Institution: UCL

Unit of Assessment: 5 - Biological Sciences

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

Period when the underpinning research was undertaken: 2000 - 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Christine Orengo	Professor of Bioinformatics	1991 - present
Janet Thornton	Professor of Bioinformatics	1986 - 2006
Ian Sillitoe	Principal Research Associate	2005 - present

Period when the claimed impact occurred: 1 August 2013 - 31 December 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of Impact

The CATH (Class, Architecture, Topology (fold family), Homologous superfamily) classification of protein domains, was developed at UCL's Institute of Structural and Molecular Biology by Professors Christine Orengo and Janet Thornton. This led to the development of the UCL-hosted online CATH database, which receives >22,500 unique visitors per month. CATH is also a partner resource in InterPro – the most frequently accessed protein function annotation server, with approximately 716,000 unique visitors per year. Outside academia, CATH is widely used across the global pharmaceutical industry for drug design and research and development. It is also used to assess impacts of mutations in proteins supporting clinical diagnostics (e.g. hypercholesterolemia). CATH has informed policy on the host range of SARS-CoV2, and led to significant efficiencies in drug discovery.

2. Underpinning research

Proteins are involved in all major biological processes. Knowing their structure and function is essential for detecting pathogenic changes, for example mutations in protein sites, and for designing drugs. However, less than 10% of proteins have detailed experimental characterisation, even in humans. CATH algorithms predict the structures and functions of proteins by identifying evolutionarily related proteins (homologues), the properties of which are likely to be similar and which have already been experimentally characterised – such as in *Drosophila* (fruit flies) or *C. elegans* (nematode worms).

The accuracy of functional inheritance in CATH is enhanced by explicitly sub-classifying evolutionary superfamilies into functional families (FunFams). This is done using entropy-based algorithms to distinguish groups of relatives having differentially conserved residues likely to be important for function (FunSites) [R1, R2].

In a medical context, knowing how close disease-associated mutations are to functional sites (FunSites) can explain their damaging effects and suggest likely pathogenesis. CATH algorithms exploited this evolutionary and functional data to detect putative cancer genes, data from which are valuable for personalised therapeutics [R3], and identified mutations implicated in antibiotic resistance. Furthermore, FunFams can facilitate drug repurposing to target disease genes, by providing valuable data for pharmaceutical companies interested in repurposing as a cost effective mechanism for selecting drugs. FunFams can also identify drug targets which are less likely to be associated with side effects [R4], providing information that is valuable for drug design. CATH methods and data are being exploited in the NHS-funded Genomics England Functional Effects Domain, and in a large-scale analysis of lung cancer data to uncover mechanisms of cancer evolution: the GBP 14m, 9-year Cancer Research UK-funded TracerX project.

FunFam classification also allows accurate detection of functionally important sites to guide mutagenesis experiments for synthetic biology and is being used to enhance functional sites in bacterial enzymes capable of degrading plastics and pesticides. It has also highlighted sites involved in SARS-CoV-2 infection [R5].

CATH computational approaches exploit very large-scale protein structure and sequence data. Homologues are grouped into superfamilies where structural and sequence similarity indicate descent from a common ancestor. Classification was initially based on structural data and Orengo and colleagues then developed sequence-based protocols to assign genome sequences to particular superfamilies, expanding the classification by >300-fold. CATH currently recognises 150 million domain sequences in approximately 5,500 superfamilies, accounting for approximately 70% of sequences from completed genomes and approximately 60% from human. It is widely used by researchers to infer structural and functional properties (papers cited >11,000 times).

Sub-classification into functional families (FunFams) considerably extended CATH's value. Recent machine learning algorithms have exploited FunFams to improve detection of functional sites (FunSites) [R2] and FunSites are being incorporated in two highly accessed resources (PDBe, with 60,000 users per month and UniProt, with 135,000 users per month at the European Bioinformatics Institute). Both capture this data to facilitate disease diagnostics and personalised medicine.

CATH is the only resource which is capable of performing functional sub-classification on such a large scale, identifying 220,000 families each with at least one experimentally characterised protein. Validation has shown high structural and functional coherence across FunFams, allowing much more accurate predictions to be made. CATH methods ranked in the top three (out of 150) in international assessments of molecular function prediction [R6], and first in 2020.

