Extending CATH: increasing coverage of the protein structure universe and linking structure with function

Alison L. Cuff1*, Ian Sillitoe1, Tony Lewis1, Andrew B. Clegg1, Robert Rentzsch1, Nicholas Furnham2, Marialuisa Pellegrini-Calace3, David Jones4, Janet Thornton2 and Christine A. Orengo1

1Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower Street, London, WC1E 6BT, 2European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, 3Department of General and Environmental Physiology, University of Bari, Bari, Italy and 4Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK

Received September 15, 2010; Accepted October 6, 2010

ABSTRACT

CATH version 3.3 (class, architecture, topology, homology) contains 128 688 domains, 2386 homologous superfamilies and 1233 fold groups, and reflects a major focus on classifying structural genomics (SG) structures and transmembrane proteins, both of which are likely to add structural novelty to the database and therefore increase the coverage of protein fold space within CATH. For CATH version 3.4 we have significantly improved the presentation of sequence information and associated functional information for CATH superfamilies. The CATH superfamily pages now reflect both the functional and structural diversity within the superfamily and include structural alignments of close and distant relatives within the superfamily, annotated with functional information and details of conserved residues. A significantly more efficient search function for CATH has been established by implementing the search server Solr (http://lucene.apache.org/solr/). The CATH v3.4 webpages have been built using the Catalyst web framework.

DESCRIPTION OF CATH HIERARCHY AND CURRENT POPULATION OF DATABASE

CATH (class, architecture, topology, homology) is a hierarchical protein domain classification system (1), where class reflects the amino acid composition, architecture the general shape of the protein domain and topology the way in which the protein folds into this architecture.

Homology captures evolutionary relationships between protein domains. Protein structures are taken from the Protein Data Bank (PDB) and decomposed into chains which, in turn, are split into domains. Domains are classified into homologous superfamilies using a combination of in-house algorithms which exploit structure, sequence and functional information. Fold groups and remote homologues (<80% sequence identity) are validated by manual curation. The class and architecture of the protein are manually specified (1).

The latest version of CATH (CATH v3.3) has expanded by 123 new folds, 199 new superfamilies and 14 473 new domains over the previous release. Table 1 shows the current population of different levels in the CATH hierarchy. CATH v3.4 has 22 988 more domains than CATH v3.3.

Figure 1(a) is of a ‘CATHerine wheel’ plot showing the population of non-homologous structures, i.e., the structures representing each homologous superfamily, within the different hierarchical layers in CATH v3.3. Figure 1(b) shows the increase in the number of superfamilies between CATH v3.2 and CATH v3.3. Folds with the greatest increase in superfamily numbers include the α-β plaits, four α-helix bundles and SH3 type β-barrels.

In version 3.3, 36.2% of the new domains classified into CATH superfamilies fall within the top 10 most highly populated folds which currently account for 35.7% of all non-homologous domain structures in CATH.

INCREASING COVERAGE OF PROTEIN FOLD SPACE

The curation of both CATH v3.3 and CATH v3.4 has largely focused on classifying structures solved by

*To whom correspondence should be addressed. Tel: +44 20 7679 3890; Fax: +44 20 7679 7193; Email: cuff@biochem.ucl.ac.uk; alisoncuff@yahoo.co.uk
structural genomics (SG) initiatives. One of the principal aims of the SG initiatives is to discover all the folds that exist in the protein structure universe (3). Previous analyses by our group have shown that a large proportion of structural superfamilies in nature are likely to be already represented in CATH, i.e., CATH superfamilies already account for a large proportion of domain sequences (up to 80%) in completely sequenced genomes (4). Indeed, there has been a gradual decrease in the number of new folds identified over the last decade (5). Currently, <2% of structures solved by traditional structural biology represent novel fold groups (5,6). By contrast, various studies (6–8) have shown that a higher proportion of any novel folds are represented by SG structures. Using a normalized root mean square deviation (RMSD) of 5Å to determine structural novelty, a recent study has shown that 28% of SG domains have novel structures compared with only 3% of non-SG domains (6).

In CATH v3.4, 1633 new SG structures have been classified, resulting in 99 new superfamilies and 39 new fold groups.

A significant proportion of domain sequences in completely sequenced genomes, currently unrepresented in CATH, are predicted to be transmembrane proteins (9). Structural classification of membrane proteins is more difficult than for soluble proteins due to the limited number of structural arrangements and their tendency to be structurally similar regardless of evolutionary history or function (9). Transmembrane proteins are also difficult to determine using experimental methods and some SG centres are specifically focusing on these types of proteins as targets (10). Many are α-helical proteins, comprising a single transmembrane helix, a helix hairpin or a 4 α-helix transmembrane bundle (9).

