Nucleic Acids Research, 2005, Vol. 33, Database issue ***D247–D251***
doi:10.1093/nar/gki024

The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis

Frances Pearl, Annabel Todd, Ian Sillitoe, Mark Dibley, Oliver Redfern, Tony Lewis, Christopher Bennett, Russell Marsden, Alistair Grant, David Lee*, Adrian Akpor, Michael Maibaum, Andrew Harrison, Timothy Dallman, Gabrielle Reeves, Ilhem Diboun, Sarah Addou, Stefano Lise, Caroline Johnston, Antonio Sillero, Janet Thornton[1] and Christine Orengo

Biochemistry and Molecular Biology Department, University College London, University of London, Gower Street, London WC1E 6BT, UK and [1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 15, 2004; Revised and Accepted September 21, 2004

ABSTRACT

The CATH database of protein domain structures (http://www.biochem.ucl.ac.uk/bsm/cath/) currently contains 43 229 domains classified into 1467 superfamilies and 5107 sequence families. Each structural family is expanded with sequence relatives from GenBank and completed genomes, using a variety of efficient sequence search protocols and reliable thresholds. This extended CATH protein family database contains 616 470 domain sequences classified into 23 876 sequence families. This results in the significant expansion of the CATH HMM model library to include models built from the CATH sequence relatives, giving a 10% increase in coverage for detecting remote homologues. An improved Dictionary of Homologous superfamilies (DHS) (http://www.biochem.ucl.ac.uk/bsm/dhs/) containing specific sequence, structural and functional information for each superfamily in CATH considerably assists manual validation of homologues. Information on sequence relatives in CATH superfamilies, GenBank and completed genomes is presented in the CATH associated DHS and Gene3D resources. Domain partnership information can be obtained from Gene3D (http://www.biochem.ucl.ac.uk/bsm/cath/Gene3D/). A new CATH server has been implemented (http://www.biochem.ucl.ac.uk/cgi-bin/cath/CathServer.pl) providing automatic classification of newly determined sequences and structures using a suite of rapid sequence and structure comparison methods. The statistical significance of matches is assessed and links are provided to the putative superfamily or fold group to which the query sequence or structure is assigned.

DESCRIPTION OF THE CATH HIERARCHY AND CURRENT POPULATION STATISTICS

The CATH database is a hierarchical classification of domains into sequence- and structure-based families and fold groups. Table 1 shows the population of the latest release of CATH (Version 2.5.1, released January 2004). In the lowest level of the hierarchy, sequences are clustered according to significant sequence similarity (35% identity and above, the S-Level). At higher levels, domains are grouped according to whether they share significant sequence, structural and/or functional similarity (homologous superfamilies, H-Level) or just structural similarity (fold or topology group, the T-level). Fold groups sharing similar architectures, i.e. similarities in the arrangements of their secondary structures regardless of connectivity are then merged into the common architectures (the A-Level). At the top of the hierarchy, domains are clustered depending on their class, i.e. the percentage of $\alpha-$helices or β-strands (the C-Level).

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

Table 1. Populations of the different levels in the CATH hierarchy

Class	1	2	3	4	Total	(5)
A	5	19	12	1	37	(n/a)
T	227	139	361	86	813	(n/a)
H	433	286	659	89	1467	(n/a)
S	957	961	2008	110	4036	(1071)
All	9013	12 962	20 411	843	43 229	(12 475)

IMPROVED CLASSIFICATION PROTOCOLS

Below we describe some new CATH associated resources and protocols that increase the speed and reliability of classifying newly determined protein structures in the CATH database.

Validation of homologues using the CATH dictionary of homologous Superfamilies (DHS)

The CATH associated Dictionary of Homologous Superfamilies (DHS) (http://www.biochem.ucl.ac.uk/bsm/dhs/) was established in 1997 (1) and contains a variety of sequence, structural and functional information for each superfamily in CATH. It was updated recently for CATH version 2.5.1, which contains 1467 homologous superfamilies, 334 of which are populated with three or more remote homologues (<35% sequence identity). The DHS contains information on all the pairwise sequence similarities and structural similarities for all pairs of relatives in each superfamily. Sequence similarity is recorded by sequence identity and *E*-value. Structural similarity is recorded by pairwise SSAP score (2) and also, by *E*-values determined against a distribution of scores obtained by comparing all non-redundant structures with each other.

