Institution: Birkbeck, University of London

Unit of Assessment: UoA 11 Computer Science and Informatics

Title of case study: Informatics support for the management and integration of large-scale life sciences data

1. Summary of the impact

Research carried out at Birkbeck's Department of Computer Science and Information Systems since 2000 has produced techniques for the management and integration of complex, heterogeneous life sciences data not previously possible with large-scale life sciences data repositories. The research has involved members of the department and researchers from the European Bioinformatics Institute (EBI) and University College London (UCL) and has led to the creation of several resources providing information about genes and proteins. These resources include the BioMap data warehouse, which integrated the CATH database – holding a classification of proteins into families according to their structure, the Gene3D database – holding information about protein sequences, and other related information on protein families, structures and the functions of proteins such as enzymes. These resources are heavily utilised by companies worldwide to explore relationships between protein structure and protein function and to aid in drug design.

2. Underpinning research

The common theme in the research underpinning this case study has been the development of techniques to support the integration, maintenance, analysis and mining of distributed, highly heterogeneous, large-scale resources holding life sciences data. (See http://www.ebi.ac.uk/luscombe/docs/imia_review.pdf for an overview of the computational issues arising in the life sciences and the common terminology used.)

The CATH classification of protein structures aims to aid the prediction of the molecular functions of proteins by identifying structurally similar proteins that are likely to be also functionally similar. A first version of CATH was developed in the 1990s by Professors Christine Orengo and Janet Thornton at UCL's Department of Biochemistry and Molecular Biology, but the initial file-based system used to hold the data limited the ability to integrate related data on protein sequences, families and functions. This initial problem led to research in 2000-2001 at Birkbeck that investigated techniques to enable the integration of data on the evolution, structure and functions of proteins with protein sequence data in order to improve understanding of the molecular nature of disease. Birkbeck staff Nigel Martin and Roger Johnson worked in collaboration with Prof. Orengo from UCL. A flexible graph-oriented approach was developed to enable the integration of protein sequence data with related data about protein structures, protein functions, and taxonomic data such as classifications of the terms used to describe protein functions. The approach was based on a graph representation of the relationships between entities through meta-level tables, with application-independent semantics, together with an application-dependent base table level, a materialized view level for query performance enhancement, and a further application-dependent view level enabling applications to access the integrated data in ways best suited to their data analysis requirements. This graph-oriented approach was used to develop a relational database for holding the CATH classification of protein structures [1].

A further problem arose from the increasing quantity and complexity of life sciences data being generated from genome sequencing projects and technologies allowing scientists to generate their own data, for example gene expression data measuring the activity of individual genes. This data was typically stored in independent repositories which needed to be integrated in order to support the analysis and mining of their combined information. Integrating such repositories gives rise to two significant problems. First, the data to be integrated is often highly heterogeneous, with inconsistent identifiers for the same biological entity, making the task of identifying and merging related data a complex one. Second, the repositories from which the integrated data is obtained continue to be updated, so the problem arises of how the integrated repository can be "synchronised" so that it correctly reflects changes to the data in the source repositories. For

example, gene sequence data extracted from a primary repository held at a site such as the EBI could be copied to and integrated with the specialised protein structure classification data at UCL, but keeping the UCL integrated data synchronised to reflect updates to the primary EBI repository was a challenging problem. Research undertaken from 2001 to 2007 developed solutions for overcoming these problems. The Birkbeck staff involved were N. Martin, A. Poulovassilis, A.Shepherd (Postdoc February to September 2002), M.Maibaum (Postdoc April 2003 to July 2004), G.Rimon (Postdoc March 2003 to January 2005), E.Sideris (Postdoc December 2005 to July 2007) and H. Fan (PhD student October 2001 to September 2005, and part-time RA October 2004 to December 2005). Our collaborators included Prof. Orengo from UCL and Dr. P. Keller from the EBI.

