaded FASTA file). For the synchronous service, the results are returned directly in the requested format (see Figure 1 for example responses). For instance, most browsers (notably excluding Google Chrome) will request an HTML page, which will produce a graphical visualization of the domain architecture; on the other hand, *wget* will default to a plain text (CSV) response. The asynchronous service returns a job identifier and the URI for retrieving the results. A request to this URI will either return a message indicating that the job is incomplete, or a list of paths for downloading output files (for details see http://gene3d.biochem.ucl.ac.uk/Gene3DScanSvc).

About the sequence annotation services. The sequence annotation services consist of a set of tools used to search protein sequences for non-structural functional elements, e.g. transmembrane regions. They are not intended to provide a comprehensive interface to the tools, but to provide a simple one-stop-shop providing residue annotations for researchers annotating a small number of sequences and to allow easier comparison to homologues already in Gene3D. Currently, transmembrane [TMHMM (14)], coiled-coils [MARCOIL (15)] and disordered region [Anchor (16)] prediction services are provided, along with a 'meta-service' that runs all three predictors in one request.

About the data services. As well as providing the means to identify functional regions and domains in novel protein sequences, Gene3D also provides a wealth of integrated protein annotations. The domain content of genomes along with associated molecular and enzymatic functions can be used to support a variety of genomic, metagenomic and evolutionary studies; for instance, to examine the functional content of the Last Universal Common Ancestor (LUCA), or to examine the effects of splicing and other sequence polymorphisms (17–19).

There are currently nine data download services: *CATH Superfamily Members*, *CATH Structural Domains Mapped to UniProt*, *Domain Assignments and Protein Architectures*, *Enzyme Commission Code Assignments*, *Functional Residues*, *Detailed Domain Assignments for Complete Genomes*, *CATH-Gene3D Phylogenetic Profiles*, *GO Functional Annotations* and *Pfam Families with No Structural Representatives*. Most of these services provide simple links between superfamily identifiers and representatives or function descriptions. For instance, the *GO Functional Annotations* service allows users to retrieve superfamilies with the same function, or retrieve all functions for a superfamily that are supported by a Traceable Author Statement (TAS) (e.g. http://gene3d.biochem.ucl.ac.uk/Gene3DDataServices/rest/GoFunctions/v9.1.0/superfamilies/1.10.10.10/TAS will return all TAS-supported GO terms for superfamily 1.10.10.10). The *CATH-Gene3D Phylogenetic Profiles* service provides two types of phylogenetic profile: (i) the distribution of species within a superfamily and (ii) the distribution of domain superfamilies for a specific genome. Larger background sets are also provided, such as the distribution of superfamilies found in bacteria or generally across UniProt. *Pfam Families With No Structural Representatives* provides the set of Pfam-A matches [pre-calculated by SIMAP (20)] that do not overlap a Gene3D domain assignment. *CATH Structural Domains Mapped to UniProt* allows access to the high-quality curated set of CATH domain assignments as described in 'The CATH-Gene3D Resources' (above) and can be queried via domain identifier or UniProt accession.

Constructing the data service URIs. The data services are designed to be both comprehensive and simple to use with data extractable as individual annotations (e.g. a protein's domain content) or as large-scale sets (e.g. all Pfam domains that do not overlap with a Gene3D domain). For all services, the root of the service returns the available versions—e.g. http://gene3d.../rest/GoFunctions returns a list of Gene3D releases. Appending the version code then returns the data types that can be queried—e.g.../GoFunctions/v9.1.0 returns 'superfamilies', 'go-terms', 'sequence-md5s'. Appending 'superfamilies' will return the list of superfamilies for v9.1.0. And so

on, until at the leaf nodes, a detailed description is returned of the selected entity. If a link returns a '404 Not Found' error it means that there was not a corresponding bit of information. For instance, http://gene3d.biochem.ucl.ac.uk/Gene3DDataServices/rest/GoFunctions/v9.1.0/superfamilies/2.10.10.10/TAS will return a '404' since there are no GO functions with TAS for superfamily 2.10.10.10.

DISCUSSION

We have produced a suite of web services to support researchers in a range of tasks at the crossroads of structure, function and evolution. This includes incorporation of structure-based homology assignments into annotation pipelines, and mining of functional associations to improve function prediction tools. The design of the services has focussed on simplicity and modularity. This provides a stable and reliable interface that can easily be extended to incorporate new tools and data sets, as well as allowing for expansion of back-end resources to support increasing user demand.

ACKNOWLEDGEMENTS

We would like to thank Pfam for providing their domain drawing JavaScript library to us.

FUNDING

The European Commission's ENFIN (grant number LSHG-CT-2005-518254 to J.L.); IMPACT Networks of Excellence funded under the framework programme 7 (to C.Y.); EMBRACE (to I.S.); Wellcome Trust (081989/Z/07/Z to I.S.); National Institutes of Health CSGID (HHSN272200700058C to P.C.). Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
2. Lees,J., Yeats,C., Redfern,O., Clegg,A. and Orengo,C.A. (2010) Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.*, **38**, D296–D300.
3. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropolous,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
4. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
5. 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.
6. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
7. Yeats,C., Redfern,O.C. and Orengo,C.A. (2010) A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics*, **26**, 745–751.
8. Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
9. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
10. Wilson,D., Pethica,R., Zhou,Y., Talbot,C., Vogel,C., Madera,M., Chothia,C. and Gough,J. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.
11. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
12. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
13. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
14. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
15. Delorenzi,M. and Speed,T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**, 617–625.
16. Mészáros,B., Simon,I. and Dosztányi,Z. (2009) Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Comput. Biol.*, **5**, e1000376.
17. Ranea,J.A., Sillero,A., Thornton,J.M. and Orengo,C.A. (2006) Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.*, **63**, 513–525.
18. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.I., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
19. Izarzugaza,J.M., Baresic,A., McMillan,L.E., Yeats,C., Clegg,A.B., Orengo,C.A., Martin,A.C. and Valencia,A. (2009) An integrated approach to the interpretation of single amino acid polymorphisms within the framework of CATH and Gene3D. *BMC Bioinformatics*, **10**, S5.
20. Rattei,T., Tischler,P., Arnold,R., Hamberger,F., Krebs,J., Krumsiek,J., Wachinger,B., Stümpflen,V. and Mewes,W. (2008) SIMAP–structuring the network of protein similarities. *Nucleic Acids Res.*, **36**, D289–D292.