



European Bioinformatics Institute
Annual Scientific Report 2009

Annual Scientific Report 2009

European Bioinformatics Institute

EMBL-European Bioinformatics Institute

Wellcome Trust Genome Campus, Hinxton

Cambridge CB10 1SD

United Kingdom

Tel. +44 (0)1223 494444, Fax +44 (0)1223 494468

www.ebi.ac.uk

EMBL Heidelberg

Meyerhofstraße 1

69117 Heidelberg

Germany

Tel. +49 (0)6221 387 0, Fax +49 (0)6221 387 8306

www.embl.org

info@embl.org

EMBL Grenoble

6, rue Jules Horowitz, BP181

38042 Grenoble, Cedex 9

France

Tel. +33 (0)4 76 20 72 69, Fax +33 (0)4 76 20 71 99

EMBL Hamburg

c/o DESY

Notkestraße 85

22603 Hamburg

Germany

Tel. +49 (0)40 89 902 110, Fax +49 (0)40 89 902 149

EMBL Monterotondo

Adriano Buzzati-Traverso Campus

Via Ramarini, 32

00015 Monterotondo (Rome)

Italy

Tel. +39 06900 91402, Fax +39 0690091406

Texts:

EMBL-EBI Group and Team Leaders

Layout, editing and cover design:

Vienna Leigh, EMBL Office of
Information and Public Affairs

Louisa Wright, EMBL-EBI Outreach Programme Project Leader

Contents

SECTION 1: INTRODUCTION	5
Foreword	7
Highlights of 2009	9
SECTION 2: SERVICES IN 2009	15
The Activities of the PANDA Group	17
The European Nucleotide Archive Team	31
Vertebrate Genomics	37
The Ensembl Genomes Team	45
The Proteomics Services Team	51
The InterPro Team	57
Computational Chemical Biology: the ChEMBL Team	61
Cheminformatics and Metabolism	65
Database Research and Development	71
The GO Editorial Office	75
The Microarray Informatics Team	79
The Microarray Software Development Team	85
The Protein Data Bank in Europe (PDBe) Team	89
Developing and Integrating Tools for Biologists	93
Literature Resource Development	97
SECTION 3: RESEARCH IN 2009	103
The Bertone Group: differentiation and development	105
The Enright Group: functional genomics and analysis of small RNA function	111
The Goldman Group: evolutionary tools for sequence analysis	117
The Le Novère Group: computational systems neurobiology	123
The Luscombe Group: genome-scale analysis of regulatory systems	129
The Rebholz-Schuhmann Group: semantic standardisation of the scientific literature	135
The Thornton Group: computational biology of proteins	141
SECTION 4: SUPPORT IN 2009	147
Outreach and Training	149
Industry Support	159
Systems and Networking	163
External Services Team	165

SECTION 5: FACTS AND FIGURES	169
Services and Research	171
Publications	177
Major Database Collaborations	185
Scientific Advisory Boards	187
External Seminar Speakers	189
INDEX	191

Section 1

Introduction

5

Foreword
Highlights of 2009

7
9



Janet Thornton

Director



Graham Cameron

Associate Director



Foreword

7

Welcome to EMBL-EBI's 2009 Annual Scientific Report.

With the emergence of ever-more powerful sequencing and the need for improved translation of biological discoveries into applications, EMBL-EBI's mission, to provide bioinformatics services for biomolecular data, to perform basic bioinformatics research, to provide user training and to support industry, has never been more important.

This year has seen major developments in capturing and storing data from next-generation sequencing machines. These are powering many international projects, including the 1000 Genomes Project to reveal the molecular basis of human variation. With our international colleagues, the EBI is providing the public archive for the data generated. The computational and storage needs of all these projects are tremendous, and require an enormous growth in both the computer power and the storage needed to fulfil these commitments.

This year has also seen ChEMBLdb go live. This is a freely available web resource for chemical biology and drug discovery research. ChEMBLdb provides worldwide access to a large number of medicinal chemistry lead optimisation experiments (usually known as Structure Activity Relationship, or SAR data) reported in the primary literature. The ChEMBLdb resource is highly complementary to existing resources such as UniProt, PDBe and ChEBI, but adds an important new capability to address the needs of the pharmaceutical and biotechnology industries, and the academic chemical biology communities with chemical data.