For this research, we began by extending the facilities for heterogeneous data transformation and integration provided by the AutoMed system [2] (developed in the early 2000s at Birkbeck and Imperial College, with lead investigators Poulovassilis and Dr P. McBrien from Imperial). We designed a new data clustering approach that enabled the integration of data sources which may have inconsistent identification of biological entities [3]. With this approach, each data source may be either a structured database, such as a relational database or a semi-structured file. Additional data resources are created to hold the clustering information, which are maintained as relational databases. Access to the data sources and to the Cluster data resources is handled in the same way using AutoMed, and any number of data sources and clustering methods can be integrated.

These new techniques were used to enable the creation and incremental maintenance of the BioMap data warehouse at UCL [4], which integrates heterogeneous data from multiple repositories. Starting from the MGED MAGE-OM object model and the ArrayExpress relational model for the deposition of gene expression data developed at the EBI, we designed an extended object model and an associated relational model to support the searching of gene expression data and its integration with related protein family, structure and function data. Based on these models, the data sources that were integrated into BioMap included the CATH classification of protein structures [5], the Gene3D database providing comprehensive information about the structure and function of most available protein sequences [6], and atomic-level protein structure data from the MSD data warehouse at the EBI.

The integration of the classification of protein structures of CATH with the annotated sequence data of Gene3D enabled structural information to be assigned to many protein sequences without a known structure, giving insight into the likely function of those sequences. The combined CATH/Gene3D resource is now widely used by researchers in the pharmaceutical industry to explore the relationships between protein structure and protein function, and to aid in drug design.

In summary, the research carried out at Birkbeck has produced techniques for the management and integration of complex, highly heterogeneous resources holding life sciences data not previously possible with large-scale life sciences data repositories.

3. References to the research

Publications (Birkbeck authors shown in bold)

[1] PFDB: A generic protein family database integrating the CATH domain structure database with sequence based protein family resources. **A J Shepherd, N J Martin, R G Johnson**, P Kellam, C A Orengo. *Bioinformatics*, 18, 2002, pp 1666-1672. DOI: 10.1093/bioinformatics/18.12.1666

[2] Data Integration by Bi-Directional Schema Transformation Rules. P McBrien, **A Poulovassilis**. Proceedings 19th International Conference on Data Engineering (ICDE 2003), pp 227-238. DOI: 10.1109/ICDE.2003.1260795

[3] Cluster based integration of heterogeneous biological databases using the AutoMed toolkit. **M Maibaum, L Zamboulis, G Rimon**, C.Orengo, **N Martin, A Poulovassilis**. Proceedings 2nd International Workshop Data Integration in the Life Sciences (DILS 2005), pp 191-207. DOI:

10.1007/11530084_16

[4] BioMap: Gene Family based Integration of Heteregeneous Biological Databases Using AutoMed Metadata. **M Maibaum, G Rimon**, C Orengo, **N Martin, A Poulovasillis**. Procedings 15th International Workshop on Database and Expert Systems Applications (DEXA 2004), pp 384-388. DOI: 10.1109/DEXA.2004.1333504

[5] The CATH database: an extended protein family resource for structural and functional genomics. F M G Pearl, C F Bennett, J E Bray, A P Harrison, **N Martin**, A Shepherd, ISillitoe, J Thornton, C A Orengo.*Nucleic Acids Res*, 31(1), 2003, pp 452-455. DOI: 10.1093/nar/gkg062

[6] Gene3D: comprehensive structural and functional annotation of genomes. C Yeats, J Lees, A Reid, P Kellam, **N Martin**, X Liu and C Orengo. *Nucleic Acids Research*, 36, 2008, pp D414-D418. DOI: 10.1093/nar/gkm1019

Research Grants received by Department of Computer Science & Information Systems, Birkbeck

Project name: Structural and Functional Annotation of Genome Data through Synchronised Data Warehouses.
Funder: BBSRC.
Duration: 2001-2003.
Amount: £65,356
(Complementary grants to the EBI £118,064 and UCL £67,324)

Project name: Integrating transcriptomics and structural data.
Funder: Wellcome Trust.
Duration: 2003-2006.
Amount: £139,722
(Complementary grants to the EBI £167,386 and UCL £295,610)

