D310–D314 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 7 November 2008*
doi:10.1093/nar/gkn877

The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies

Alison L. Cuff[1,*], Ian Sillitoe[1], Tony Lewis[1], Oliver C. Redfern[1], Richard Garratt[2], Janet Thornton[3] and Christine A. Orengo[1]

[1]Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK, [2]Institute de Fisica de Sao Carlos, Universidade de Sao Paulo, Caixa Postal 369, Sao, Carlos, SP, Brazil and [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received September 15, 2008; Revised October 16, 2008; Accepted October 17, 2008

ABSTRACT

The latest version of CATH (class, architecture, topology, homology) (version 3.2), released in July 2008 (http://www.cathdb.info), contains 114 215 domains, 2178 Homologous superfamilies and 1110 fold groups. We have assigned 20 330 new domains, 87 new homologous superfamilies and 26 new folds since CATH release version 3.1. A total of 28 064 new domains have been assigned since our NAR 2007 database publication (CATH version 3.0). The CATH website has been completely redesigned and includes more comprehensive documentation. We have revisited the CATH architecture level as part of the development of a 'Protein Chart' and present information on the population of each architecture. The CATHEDRAL structure comparison algorithm has been improved and used to characterize structural diversity in CATH superfamilies and structural overlaps between superfamilies. Although the majority of superfamilies in CATH are not structurally diverse and do not overlap significantly with other superfamilies, ~4% of superfamilies are very diverse and these are the superfamilies that are most highly populated in both the PDB and in the genomes. Information on the degree of structural diversity in each superfamily and structural overlaps between superfamilies can now be downloaded from the CATH website.

CURRENT POPULATION OF THE CATH HIERARCHY

CATH (class, architecture, topology, homology) is a hierarchical protein domain classification (1) where domains are classified manually by curators, guided by prediction algorithms (such as structure comparison). Each protein structure is decomposed into one or more chains which in turn are split into one or more domains before being classified into homologous superfamilies according to both structure and function. At the Class, or C-level, the domains are classified simply on the basis of their secondary structure content [whether they are mostly α-helical (Class 1) or β-sheet (Class 2), contain a significant percentage of both secondary structure elements (Class 3) or contain very little secondary structure (Class 4)]. The domains within each class are then sorted according to their architecture—that is similarities in the arrangements of secondary structures in 3D space. Each architecture (A-level) is further broken down into one or more topology, or fold, groups (T-level), where the connectivity between these secondary structures are taken into account. The domains are then classified into their respective homologous superfamilies (H-level) according to similarities in sequence, structure and/or function. Clustering performed at the H-level (>35% sequence identity and above) then produces one or more sequence families for each of the homologous superfamilies (S-level). Table 1 below shows the current population of different levels in the CATH hierarchy.

A PERIODIC TABLE OF CATH ARCHITECTURES

A visual snapshot of the domain architectures in the CATH database is now captured in a new 'Protein Chart' (2). This chart, inspired by Taylor's 'Periodic Table' of protein structures devised in 2002 (3), shows fold representatives of all the most regular domain architectures currently classified in the CATH database. It is organized so that the smallest representative for any given architecture is at the top of the chart and the largest at the bottom, giving a guide to the variation in size and

*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

structure that can occur. Functional information and population statistics for each architecture are provided in a table accessible from the CATH web site (http://www.cathdb.info/download#version_v3.1).

Using the chart, we have identified nine new architectures classified since CATH architectures were first presented in 1997 (1) (Figure 1). These new architectures are not highly populated accounting for only ~4% of predicted CATH domain sequences in the genomes.

In the mainly-α class, the α-solenoid architecture (1.40) contains only one superfamily. Domains provide an α-helical scaffold for a central hydrophobic cavity, which contains light harvesting molecules (4). The αα-barrel (1.50) contains 2 α-helical layers, with long loops that create a tunnel. They are typically glycosyl hydrolases (5). The α-horseshoe (1.25) is a super helical structure made up of a number of 3 α-helical orthogonal bundle repeats.