The EBI's research is diverse and flourishing, producing both exciting discoveries and the development of powerful new tools to handle the flood of data. We have contributed to understanding how genome structure affects its function; to defining the repertoire of transcription factors, which provide regulatory mechanisms underlying biological processes in humans; and to elucidating the effect of deleting a component of the nutrient-responsive mTOR (mammalian target of rapamycin) signalling pathway, which led to increased life span and resistance to age-related pathologies in mice. In parallel we have developed new resources, including leading the worldwide effort to develop a Systems Biology Graphical Notation, which is a new 'language' to describe and visualise all kinds of biological knowledge, from gene regulation to metabolism and cellular signalling.

Our services are being used ever-more frequently by scientists in Europe and worldwide, with around 3 million web hits per day on the EBI website and over one million compute jobs per month run at the EBI. As the data resources develop, our commitment to providing training for users increases. This year we have run many new workshops and courses, both at Hinxton and throughout Europe, and participated in many conferences, in total reaching an estimated 24,000 researchers. Our Industry Programme has also grown, invigorated by the new computational chemistry developments at the EBI and bringing new ideas both for services and for workshops. The Innovative Medicines Initiative (IMI), run by the European pharmaceutical industry and the European Commission, offers a welcome opportunity to address the needs of this sector more directly with explicit grant funding, involving the EBI in several new projects.

The additional funding made available by EMBL Council in this indicative scheme is being used to consolidate the core set of data resources provided by EMBL-EBI, but longer-term funding is still precarious. The ELIXIR preparatory phase, which aims to develop a plan to construct and operate a sustainable infrastructure for biological information in Europe, is beginning to bear fruit. However, although two countries have already committed to this project, there is still much work to be done to define the scope and secure stable funding for this infrastructure for the future.

All our efforts at the EBI rely on extensive interactions with colleagues in Europe and throughout the world. The deposition of new data, the daily exchange of information between data resources, the joint development of software tools, the sharing of curation tasks and the challenges of collaborative research have built an extensive community of collaborators. It remains our privilege and pleasure to work with them and together we will continue to aim for excellence in all that we do.

*Janet Thornton, Director
Graham Cameron, Associate Director*

Janet Thornton, Director

*PhD 1973, King's College & National Inst. For Medical Research, London.
Postdoctoral research at the University of Oxford, NIMR & Birkbeck College, London.
Lecturer, Birkbeck College 1983–1989.
Professor of Biomolecular Structure, University College London since 1990.
Bernal Professor at Birkbeck College, 1996–2002.
Director of the Centre for Structural Biology at Birkbeck College and University College London, 1998–2001.
Director of EMBL-EBI since 2001.*

Graham Cameron, Associate Director

*Applications Programmer, EMBL Data Library, 1983–1983.
Database Administrator, EMBL Data Library, 1983–1986.
Manager, EMBL Data Library, 1986–1992.
Project Leader overseeing the creation of EMBL-EBI Outstation, 1993–1994.
Head of Services, EMBL-EBI, 1994–1998.
Joint Head of EMBL-EBI, 1998–2001.
Associate Director of EMBL-EBI since 2001.*

Janet Thornton*Director***Graham Cameron***Associate Director*

Highlights of 2009

SERVICES

Enabling optimal exploitation of biomolecular information is at the heart of EMBL-EBI's mission. The EBI services include the provision of biological databases and tools to explore them. Our constituency includes academic and commercial researchers throughout Europe and the world, and we form a European node in many global data-sharing collaborations (figure 1). These service activities are accompanied by extensive outreach and training (concentrated mostly in Europe) with a dedicated team (see page 149), and industry users receive targeted support through the EBI's Industry Programme (page 159).

2009 has been an interesting year for these services. The familiar ever-increasing data flow rates perpetuate exponential database growth curves, which are already etched on the retinas of service providers and funders. However, in 2009 striking qualitative changes have accompanied the familiar quantitative changes. This has resulted in new data resources, restructuring of existing resources and the demise of some obsolete resources.