4. Details of the impact

The research outlined in Section 2 has enabled the construction of several bioinformatics resources which have been widely used by healthcare, pharmaceutical, life sciences, bioinformatics and technology companies and research institutes worldwide:

(1) The combined CATH/Gene3D resource has very significant worldwide use (details are given in the CathDB Usage Data Analysis Report, July 2013 – see Section 5). Taking 2012 as an example, 66162 unique visitors visited the CATH/Gene3D website hosted at UCL, with the 6 major sources of visitors being the UK, USA, Ukraine, India, Denmark and Japan. Analysis of the visitors shows significant use by commercial companies during this period:

- Use by companies in the pharmaceutical and life science sectors includes a global healthcare company which manufactures and markets pharmaceutical products and services (1212955 hits), an Indian conglomerate with a group member developing biopharmaceuticals (34214), a company that provides software testing and simulation DNA techniques (22380), a bioinformatics software company that develops software for DNA, RNA and protein sequence analysis (4913), and a company specialising in the commercialization of emerging technologies in the life sciences (731).
- Additional commercial users in the engineering and technology sectors include a petrochemical company (2695 hits), a global technology company (1775), and an information and communications technology support services company working with bioinformatics centres (672).

Use from major research institutes include Bigelow Laboratory for Ocean Sciences (53535 hits), the EBI (12655), Institute National de la Recherche Agronomique (7309) and Forschungszentrum Julich (5358).

Many visitor hits access the tools supported by the resource. These include search tools, for example to enable search of the CATH classification of protein structures, and analysis tools, for example to enable a protein structure to be submitted to identify the closest structure within the CATH classification. During 2012, the 10 most widely-used search and analysis tools supported by the resource were accessed on 4311 occasions by commercial companies and on 1764 occasions by research institutes.

(2) CATH/Gene3D is now itself integrated with other major bioinformatics databases as a partner of InterPro (http://www.ebi.ac.uk/interpro/), the primary protein family and function annotation server. InterPro is a significant international resource for structural and functional classification of proteins. It is used directly by major sequence databases and genomics projects for large-scale genome analysis, as well as for the classification of individual protein sequences via its web interface. The InterProScan service, which enables a protein sequence to be submitted for analysis of its structure and function against the InterPro member databases, averaged more than two million sequence searches per month in 2011.

(3) Links to CATH/Gene3D are provided by many other international bioinformatics resources, e.g. Protein Databank (PDB), Protein Structure Initiative (PSI), Pfam, KEGG. Since 2008 Professors Orengo and Thornton have given talks and have participated in workshops publicising CATH across Europe and in the USA, Japan, South Korea and India, including the EBI industry programme and several EMBO workshops at the EBI, which have been attended by participants from industry. They have given talks on CATH to computational biologists at several pharmaceutical companies. Prof. Orengo was a speaker at the Gordon Conference on Computer Aided Drug Design in 2009, and gave a talk in 2011 on CATH at MedImmune, the Research and Development arm of the global biopharmaceutical company AstraZeneca; these two talks publicising CATH were attended by a significant number of researchers from the pharmaceutical industry. Letters of support from the CEO of Acpharis and the former Director of the Computational Biology and Chemistry Department of the Merck Research Laboratories in Italy attest to the widespread use of the CATH resource in the pharmaceutical sector.

5. Sources to corroborate the impact

Claim (1)

- CATH Technical Manager
  Institute of Structural and Molecular Biology
  University College London

- CathDB Usage Data Analysis Report, July 2013. Available from The Business Engagement and Impact Manager, School of Business, Economics and Informatics, Birkbeck, University of London

Claim (2)

- Sarah Hunter et al. (2011). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research 2011 40(1),* pp D306-D312*.* DOI: 10.1093/nar/gkr948

Claim (3)

- CATH Project Leader
  School of Life and Medical Sciences
  University College London

- Letter of support from the CEO of Acpharis

- Letter of support from the former Director of the Computational Biology and Chemistry Department of the Merck Research Laboratories in Italy.