CATH was integrated with another world leading structure classification, SCOP, to give the most comprehensive protein structure classification information available (Genome3D) providing consensus structural data valuable for pharma and biotech companies.

3. References to the research

[R1] Das, S., Lee, D., Sillitoe, I., Dawson, N.L., Lees, J.G., Orengo, C.A. (2015). 'Functional classification of CATH superfamilies: a domain-based approach for protein function annotation'. *Bioinformatics*. **31**(21), 3460-7. DOI: http://doi.org/10.1093/bioinformatics/btv398. Epub 2015 Jul 2.PMID: 26139634.

[R2] Das, S., Scholes, H.M., Sen, N., Orengo, C. (2020). 'CATH functional families predict functional sites in proteins'. *Bioinformatics*. Nov 2:btaa937. DOI: http://doi.org/10.1093/bioinformatics/btaa937. Online ahead of print. PMID: 33135053.

[R3] Ashford, P., Pang, C.S.M., Moya-García, A.A., Adeyelu, T., Orengo, C.A. (2019). 'A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations'. *Scientific Reports*. **9**(1), 263. DOI: http://doi.org/10.1038/s41598-018-36401-4.

[R4] Moya-García, A., Adeyelu, T., Kruger, F.A., Dawson, N.L., Lees, J.G., Overington, J.P., Orengo, C., Ranea, J.A.G. (2017). 'Structural and Functional View of Polypharmacology'. *Scientific Reports*. **7**(1), 10102. DOI: http://doi.org/10.1038/s41598-017-10012-x.

[R5] Lam, S.D., Bordin, N., Waman, V.P., Scholes, H.M., Ashford, P., Sen, N., van Dorp, L., Rauer, C., Dawson, N.L., Pang, C.S.M., Abbasian, M., Sillitoe, I., Edwards, S.J.L., Fraternali, F., Lees, J.G., Santini, J.M., Orengo, C.A. (2020). 'SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals'. *Scientific Reports*. **10**, 16471.DOI: http://doi.org/10.1038/s41598-020-71936-5

[R6] Zhou, N., Jiang, Y.,…Rost, B., Brenner, S.E., Orengo, C.A., Jeffery, C.J., Bosco, G., Hogan, D.A., Martin, M.J., O'Donovan, C., Mooney, S.D., Greene, C.S., Radivojac, P., Friedberg, I. (2019). 'The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens'. *Genome Biology*. **20**(1), 244. DOI: http://doi.org/10.1186/s13059-019-1835-8.

4. Details of the impact

The CATH classification, made available on a UCL-hosted website maintained by the Orengo group, is an internationally renowned resource and one of the world's leading protein structure classifications in this field [S1]. There are more than 22,500 unique web visits to CATH per month and 2,000,000 pages are accessed per month [S1]. Two-thirds of these web visits are from industry-based sites [S1]. CATH has been made an ELIXIR Europe-wide Core Data Resource (CDR) [S2], one of only three European national resources to be endorsed as meeting the highest standards in data quality and data access established by the ELIXIR (European Strategy Forum on Research and Innovation), - excellent science, community served, high quality of service and impact - and, notably, the only CDR to be endorsed in the UK. [TEXT REMOVED FOR PUBLICATION].

CATH data is further disseminated through the InterPro web server, at the European Bioinformatics Institute (EBI). Outside academia, InterPro is one of the most widely used web portals by biologists in industry, with over 723,000 unique visitors per year. It combines protein family data from multiple resources to assign greater confidence. CATH data is also disseminated via the web portal of the international protein structure resource, the Protein Databank (PDB), with over 4,411,871 unique users/year, and UniProt, a major source of protein functional data with over 10,800,000 unique users/year (2,220,00 of which are from industry). Further links to CATH are provided by many international web-based computational biology resources, for example Pfam, BRENDA [S1]).