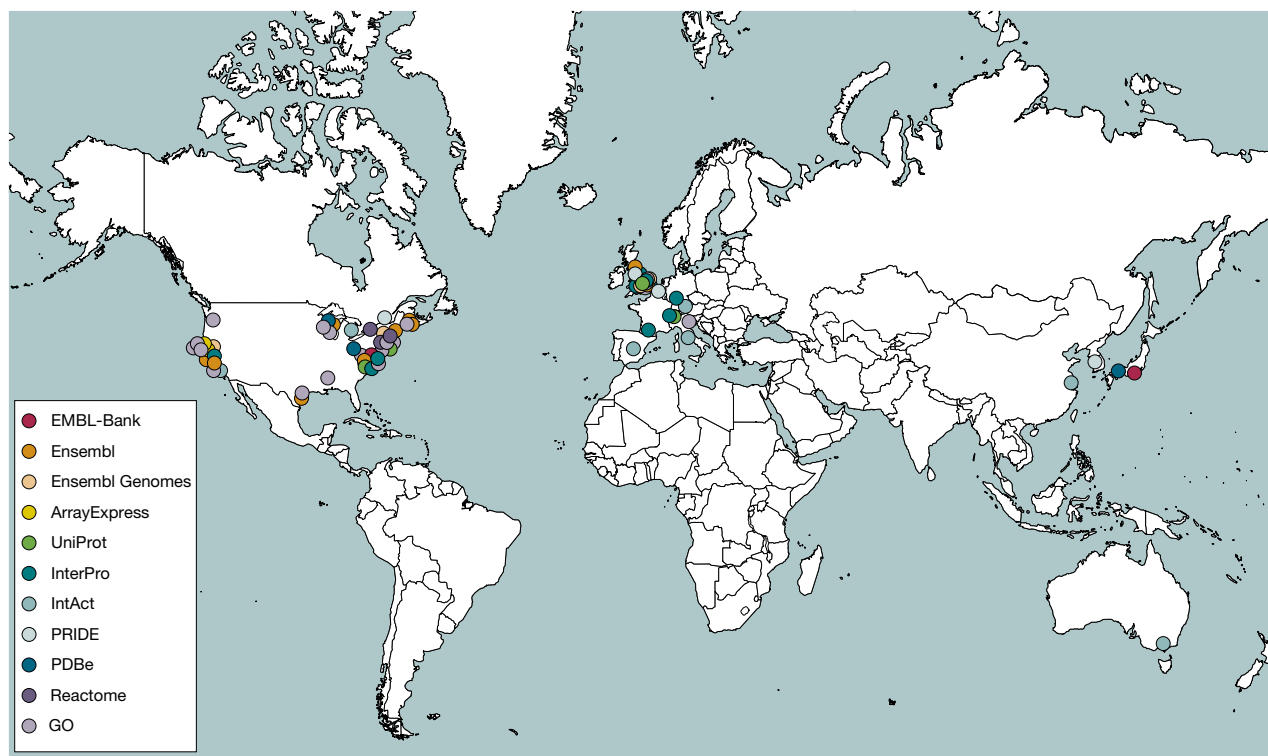


Figure 1. Map showing location of EMBL-EBI's collaborations for the major databases.

One of the most striking new resources is ChEMBLdb, a database of bioactive compounds (drugs and drug-like molecules) and their quantitative properties and bioactivities. Computer methods for the storage and exploitation of chemical information – now called cheminformatics – pre-date the existence of bioinformatics, but hitherto chemical data have primarily resided in proprietary systems and used closed/proprietary formats. With the aid of funding from the Wellcome Trust, the EBI has acquired the datasets of Inpharmatica/Galapagos, which provide the foundation for crucial new public chemical resources, of which ChEMBLdb is the first. This effort, led by a new team leader, John Overington, will have a profound enabling effect on biological research.

Rationalisation of our services has continued in 2009. The structure of our core databases reflects the central dogma on genes, transcripts, proteins, structures of biomacromolecules and the choreography of their behaviour and interactions with each other and with small molecules. Having already integrated our protein and nucleotide service activities, we have further restructured our DNA activities and services in 2009.

We now distinguish between the Ensembl family of databases – which provide genomic data organised and annotated by the in-house Ensembl pipeline and the European Nucleotide Archive (ENA) – the collection that stores, organises and makes available data submitted by the scientific community, along with submitter annotation where available.

The Ensembl databases include:

- Ensembl, a joint project with the Wellcome Trust Sanger Institute presenting vertebrate and other eukaryotic data;
- Ensembl Genomes, a new database launched in 2009, which exploits the framework of Ensembl and works with organism-specific communities to curate genomes from metazoa, protists, bacteria, plants and fungi.

