Institution: University College London / Birkbeck College

Unit of Assessment: 5 - Biological Sciences

Title of case study: CATH structural classification of proteins aids drug discovery in the pharmaceutical industry

1. Summary of the impact (indicative maximum 100 words)

The CATH classification of protein structure, developed at the Institute of Structural and Molecular Biology, UCL, by Janet Thornton and Christine Orengo, has been used widely across the pharmaceutical industry and academia to guide experiments on proteins. This has led to significant cost and time savings in drug discovery. The UCL-hosted online CATH database receives around 10,000 unique visitors per month, and is a partner in InterPro – the most frequently accessed protein function annotation server available.

2. Underpinning research (indicative maximum 500 words)

Less than 10% of proteins have detailed experimental characterisation – even in human – and computational approaches have emerged to predict the function of a protein by identifying evolutionarily related proteins (homologues) whose functions are likely to be similar and which have already been experimentally characterised (e.g. in fly or worm).

One of the first classifications of proteins, CATH (denoting protein Class, Architecture, Topology, Homologous superfamily), was established by Orengo and Thornton in 1994 **[1, 2]**. CATH groups homologues according to their structural and therefore likely functional similarity, using a combination of automated and manual procedures. The major level in CATH (homologous Superfamily) groups together proteins which have clearly descended from a common ancestor. The early success of this work led to Orengo being awarded an MRC Senior Fellowship in 1995 (for 10 years in total) to extend the classification. Further research led to powerful algorithms for predicting which genome sequences could be classified in CATH. CATH now classifies ~20 million protein sequences – ~70% of domain sequences from 2,000 completed genomes and ~60% of domain sequences from human.

In 1995, individual structures in CATH were provided with highly valuable additional data in the form of PDBsum webpages **[3]** with LIGPLOT analyses **[4]**, developed by postdoctoral researcher Roman Laskowski in the Thornton group, and the ProCHECK suite was developed to assess structure quality **[5]**.

CATH has been involved in several national and international consortia and Networks of Excellence providing structural/functional annotations for proteins. These integrated multiple resources to improve confidence and ensure much greater coverage. TrEMBLOR, Integr8, IMPACT, and BIOSAPIENS consortia disseminated the integrated data widely to academia and industry via web sites and web servers. Other consortia (e.g. ENFIN, IMI-EUROPAIN) promoted collaborations with experimental groups in academia and industry, to ensure that the function prediction tools were adopted more widely by experimentalists working on biological systems related to human health and disease.

Since 2000 the Thornton and Orengo groups have exploited CATH in collaboration with structural biology groups in the NIH-funded Protein Structure Initiative (PSI) – a structural genomics initiative which targets proteins in bacteria, associated with pathogenesis, for structure determination. PSI is a large strategic project, (funding of >$150 million) and involving >20 research groups and several hundred scientists across USA and Europe.

In 2009, further research allowed sub-classification of functional families in CATH. CATH function prediction methods ranked 7th worldwide (out of 56) in an international assessment **[6]**. The unique combination of structure and sequence data in CATH allows accurate detection of functionally important sites to guide mutagenesis experiments and explain the damaging effects of mutations associated with disease.

CATH is being integrated with the only other world-leading structure classification, SCOP, to give the most comprehensive classification available (Genome3D). Other major structure prediction resources are being included (Gene3D, SUPERFAMILY, GenTHREADER, PHYRE) to provide consensus annotations via a common web portal. Genome3D is expected to be widely used by Industry.

Christine Orengo has worked at UCL since 1991 and is currently Professor of Bioinformatics; Janet Thornton was at UCL until 2001, when she moved to the European Bioinformatics Institute.

3. References to the research (indicative maximum of six references)

[1] Orengo CA, Jones DT, Thornton JM. Protein Families and Domain Superfolds. Nature. 1994, 372, 631-634. http://dx.doi.org/10.1038/372631a0

[2] Orengo CA, Michie AD, Jones DT, Swindells MB Thornton JM. CATH: A Hierarchic Classification of Protein Domain Structures. Structure. 1997, 5, 1093-108. http://dx.doi.org/10.1016/S0969-2126(97)00260-8

[3] Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci. 1997, 22, 488-90. http://dx.doi.org/10.1016/S0968-0004(97)01140-7

[4] Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng.1995, 8, 127-134. http://dx.doi.org/10.1093/protein/8.2.127

[5] Laskowski RA, MacArthur MW, Moss DS, Thornton JM. *PROCHECK*: a program to check the stereochemical quality of protein structures *J. Appl. Cryst.* 1993. 26, 283-291. http://dx.doi.org/10.1107/S0021889892009944

[6] Radivojac, P, Clark, W T, Oron, TR, Schnoes, AM, Wittkop T, Sokolov A, Rentzsch R, Orengo, CA, et al. A large-scale evaluation of computational protein function prediction. Nature Methods 2013 10(3), 221-227. http://dx.doi.org/10.1038/nmeth.2340

There are more than 2,195 citations for the main CATH Classification publication in Structure [2] and a total of 7,058 citations for all publications reporting developments to the CATH classification and insights derived from CATH analyses. PROCHECK and PROCHECK-NMR have been cited 16,337 and 3,205 times, respectively. LIGPLOT has been cited 2,331 times.