Multiple structure alignments are derived for structurally coherent subgroups of relatives, having a pairwise SSAP score of >85 against all relatives in the subgroup. These are generated using the CORA algorithm (3) and displayed using CORAplot (3). The current DHS contains 671 structural alignments from 416 superfamilies. Highly conserved sequence positions, which may be associated with functionally important sites, are highlighted.

Two new methods have been devised to illustrate the degree of structural divergence across the superfamily. Both exploit a multiple structure alignment to identify equivalent secondary structures across the superfamily and inserted secondary structures. Plots give information on highly conserved secondary structures that are diagnostic for the particular superfamily and on the degree of structural embellishment occurring in diverse relatives. Putative homologues to a particular CATH superfamily can be aligned against structural relatives in order to determine whether their structural characteristics fall within the range of structural diversity observed across the superfamily. Information on the population of the superfamily is also provided so that users can gauge how well the superfamily has been sampled to date.

Functional annotations are also provided for each superfamily in the DHS by recruiting relevant functional data from the Protein Data Bank (PDB) (4), GenBank (5), ENZYME (6), KEGG (7) and Gene Ontology (8) databases. The more than 10-fold expansion in the extended CATH database (from 43 299 CATH structural domain sequences to 616 470 by including related GenBank sequences and genome sequences) has significantly increased the amount of functional data available for a particular superfamily.

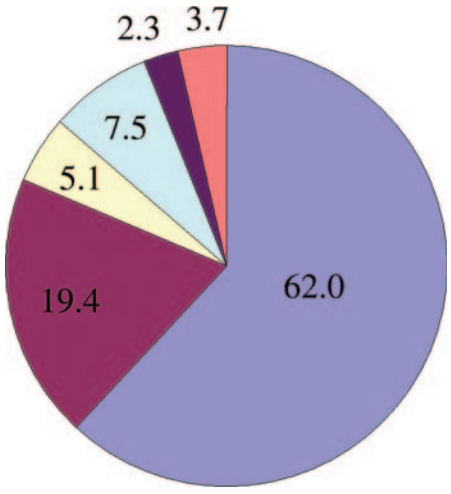

Figure 1. The proportion (%) of structures from the PDB that have been classified in CATH over the last two years using different sequence comparison or structure comparison methods. Blue segment: PDB sequences with 95% sequence identity or more to existing CATH domains, recognized using SSEARCH. Magenta segment: PDB sequences with 30% sequence identity or more to existing CATH domains, recognized using SSEARCH. Yellow segment: PDB entries that can be assigned to existing CATH superfamilies by scanning the HMM library. Green segment: PDB entries that can be assigned to CATH superfamilies by structure comparisons against CATH representatives using SSAP. Purple segment: PDB entries that can be assigned to CATH fold groups by structure comparisons against CATH representatives using SSAP. Orange segment: PDB entries that do not match any CATH structure and represent novel folds.

Expansion in the functional information together with more informative descriptions of structural variability in each CATH superfamily considerably assists in validating new homologues classified in CATH. Furthermore, links to the DHS are provided for structural matches identified using the CATH server.

Improved detection of remote homologues using an extended CATH-HMM model library

Profile based methods for sequence comparison were developed in the early 1980s and allowed recognition of more distant homologues than pairwise based approaches (9). Benchmarking of several publicly available methods, including those using position-specific scoring matrices and hidden Markov models (HMMs) have been undertaken by several groups (10,11). These approaches used datasets of distant homologues selected from the structural classifications, such as SCOP and CATH, to determine the sensitivity of various profile based methods, e.g. HMMs (12) and PSI-BLAST (13).

We recently used a dataset of remote structural homologues from the CATH database (<35% sequence identity), which had been validated by structure comparison and manual inspection to assess the performance of several HMM based strategies (Strategies for Improved Fold and Superfamily Recognition in Genome Annotation; I. Sillitoe, personal communication). HMMs were built using the SAM-T technology developed by Karplus *et al.* (14). A total of 23 876 HMM models were built for representative sequences from each sequence family in the extended CATH database (containing 616 470 domain sequences). The extended model library gives a 10% increase in coverage for remote homologue detection