Table 1. Release statistics for CATH version 3.2

Class	Architecture	Topology	Homologous superfamily	S35 family
1	5	310	682	2078
2	20	196	438	2062
3	14	512	956	4558
4	1	92	102	173
Total	40	1110	2178	8871

In the mainly-β class, we identify a new β-propellor—the 5-bladed propeller (2.115). A new sandwich architecture, the 3-layer βββ-sandwich is made up of three anti-parallel β-sheets layered into three adjacent stacks with an immunoglobulin-like sub-domain. Most are rieske iron–sulphur proteins (7)

Four new architectures are classified in the α-β class. Super-rolls are made from twisted anti-parallel β-strands capped by 2 α-helices. All classified domains bind to and neutralize lipopolysaccharides in the outer-membrane of gram-negative bacteria (8). The 3-layer (βαβ) sandwich architecture contains 10 domains in three different folds. The most highly populated fold is largely comprised of bacterial heat shock proteins. The αβ-prism is made up from a repeating folding unit composed of two parallel α-helices and a β-sheet. Domains with this fold are commonly found in 5-enolpyruvylshikimate-3-phosphate synthase and UDP-*N*-acetylglucosamine enolpyruvyl transferase (9). 5-Stranded αβ-propellers are composed of ββαβ repeats arranged in a circular fashion surrounding a channel in the centre of the structure (10).

INCREASING THE PROPORTION OF NOVEL STRUCTURES CLASSIFIED IN CATH

Recent analyses of CATH domain annotations in Gene3D (11) showed that between 80–90% of domain sequences in completely sequenced genomes can be assigned to a structural family in CATH. This suggests that the CATH

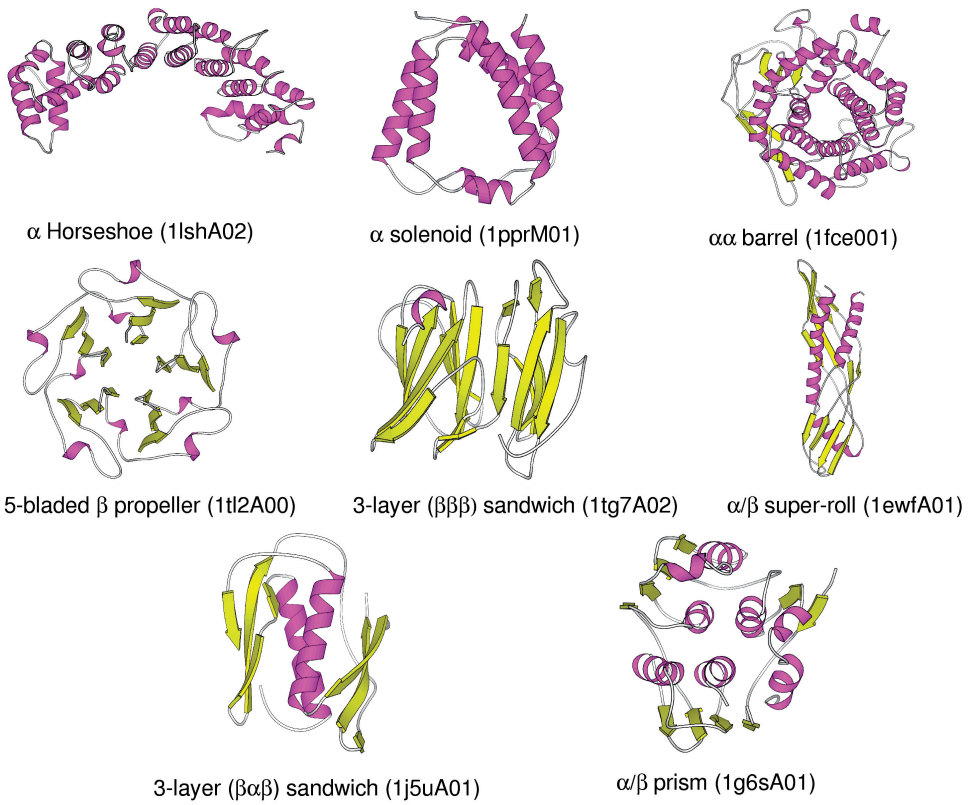

Figure 1. Some of the architectures new to CATH since 1997.

database now provides a reasonably comprehensive structural view of the protein universe. Those protein families that have yet to be represented structurally are likely to be transmembrane or disordered proteins.