Fellowships funding this work

- Topological Fingerprints for Protein Fold Families and Their Application to Structure Analysis, Prediction and Determination. MRC Senior Fellowship. G117/97. £550,181 (Nov 1995 – Nov 2000) (Professor Orengo).
- Predicting structures and functions for the genomes. MRC Senior Fellowship. Ref. G117/384. £620,000 (Nov 2000 – Nov 2005) (Professor Orengo).

4. Details of the impact (indicative maximum 750 words)

Breadth of use of CATH, PDBsum and LIGPLOT

The CATH classification is made available on a UCL-hosted website, maintained by the Orengo group **[a]**. This internationally renowned resource is one of the leading protein structure classifications in the field. Web access to CATH ranges from 8,900 (Google Analytics) to 22,500 (Webalyzer) unique visits per month and the number of pages accessed varies from 85,000 to 2m per month depending on the method used for analysing access. Two thirds of these web accesses are from industry-based sites **[b]**.

CATH data is further disseminated through DAS and the InterPro web server, at the European Bioinformatics Institute (EBI). InterPro is one of the most widely used web portals by biologists in industry and academia, with more than five million web page accesses per month. It combines protein family data from multiple resources to assign greater confidence. DAS was established by 30 European partners as part of the Biosapiens network, headed by Janet Thornton whilst the InterPro server is being developed by a consortium of eight European partners, including CATH. Information from CATH is also disseminated via the web portal of the international Protein Databank (PDB) resource **[c]**, the primary source of protein structures. Further links to CATH are provided by many international web-based computational biology resources (e.g Pfam, BRENDA **[b]**).

Thornton and Orengo have given talks on CATH and PDBsum in Europe, the United States, Japan, South Korea and India, including through the EBI Industry program **[d]**, and to computational biologists at several pharmaceutical companies. Orengo spoke at a Gordon Conference on drug design in 2009, which was attended by a significant number of researchers from the pharmaceutical industry **[e]**.

CATH, PDBsum and the threading algorithm were three of the four major UCL bioinformatics resources used to establish the UCL company Inpharmatica in 1998. This was involved in predicting structures and functions for proteins via the 'Biopendium'. Inpharmatica sold this and other related software packages to several large pharmaceutical companies including Pfizer, Astra Zeneca and Glaxo-Wellcome. Inpharmatica was acquired by Galapagos in 2006 **[f]**.

More recently, the latest structure comparison algorithms underpinning CATH (CATHEDRAL) and LIGPLOT have been distributed directly to pharmaceutical companies including *[Text removed for publication]*. Together these tools have generated *[Text removed for publication]* in licence income to UCL within the census period **[g]**. PDBsum is widely used by pharmaceutical companies with 34,000 unique visitors/month in total (1.1m web hits).

[Text removed for publication]

Papers published by Thornton and Orengo have been cited 13 times across 11 patent documents published in the assessment period, indicating the commercial relevance of their work. The patents are filed across the USA, Europe and Internationally through the PCT system and are assigned to GSK Ltd, Biogen Idec Inc. and Pharnext **[i]**.

Specific applications to problems in the pharmaceutical industry

CATH is routinely used by the pharmaceutical industry to identify the structures of proteins implicated in disease. CATH prediction methods and domain assignments are widely used by industry, as are the methods for analysing the structures (e.g. LIGPLOT, PROCHECK). CATH is widely used by researchers in the pharmaceutical industry to explore protein structure function relationships and to aid in drug design. Structure is highly conserved during evolution and 3D information can give better clues to the molecular mechanisms associated with a protein's function than purely sequence data. Since very few human proteins are experimentally characterised (<10%), CATH and related resources are used to search for homologues with known, experimentally validated functions and the structural data associated with these relatives can be then be exploited to build 3D models for the human proteins. In addition, the CATH classification can be searched with the structures of proteins which are potential drug targets to identify close relatives which might also bind the target drug, giving rise to side effects.

[Text removed for publication]

CATH has been exploited by the NIH-funded, international, structural genomics initiatives to select proteins from pathogenic organisms for structure determination to aid drug design. Since 2000, more than 3,000 protein structures have been solved by these initiatives representing about 30% of the unique structures deposited into the PDB from all sources, worldwide, during this period. These structures have shed light on how structure is linked to function and provided important details of binding sites in proteins implicated in cancer and pathogen associated diseases.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[a] CATH classification website: http://www.cathdb.info/

[b] Web stats report, available on request; includes URLs for major resources linking to CATH and PDBsum, including InterPro, PDB, DAS, Knowledgebase for PSI, Pfam.

[c] Letter confirming PDB hosting of CATH and its usefulness from the Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey USA. Copy available on request.

[d] EBI industry programme: http://www.ebi.ac.uk/industry

[e] Gordon Research Conference on Computer Aided Drug Design, 2009, Programme: http://www.grc.org/programs.aspx?year=2009&program=cadd

[f] Details of sale of Inpharmatica: www.uclb.com/what-we-do/companies/inpharmatica

[g] A report from UCL Business PLC on commercial licences is available on request.

[h] [Text removed for publication]

[i] Report from Cambridge IP Ltd. Copy available on request.

[j] [Text removed for publication]

[k] [Text removed for publication]

[l] [Text removed for publication]