The European Nucleotide Archive includes:

- the long-established EMBL-Bank Nucleotide Sequence Database;
- the Trace Archive for raw capillary sequencing data;
- the Sequence Read Archive for unassembled data from next-generation sequencing.

As a consequence of this rationalisation and other developments, we will soon discontinue four databases:

- **Integr8** – a portal for species with completely deciphered genomes, which will be made redundant by developments in the Ensembl, Ensembl Genomes and UniProt interfaces.
- **Genome Reviews** – related to the Integr8 database and which provides standardised annotation of non-vertebrate genomes. It too will be made redundant by developments in the Ensembl, Ensembl Genomes and UniProt interfaces.
- **ASTD** – documents Alternative Splicing and Transcript Diversity, information which will in future be provided by Ensembl.
- **IPI** – the International Protein Index has provided a top-level guide to the main databases that describe the proteomes of higher eukaryotic organisms. It is rendered redundant by the adoption of standardised proteome sets in Ensembl and UniProt.

New challenges are nowhere more apparent than in the human genome data. The study of genetic variation and its phenotypic correlates has spawned the 1000 Genomes Project, which threatens to drown us in data; and the European Genome-phenome Archive (EGA), which will connect genetic and phenotypic information of individuals and thus require infallible systems to ensure only authorised access.

These studies of human variation are one exploitation of today's ability to repeat biomolecular assays across different samples and conditions, or carry out multiple assays on the same sample or samples. This is revolutionising biomolecular science. For instance, the same individual can be genotyped and profiled for gene expression and the results compared across individuals. Hitherto sample information has been labelled 'metadata' with an implication that it is not at the heart of the molecular science. However, these data play an increasing role in research, and we must capture them systematically and securely.

We loosely use the term sample data to refer to a range of such metadata. They are recorded in many databases, including ArrayExpress, EGA, PRIDE, Ensembl, and the ENA and can potentially be used to link different datasets from the same sample. In 2009 we have initiated a project that will combine and exploit work from a number of EBI developments to create a single sample database to support existing databases and to provide a sample-based entry point for database searches.

Other significant developments in the year include:

- the passing of the leadership of the PDBe database from Kim Henrick to Gerard Kleywegt. This transition has been smooth, and Gerard is already well embedded in the wwPDB collaboration. His previous experience has enabled

him to come up to speed very fast, and there has certainly been no pause in the commitment to good services layered on high quality data;

- the appointment of Johanna McEntyre to head our Literature Services team. She joins us from the National Center for Biotechnology Information (NCBI), our collaborators in the USA, and also has extensive experience in publishing, so she too has been able to hit the ground running. Building on work from Peter Stoeck and Dietrich Rebholz-Schuhmann, her immediate priorities are to consolidate and advance our literature services by expanding content, improving searching, enhancing citation-based links within literature and to the other EBI databases as well as leveraging Dietrich Rebholz-Schuhmann's text-mining work, and strengthening the UK PubMed Central project;
- a new database, Gene Expression Atlas, has been launched as a part of the ArrayExpress infrastructure. The Gene Expression Atlas allows users to query gene expression under a range of biological conditions, including different cell types, developmental stages, physiological states, phenotypes and disease states. While the ArrayExpress Archive makes data from high-throughput functional genomics assays available to microarray-data experts, Gene Expression Atlas presents this information in a format accessible to any biologist.

These reorganisations all contribute to the integration of our information, a long-standing priority, which is rendered ever-more urgent by the diversity of -omics data available, and the holistic emphasis of systems biology. During the year we have initiated a new project that will exploit the end-to-end comprehensive data at the EBI to create a searching and browsing portal providing gene-centric summary information across our databases. This will become available in 2010.

Aside from the integration of resources produced at the EBI, we have also made strides in exploiting web services technology Europe-wide, through the European Commission-funded EMBRACE network. This already creates interoperability between over 800 tools throughout Europe, including those from its sister Commission-funded network BioSapiens.