The fact that most major folds are represented in CATH is reflected in the continual decrease in the proportion of non-redundant structures found to adopt novel folds. Table 2 gives the number of new folds identified over the last 10 years and the percentage of non-redundant structures deposited adopting a novel fold.

IMPROVEMENTS TO THE STRUCTURAL COMPARISON METHOD CATHEDRAL

We have further improved our domain boundary prediction and fold assignment algorithm, CATHEDRAL (12), which is used to guide curators in the manual classification process. Whole PDB structures are scanned against a library of representatives from the CATH database to recognize constituent domains. CATHEDRAL initially performs rapid secondary structure comparison against the library to identify putative fold matches, which are then more accurately aligned at the residue level using dynamic programming. A support vector machine (SVM) is used to combine different measures of structural similarity and rank hits to the query structure. All domains predicted to be genuine hits by the SVM are assigned in an iterative fashion to identify constituent folds and domain boundaries from the residue-based structural alignments. Hits are allowed to overlap by up to 30 residues and conflicts are resolved by a new algorithm that moves along the overlapping region and assigns each residue to the closest domain.

STRUCTURAL DIVERSITY AND THE VALIDITY OF THE CATH HIERARCHY

There has been much debate on the existence of a protein fold continuum and the validity of a hierarchical protein classification system (13–16). Greene *et al.* (17) previously explored the concept of 'lateral links' across the CATH hierarchy as a way of capturing structural relationships between superfamilies. More recent in-house analyses have shown that, within some of the most highly populated superfamilies, significant structural changes have occurred. Typically, the domains within a given superfamily possess a 'common structural core' comprising 40–50% of the residues in the structure, but there can be considerable structural embellishments to this core and some domains can be up to three times larger than the typical representative of the family (18). In some cases, the embellishments are so considerable that the domain in question can be considered to exhibit a different fold to the other domains in the family.

Due to the improvements made to CATHEDRAL (12), we have been able to perform a database-wide analysis of the similarities between all protein structures in the CATH database. This has been used to examine the extent to which superfamilies diverge structurally and determine which superfamilies overlap with one another. Domains in each superfamily were first assigned to 'structurally similar groups' (SSGs), whereby a domain is assigned to a particular SSG if they exhibit significant structural similarity with other domains in that group (Cuff,A.L. *et al.*, submitted for publication). That is, if they share a normalized RMSD (SiMAX) structure comparison score of <5 Å (Cuff,A.L. *et al.*, submitted for publication). Superfamilies with five or more SSGs were deemed to be structurally diverse.

The majority of homologous superfamilies (~96%) in the database are structurally conserved and structurally coherent, that is, they contain less than five SSGs and do not overlap with any other superfamily. However, the ~4% of CATH superfamilies that do show considerable structural diversity, are those which are the most highly populated in CATH, accounting for 40% of domain sequences in the genomes (Figure 2) (Cuff,A.L. *et al.*, submitted for publication).