CATH v3.4 includes 2274 new transmembrane proteins, accounting for 71 new superfamilies and 22 new fold groups. Most of the newly classified superfamilies (62%) are α-helical in nature, with some 24% being single transmembrane helix superfamilies (see Table 2). A list of membrane-associated CATH superfamilies, with links to their individual superfamily pages, is now available to

| Table 2. Table showing the number of new superfamilies created from transmembrane proteins that have been classified for CATH v3.4 |
| Architecture | Number of new superfamilies |
|---------------|-----------------------------|
| Single transmembrane helix | 17 |
| Helix hairpin | 8 |
| α-Orthogonal bundle | 9 |
| α-Up-down bundles | 10 |
| β-Barrel | 8 |
| α/β-Roll | 2 |
| Two-layer β-sandwich | 3 |
| Two-layer α/β-sandwich | 9 |
| Few secondary structures | 3 |
| β-ribbon | 1 |
| Single sheet | 1 |
view at http://www.cathdb.info/sfam/membrane/ and will be added as an option on the CATH portal for CATH version 3.4.

REDESIGNING THE WEBSITE

Historically, CATH has provided information on protein structures only. Information on CATH superfamily sequence relatives is currently obtainable from CATH’s ‘sister’ site, Gene3D (11). Multi-domain architectures (MDA) and taxonomic distribution for CATH superfamilies are also provided through Gene3D as are a number of protein functional annotations. These include protein–protein interaction (PPI) data (12), GO functional assignments (13), KEGG pathways (14) and FunCAT functional descriptions (15).

Current work on the CATH website includes the development of a single web-based portal through which users can access the data provided by both CATH and Gene3D. All the usability of the original site is being maintained, including the CATHEDRAL (16) and SSAP (17,18) web servers for structural comparison. Users are able to browse through the hierarchy in the same manner as previous incarnations of the website.

The CATH superfamily pages, however, have been completely redesigned in order to provide the functional information previously only available though Gene3D and structural diversity known to exist within some superfamilies (see Figure 2). Beta pages for CATH version 3.4 for the HUP superfamily (cath code 3.40.50.720) are available for viewing (http://beta.cathdb.info/cathnode/3.40.50.720). Previous research carried out by our group has shown that the 100 most structurally diverse superfamilies in CATH are also the most highly populated, accounting for around 40% of the domain sequences in the genomes (see Figure 3) (19). Integrating sequence data more seamlessly with the structural data allows us to identify the most structurally and functionally diverse superfamilies and the most highly populated (see Figure 4).
The largest and smallest domains within any particular superfamily are now displayed to give a snapshot of the structural variation across the superfamily (see Figure 3). A more thorough understanding of structural diversity across the superfamily can be obtained by viewing plots of structural similarity scores between pairs of relatives (see Figure 4).

Within each superfamily, we also provide information on structurally coherent groups of relatives. Structurally similar groups are identified by comparing domain structures using our in-house structure comparison algorithm [SSAP (17,18)] and using multi-linkage clustering to generate groups of ‘close’ structural clusters [superposing with normalized RMSD (20) <5Å] and clusters of structurally more distant relatives (superposing with RMSD <9Å).

In our previous analyses, a superfamily with five or more close SSGs was considered to be structurally diverse (19). By including predicted CATH domains in our CATH resource, we can see from Figure 2 that there is a correlation between structural diversity, measured by the number of close structural clusters and the sequence diversity, measured by the number of sequence clusters (domains clustered at 30% sequence identity).

As regards functional annotations, users can explore the degree of functional diversity across the superfamily by examining the range of annotations provided for all the predicted CATH sequence relatives (integrated from Gene3D) by the Enzyme Classification (21), UniProt (22), FunCAT (15), KEGG (14) and GO resources (13). Over-represented keywords are also extracted from domain and protein annotations using the Solr search engine (see section below) to give an indication of the overall functionality and functional diversity of the superfamily being viewed.

CATH v3.4 provides multiple structural alignments [displayed using the Jalview applet (23)] for both close and distant structural clusters showing conserved residues [as calculated by scorecons (24)] and functional residue data downloaded from WSas (catalytic residues and ligand binding residues) (25). Multiple sequence alignments will also be provided for a recently established functional family subclassification within the superfamily.
GeMMA clusters (26). Individual three-dimensional (3D) domain structures from these alignments can be selected for display using the Jmol applet (27), complete with annotated functional and conserved residues (see Figure 5). This resource will be expanded in the future to include other functional data, for example, relating to protein interactions and also mutation data from OMIM.

PHYLOGENETIC DISPLAY OF FUNCTIONAL CHARACTERISTICS IN CATH SUPERFAMILIES

In addition to providing multiple structural alignments, the structural clusters, representing close and distant structural relatives, respectively, are used to generate structure guided multiple sequence alignments of the sequence domains from Gene3D associated with the cluster. The alignments are presented on the CATH superfamily pages and are utilized by a new resource (FunTree) being developed in collaboration with the Thornton Group at the European Bioinformatics Institute. The expanded structure-based sequence alignments are used to generate a phylogenetic tree of the relatives. For superfamilies that contain known enzymes, functional data are displayed, including assigned enzyme commission (E.C.) number.

In addition, comparative analysis of the enzyme’s reactions, using comparisons between bond order changes and substrate substructure similarity (28), is carried out. The results of the reaction and small-molecule analysis are presented in conjunction with the sequence comparison analysis. Furthermore, the multi-domain architecture information from Gene3D is taken into consideration. This allows the analysis of the evolution of enzyme function within a CATH superfamily.