Changing information technology increases user expectations and continues to bring challenges and opportunities. Ever-improving household computer tools make people impatient with impenetrable scientific user interfaces. Advances in distributed informatics (such as grid and cloud computing, service-oriented architectures, and virtualisation) challenge traditional methodologies. For example, running programs locally is being replaced to some extent by making algorithms available via web services, which are run by remote users on data centres' machines. We have carried out tests of cloud computing, including running cloud software at the EBI and also loading the Ensembl data into the Amazon Cloud. While the time is not ripe for wholesale adoption of this methodology, we see it as immensely important that our IT development strategy retains the agility necessary to exploit new trends. For these, and other reasons, a new, UK-funded EBI data centre will outsource the provision of the physical facilities (though not, at this stage, the IT hardware).

All developments of the EBI services sit in the context of the ELIXIR project whose mission is to construct and operate a sustainable infrastructure for biological information in Europe, to support life science research and its translation to medicine and the environment, the bio-industries and society.

Indeed, ELIXIR is essential to the future of the European bioinformatics infrastructure. No matter how clever our engineering, we are faced with an inescapable truth: the scale of the bioinformatics task is outstripping the rate of information technology advance. The much-cited Moore's Law anticipates a doubling of 'bang for the buck' in the hardware market every eighteen months. However, our rising demand is not going to be satisfied by IT advances

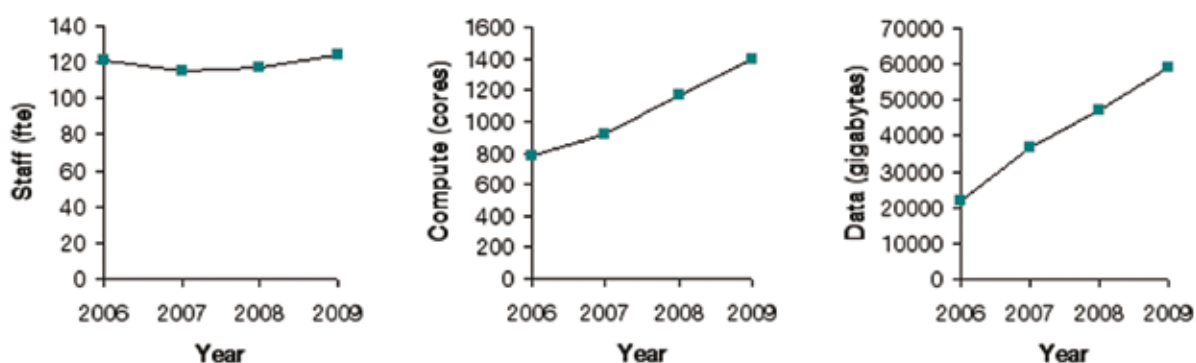


Figure 2. Combined staff, compute and storage trends over four years for EMBL-Bank, Ensembl, ArrayExpress, UniProt, InterPro and PDBe.

alone. We must use our ingenuity to engineer solutions that solve our problem within reasonable costs. Among other things, this will involve judicious sharing of the workload within ELIXIR. For this to succeed, it must be done through modularisation along natural lines of cleavage, rather than unnaturally dividing single tasks. The changing costs of six of our core databases over the last four years (figure 2) renders this modularisation strategy plausible. On one hand, we see that the staff costs can be contained for such projects, implying that growth of staff necessary for the combined infrastructure of the EBI and ELIXIR will be required for new information resources, rather than expansion of existing projects. On the other hand, we do see an inexorable rise in IT costs, both compute and storage. This will inevitably require further investment and the UK support for a new data centre is a welcome contribution. It is anticipated that ELIXIR nodes other than the EBI will create and operate some data resources in future.

As we move forward into the coming year, and develop EMBL's ideas for the next Scientific Programme (2012–2016), we can review how far we have come in recent years. The computer infrastructure has changed unrecognisably in scale, and we have received support both from the EMBL Member States' budget and UK sources to enable a robust architecture. This gives us confidence that we can cope with the coming needs of bioinformatics, and indeed continue to improve. The usage of our services continues to rise, which implies that we are continuing to serve the needs of an expanding user community. Evolving science challenges us, quantitatively and qualitatively, but, with the scientific context of EMBL's research and the local relationships with the Wellcome Trust Sanger Institute and Cambridge University, we believe that we can rise to those challenges and continue to serve science well.

RESEARCH

The eight research groups at EMBL-EBI perform computational research into many different biological questions, ranging from genome evolution and transcriptional regulation to systems modelling of basic biological processes and disease. Bioinformatics continues to diversify, often led by the development of new technologies which generate the need for new methods for data handling and interpretation. The EBI research groups are usually quite small, with two or three students and externally-funded postdocs. Their research complements the broad remit of EBI's service provision, benefiting from the in-house technical expertise provided by the larger service teams, and in turn helping to identify current challenges for researchers using our data resources. Several service teams also incorporate a small research and development component.