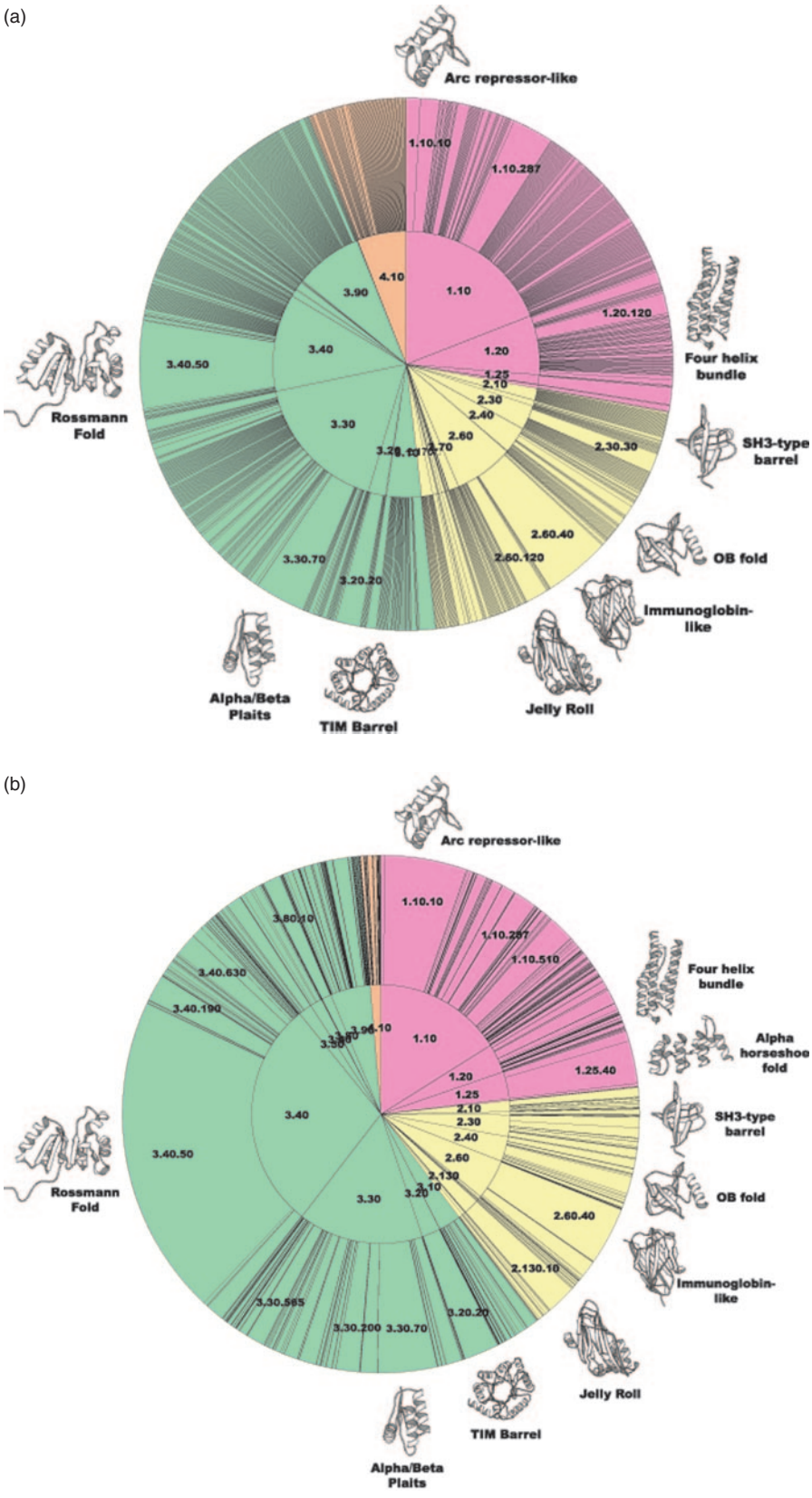

Figure 2. CATHerine wheels (**a**) illustrating the distribution of domain structures from the PDB among the different levels in the CATH hierarchy. The three classes are illustrated in colour, mainly α pink, mainly β yellow and α−β green. The inner wheel corresponds to different architectures in the classification and the outer wheel to different fold groups. Each fold group has been subdivided according to the numbers and populations of different homologous superfamilies adopting that fold. (**b**) Illustrating the distribution of CATH domains among the sequences from 150 completed genomes, in Gene3D. In this case, the fold groups labelled in the outer circle have been divided according to the number and size of close sequence families within each fold group.

compared to the standard CATH HMM model library, with a low error rate (0.1%) (I. Sillitoe, personal communication).

It can be seen from Figure 1 that on average, nearly 87% of homologues classified in CATH over the last two years could be recognized using sequence comparison methods, both pairwise sequence alignment and scans against the more sensitive extended CATH-HMM model library.