If we consider the different SSGs to represent distinct 'folds' within these superfamilies, then instead of the 1110 'fold groups' (defined by the Topology level in CATH

Table 2. Numbers of structures classified in CATH and the proportion of novel folds per year

Year of PDB release	Number PDB structures classified in CATH	Number novel folds	Novel folds (%)
1997	1584	92	5.81
1998	1876	87	4.64
1999	2226	104	4.67
2000	2549	90	3.53
2001	2766	91	3.29
2002	2821	73	2.59
2003	3668	34	0.93
2004	3711	61	1.64
2005	3198	6	0.19
2006	3163	18	0.57
2007	2802	11	0.39

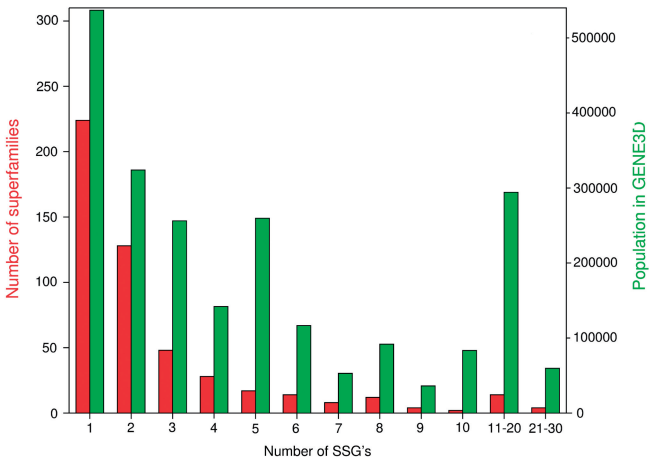

Figure 2. Relationship between the degree of structural diversity (measured by the number of SSGs) and population of the superfamilies in the genomes (number of sequences).

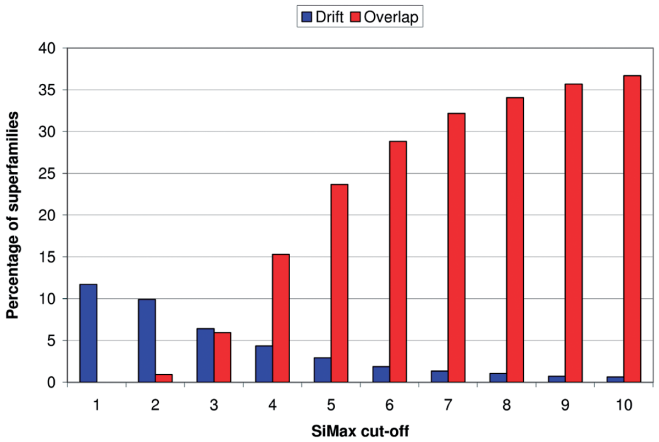

Figure 3. Plot showing the percentage of superfamilies that overlap (red) and show structural diversity, or drift (>5 SSG's) (blue) for different SiMAX cuff-offs.

version 3.2) there would be 3118 'fold groups' and some superfamilies would have multiple 'folds'. However, although examples of dramatic fold changes are known (19), they are rare and the majority of gross structural changes that occur within a superfamily result from extensive structural embellishments to the common core rather than a dramatic change within the core. Therefore, the CATH hierarchical classification is not challenged if we consider a more appropriate definition of the T-level or topology level in CATH to be a grouping of structures sharing a common fold in the core of the domain. A file containing the number of SSGs contained in each superfamily in CATH can be downloaded from (http://www.cathdb.info/download#version_v3.1)

We also investigated whether structures in different superfamilies were structurally similar (i.e. SiMAX <5Å). We observed relatively little overlap between different superfamilies and fold groups for a SiMAX threshold of <5Å. As the threshold is increased, however, more overlaps do occur between some architectures, such as the α up-down bundle, α-orthogonal bundle, β-sandwiches and αβ-sandwiches (Figure 3). This is largely due to the presence of small common super-secondary motifs, such as the α-hairpin, β-hairpin and αβ-motif. Superfamilies that exhibit no structural overlaps at all tend to have very distinctive folds, such as the β-trefoil fold, with unusual motifs or unusual combinations of common motifs.

A structural overlap matrix of SiMAX scores created via the all-against-all CATHEDRAL analysis is downloadable from the CATH website (see http://www.cathdb.info/download#version_v3.1) so that users can perform their own analyses on a CATH-based protein structure universe.