MOVING CATH TOWARDS A MODERN WEB FRAMEWORK

Since 2006, the content for the CATH website has been generated through a series of standalone CGI scripts written in the Perl programming language. Although
IMPROVEMENTS TO THE SEARCH FUNCTION IN CATH

In order to improve CATH's searchability, we have built a dedicated search engine which indexes structures and CATH classifications according to the keywords and entity IDs associated with them and the full text of their descriptions and annotations. This uses Solr (31), a search server based on the popular Lucene toolkit, enabling complex queries across various fields which support Boolean operators, phrase searches, wildcards and many other advanced search features. For each CATH release, related data from CATH, Gene3D, PDB, UniProt and other external sources are aggregated and flattened into a single 'document' per CATH entity, which is added to a Solr index. This enables even highly complicated queries to be answered very quickly.

The Solr index can also be queried for the most significant terms associated with a given entity. We use this feature to annotate search results with lists of the most representative keywords for each entity. Solr is entirely web-services based. Queries are answered via a RESTful interface allowing data to be returned in a variety of formats including XML, JSON and CSV. It drives the search functionality on the CATH website, and we plan to make it publically available for external users to query programmatically.

SUMMARY

In summary, CATH has expanded over the last 2 years to include 365 new superfamilies (176 new fold groups), 29% of which came from the SG initiatives (30% fold groups) and 28% (22%) of which were membrane families (folds). We have extended the functional information available for each CATH superfamily by integrating domain sequence relatives from Gene3D and displaying their functional annotations from various public resources (e.g. GO, EC, Kegg and FunCat). We now provide more detailed information on structural and functional diversity across each superfamily and multiple structure alignments for clusters of close and distant structural relatives. The FunTree display presents a phylogenetic perspective of enzyme superfamilies derived from a multiple sequence alignment and annotated by functional characteristics such as EC number and reaction mechanisms. Finally, access to the data in CATH has been made easier by building the webpages within the Catalyst MVC framework and the search facilities have been significantly improved by exploiting Solr and Lucene.

FUNDING

A. Cuff, M. Pellegrini-Calace, N. Furnham (BBSRC); A. Clegg (EMBRACE); T. Lewis (IMPACT, E.U); R. Rentzsch (ENFIN, E.U); I. Sillitoe (Wellcome Trust). Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH – a hierarchic classification of protein domain structures. Structure, 5, 1093–1109.
2. 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.
3. Grabowski, M., Joachimiak, A., Otwinowski, Z. and Wadew, K. (2007) Structural genomics: keeping up with expanding knowledge of the protein universe. Curr. Opin. Struct. Biol., 17, 347–353.
4. Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X. and Orengo, C. (2008) Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res., 36, D414–D418.
5. Greene, L.H., Lewis, T.E., Addon, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A. 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.

6. Dessailly, B.H., Nair, R., Jaroszewski, L., Fajardo, J.E., Kouranov, A., Lee, D., Fiser, A., Godzik, A., Rost, B. and Oreno, C. (2009) PSI-2: structural genomics to cover protein domain family space. *Structure*, 17, 869–881.

7. Todd, A.E., Marsden, R.L., Thornton, J.M. and Oreno, C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.*, 348, 1235–1260.

8. Chandonia, J.M. and Brenner, S.E. (2000) The impact of structural genomics: expectations and outcomes. *Science*, 311, 347–351.

9. Neumann, S., Fuchs, A., Mulkidjanian, A. and Frishman, D. (2010) Current status of membrane protein structure classification. *Proteins*, 78, 1760–1773.

10. Hendrickson, W.A. (2007) Impact of structures from the protein structure initiative. *Structure*, 15, 1528–1529.

11. Buchan, D.W., Shepherd, A.J., Lee, D., Pearl, F.M., Rison, S.C., Thornton, J.M. and Oreno, C.A. (2002) Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. * Genome Res.*, 12, 503–514.

12. Kerrien, S., Buchan, D., Jaroszewski, L., Fajardo, J.E., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. et al. (2004) The FunCat, a functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res.*, 32, 5539–5545.

13. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009) Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. *Bioinformatics*, 25, 1189–1191.

14. Talavera, D., Laskowski, R.A. and Thornton, J.M. (2009) WSas: a web service for the annotation of functional residues through structural homologues. *Bioinformatics*, 25, 1192–1194.

15. Lee, D.A., Rentzsch, R. and Orengo, C. (2009) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res.*, 37, 720–737.

16. Talalay, P. and Hofmann, K. (1999) Conserved domains: a search tool for finding, analyzing and classifying all the domains of a protein database. *Nucleic Acids Res.*, 27, 28–30.

17. Taylor, W.R. and Orengo, C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, 208, 1–22.

18. Orengo, C.A. and Taylor, W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, 266, 617–635.

19. Cuff, A.L., Redfern, O.C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T. and Todd, A. et al. (2009) The CATH hierarchy revisited—structural divergence in domain superfamilies and the continuity of fold space. *Structure*, 17, 1051–1062.

20. Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, 28, 304–305.

21. Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, 32, 922–923.