Expansion of CATH with sequence relatives from completed genomes and domain partnership information

We have recently devised protocols for identifying sequence relatives to CATH superfamilies in completed genomes (15). To date, nearly one million sequences from 150 completed genomes have been scanned against the CATH-HMM model library (15). Between 40 and 60% of sequences or partial sequences from each genome could be assigned to a CATH superfamily. Genome sequences were also scanned against libraries of HMM models from the Pfam database (release 10) (16) in order to extend the domain annotation of each genome sequence and provide more comprehensive information on domain partnerships.

Sequence relatives to CATH superfamilies, identified in this way are displayed in the CATH related DHS and Gene3D resources. Gene3D displays the domain composition of each gene annotated by CATH and Pfam domains. CATH family data in the Gene3D resource has revealed some intriguing insights into the expansion of superfamilies involved in metabolism and regulation in bacterial genomes (17).

Figure 2 shows that the power-law like trends first detected in the structural classifications are mirrored when sequence relatives from the genomes are also included. Considering the structural data alone, it can be seen from Figure 2a that fewer than 10 of the most highly populated folds in the CATH database account for nearly 25% of all superfamilies in the PDB. These folds were previously described as superfolds as they are adopted by many diverse homologous superfamilies (18). When genome sequences are included it can be seen from Figure 2b that the same fold groups dominate the genomes, as they are adopted by nearly 45% of all close sequence families (relatives have 35% or more sequence identity), of known structure, in the genomes.

THE CATH SERVER

A new protocol has been developed for searching CATH with a newly determined protein structure. Structures submitted to the server (http://www.biochem.ucl.ac.uk/cgi-bin/cath/CathServer.pl) are first processed by the DDMake suite of programs that generate derived data from the PDB coordinate files (e.g. secondary structure data, residue accessibilities and $\phi\psi$ data, sequence data in the FASTA format, etc.). The query sequence is scanned against the CATH-HMM model library to identify more remote homologues. Threshold *E*-values used to recognize homologues are predetermined by benchmarking with validated structural homologues from CATH (I. Sillitoe, personal communication).

If the sequence returns a significant match to any relative in one or more CATH superfamilies, representatives from all close sequence families within those superfamilies are structurally compared with the query structure using the SSAP structure alignment program (2). The top 10 structural matches, sorted in the order of SSAP score are then displayed together with information on the degree of sequence and structural similarity and with links to the CATH page and the DHS page for each CATH superfamily identified. Rasmol images are also provided for the top 10 matches.

Any query structure unmatched by the CATH-HMM library is scanned against a library of representative structures from each close sequence family in CATH using the rapid structure comparison algorithm, CATHEDRAL (19). CATHEDRAL uses a robust statistical framework based on the extreme value distributions observed for random similarities to assess significance. If the query structure significantly matches one or more CATH superfamilies, SSAP comparisons are performed for all sequence representatives in those superfamilies and the top 10 matches are displayed, as before.

ACKNOWLEDGEMENTS

F.P., I.S., M.D., A.G., T.L., A.A. and C.O. all acknowledge the Medical Research Council for their funding. A.T., D.L. and R.M. are currently supported by funding from the National Institutes of Health. G.R., O.R. and T.D. acknowledge support from the Biotechnology and Biological Sciences Research Council, and C.B. acknowledges support from the Wellcome Trust for the research described in this manuscript.

REFERENCES

1. Bray,J.E., Todd,A.E., Pearl,F.M., Thornton,J.M. and Orengo,C.A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.*, **13**, 153–165.
2. Taylor,W. and Orengo,C. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.
3. Orengo,C. (1999) CORA—topological fingerprints for protein structural families. *Protein Sci.*, **8**, 699–715.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, 23–26.
6. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
7. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
8. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
9. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
10. 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.
11. Pearl,F.M., Lee,D., Bray,J.E., Buchan,D.W., Shepherd,A.J. and Orengo,C.A. (2002) The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Sci.*, **11**, 233–244.

12. Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
13. Altschul,S.F., Madden,T.L., Schäffer,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.
14. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
15. Lee,D., Grant,A., Marsden,R. and Orengo,C. (2004) Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins*, in press.
16. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
17. Ranea,J.A., Buchan,D.W., Thornton,J.M. and Orengo,C.A. (2004) Evolution of protein families and bacterial genome size. *J. Mol. Biol.*, **336**, 871–887.
18. Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.
19. Harrison,A., Pearl,F., Sillitoe,I., Slidel,T., Mott,R., Thornton,J. and Orengo,C. (2003) Recognizing the fold of a protein structure. *Bioinformatics*, **19**, 1748–1759.