REDESIGNED WEBSITE

The CATH database can be accessed at http://www.cathdb.info. The web interface has been completely redesigned since version 3.1. Documentation, such as an FAQ, tutorials, a glossary, downloadable data files and staff webpages have also been created and are being maintained through a open source wiki software package. This will be frequently updated.

SUMMARY

In the light of our analyses on structural diversity in CATH, it is clear that the T-level provides a clustering of domain structures having similar folds in their domain cores. For each superfamily, information on the variety of different decorations to this common structural core is provided as distinct SSGs within the superfamily. Multiple structural alignments will shortly be provided for each SSG in order to highlight common secondary structures in the domain core and embellishments to this core.

FUNDING

Funding for open access charge: BBSRC.

Conflict of interest statement. None declared.

REFERENCES

1. Orengo,C.A., Michie,A.D., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH: A Hierarchic Classification of Protein Domain Structures. *Structure*, **5**, 1093–1108.
2. Garratt,R.C. and Orengo,C.A. (2007) *The Protein Chart*. Wiley-VCH, Germany. ISBN: 978-3-527-31963-3.
3. Taylor,W.R. (2002) A 'periodic table' for protein structures. *Nature*, **416**, 657–660.
4. Hofmann,E., Wrench,P.M., Sharples,F.P., Hiller,R.G., Welte,W. and Diederichs,K. (1996) . Structural basis of light harvesting by carotenoids: peridinin-chlorophyll-protein from Amphidinium carterae. *Science*, **272**, 1788–1791.
5. Rojas,A.L., Nagem,R.A., Neustroev,K.N., Arand,M., Adamska,M., Eneyskaya,E.V., Kulminskaya,A.A., Garratt,R.C., Golubev,A.M. and Polikarpov,I. (2004) Crystal structures of beta-galactosidase from Penicillium sp. and its complex with galactose. *J. Mol. Biol.*, **343**, 1281–1292.
6. Fülöp,V. and Jones,D.T. (1999) Beta propellers: structural rigidity and functional diversity. *Curr. Opin. Struct. Biol.*, **9**, 715–721.
7. Kolling,D.J., Brunzelle,J.S., Lhee,S., Crofts,A.R. and Nair,S.K. (2007) Atomic resolution structures of rieske iron-sulfur protein: role of hydrogen bonds in tuning the redox potential of iron-sulfur clusters. *Structure*, **15**, 29–38.
8. Beamer,L.J., Carroll,S.F. and Eisenberg,D. (1997) Crystal structure of human BPI and two bound phospholipids at 2.4 angstrom resolution. *Science*, **276**, 1861–1864.
9. Palm,G.J., Billy,E., Filipowicz,W. and Wlodawer,A. (2000) Crystal structure of RNA 3′-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology. *Structure*, **8**, 13–23.
10. Humm,A., Fritsche,E., Steinbacher,S. and Huber,R. (1997) Crystal structure and machanism of human L-arginine: glycine amidinotransferase: a mitochrondrial enzyme involved in creatine biosynthesis. *EMBO J.*, **16**, 3373–3385.
11. 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.
12. Redfern,O.C., Harrison,A., Dallman,T., Pearl,F.M. and Orengo,C.A. (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.*, **3**, e232.
13. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
14. Krishna,S.S. and Grishin,N.V. (2005) Structural drift: a possible path to protein fold change. *Bioinformatics*, **21**, 1308–1310.

15. Kolodny,R., Petrey,D. and Honig,B. (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr. Opin. Struct. Biol.*, **16**, 393–398.
16. Sippl,M.J., Suhrer,S.J., Gruber,M. and Wiederstein,M. (2008) A discrete view on fold space. *Bioinformatics*, **24**, 870–871.
17. Greene,L.H., Lewis,T.E., Addou,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.
18. Reeves,G., Dallman,T., Redfern,O., Akpor,A. and Orengo,C.A. (2006) Structural diversity of domain superfamilies in the CATH database. *J. Mol. Biol.*, **360**, 725–741.
19. Davidson,A.R. (2008) A folding space odyssey. *PNAS*, **105**, 2759–2760.