CATH was one of four major resources used to establish the UCL spinout company Inpharmatica. It was also used by Inpharmatica to provide data and services for predicting protein structures and functions for large pharmaceutical companies including Pfizer, AstraZeneca, Merck and Glaxo-Wellcome [S3]. [TEXT REMOVED FOR PUBLICATION] Inpharmatica has since been acquired by biotechnology company Galapagos. Since 2013, Galapagos has attracted over GBP11,000,000 in investment [S4].

The structure comparison algorithms underpinning CATH (CATHEDRAL) have been distributed directly to pharmaceutical companies, including UCB Celltech and Cubist Pharmaceuticals, generating GBP34,931 in licence income to UCL within the census period [S5]. Papers published by Orengo have been cited across eight patent documents, demonstrating the commercial relevance of the work.

Users in industry exploit the CATH prediction methods, domain assignments and methods (e.g. CATHEDRAL) for analysing their structures. For example, R&D staff in DSM Chemical Technology R&D BV used CATHEDRAL to search CATH for structural homologues of a protein they were solving; a biocatalyst valuable for the pharmaceutical industry because of its ability to synthesis various beta-hydroxy amino acids - compounds widely used in food processing and drug development [S6].

Specific applications to problems in the pharmaceutical industry

Identifying disease drug targets: CATH is routinely used by the pharmaceutical industry to identify the structures of proteins implicated in disease. Predicted functional site data is used to assess the effects of disease-associated residue mutations and to identify putative ligand (i.e. drug) binding sites and variation in drug-binding modes across families. CATH has been used, for example, by:

- **Gene Tools LL and BrainMicro LLC**, to identify domain boundaries in enzymes being structurally characterised as potential therapeutic targets for toxoplasmosis and other parasitic infections [S7];
- **Phylogica Ltd** to search fold space in the characterisation of a peptide library used for target discovery, and for identifying new cell-penetrating peptides [S8];
- **Acpharis**, to identify the domains of kinases - a major drug target family for pharmaceutical companies [S9]; and
- **AstraZeneca**, for structural analyses that identified pharmacophore binding motifs for NAD (Nicotinamide adenine dinucleotide– an critical component in metabolism) and its analogues across different protein families [S10].

Biologic drug design: CATH data is also being used for biologics in large pharma companies. [TEXT REMOVED FOR PUBLICATION].

Pathogen protein structure-based drug design: CATH FunFam technology was exploited by NIH-funded, international, structural genomics initiatives (between 2013 and 2017) to select proteins from pathogenic organisms for structure determination to aid drug design. More than 2,400 protein structures were deposited into the Protein Databank 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.

Covid-19: More recently, CATH FunFam analyses of binding sites involved in SARS-CoV-2 infection of animal hosts was used by the WHO and a UN Food and Agriculture Organisation policy unit [S11] in strategy discussions on animals at risk from infection, or which are likely to become reservoirs for the virus. These sites constitute major mechanisms of infection which are targetable by drugs. The work was also reported in several newspapers globally.

Orengo has given talks on CATH in Europe, the United States, Singapore, China, Malaysia, Africa and India, including the EBI Industry program and to computational biologists at several pharmaceutical companies.

5. Sources to corroborate the impact

[S1] CATH classification website: http://www.cathdb.info/
Web stats report, available on request; includes URLs for major resources linking to CATH including InterPro, PDB, UniProt, Pfam.
[S2] Letter from ,[TEXT REMOVED FOR PUBLICATION] confirming CATH is CDR and its usefulness.
[S3] Letter from [TEXT REMOVED FOR PUBLICATION]
[S4] Letter from [TEXT REMOVED FOR PUBLICATION].
[S5] A report from UCL Business PLC on commercial licences is available on request.
[S6] Paper - DOI: https://doi.org/10.1371/journal.pone.0124056
[S7] Paper - DOI: https://doi.org/10.3389/fcimb.2018.00352
[S8] Paper - DOI: https://doi.org/10.1016/j.cbpa.2017.03.016
[S9] Paper - DOI: https://doi.org/10.1021/acs.jmedchem.9b00089
[S10] Paper - DOI: https://www.ncbi.nlm.nih.gov/pubmed/23889609
[S11] UN FAO policy document titled 'Exposure of humans or animals to SARS-CoV-2 from wild, livestock, companion and aquatic animals' (https://doi.org/10.4060/ca9959en)