At EMBL-EBI our research includes groups focused in the following areas:

- genome function analysis, including chromatin, epigenetics, microRNAs and variation;
- evolutionary processes;
- genome-scale analysis of regulatory systems, including transcription;
- cheminformatics analysis of small molecules and their interactions, especially for design of new therapeutics;
- 3D structural basis of biological processes and disease;
- computational neurobiology;
- differentiation and development;
- exploiting the literature to extract new knowledge, through text mining.

Many exciting new discoveries and tools have been developed this year and below are a few highlights.

The impact of nucleosome positions on the evolution of genomic DNA (Washietl *et al.*, 2008) – Is the molecular evolution of DNA affected by its packaging in the nucleus? The Goldman group, in collaboration with Rainer Machné (Institute for Theoretical Chemistry, Vienna) analysed substitution rates in the yeast genome dependent on the location within maps of nucleosome positions. A statistically significant dependency could be observed; substitution rates are on average 10–15% lower in free linker regions between nucleosomes than in the DNA wrapped around nucleosomes. This striking difference is found in intergenic regions as well as in coding regions. The cause for this is not yet clear, but could reflect the fact that most known DNA repair mechanisms are impaired by nucleosomes and are more efficient in naked DNA. The observation can be interpreted as a phylogenetic 'footprint' of nucleosome positions that demonstrates for the first time that nucleosome organisation is a feature of the yeast genome conserved over evolutionary timescales.

The human transcriptional regulatory system (Vaquerizas *et al.*, 2009) – transcription factors are key cellular components that control gene expression; their activity determines how cells function and respond to cellular environments. Despite intense scientific interest in transcriptional regulation, there was no reliable repertoire of transcription factors encoded in the human genome. In collaboration with Sarah Teichmann (MRC Laboratory of Molecular Biology), Nick Luscombe's group identified and analysed 1,391 sequence-specific DNA binding factors for the human genome, their function, genomic organisation and evolutionary conservation. Their expression profiles across numerous healthy organs and tissues were assessed, allowing the definition of global and specific regulators. In addition, a comparison

of 24 eukaryotic organisms revealed how distinct classes of regulators appeared at crucial points during the evolution of the human lineage. The dataset and analyses offer useful starting points for further investigations of the regulatory mechanisms underlying biological processes in humans.

Transcriptional regulation of the fission yeast cell cycle (Aligianni *et al.*, 2009) – gene subsets are induced at specific times during the cell cycle when their function is required; mistakes in this regulation can lead to cancer. A well-known regulatory complex (MBF) is responsible for controlling the onset of DNA replication in humans and some other organisms. In a collaborative work with Jürg Bähler's group at University College London (UCL), Alvis Brazma's group has shown that in fission yeast this complex also induces *Yox1p*, the protein product of which binds to MBF and represses MBF-regulated genes. These data uncover a new negative control loop, highlighting the sophistication of cell cycle regulation and illustrating regulatory similarities and differences between organisms.

Functional genomics of ageing (Selman *et al.*, 2009) – identification of a new longevity-associated gene in mouse. In collaboration with the Functional Genomics of Ageing Consortium at UCL, the Thornton group has analysed various functional genomics datasets to understand more about the molecular basis of ageing, involving a variety of different experiments on flies, worms and mice. To understand how caloric restriction affects longevity, they showed that in mice deletion of ribosomal S6 protein kinase 1 (S6K1), a component of the nutrient-responsive mTOR (mammalian target of rapamycin) signalling pathway, led to increased life span and resistance to age-related pathologies such as bone, immune, and motor dysfunction and loss of insulin sensitivity. Deletion of S6K1 induced gene expression patterns similar to those seen in calorie restriction or with pharmacological activation of adenosine monophosphate-activated protein kinase (AMPK), a conserved regulator of the metabolic response to calorie restriction. These results demonstrate that S6K1 influences healthy mammalian life span and suggest that therapeutic manipulation of S6K1 and AMPK might mimic calorie restriction and could provide broad protection against diseases of ageing.

Resource developments in research groups

As part of their research, several groups have developed specialist databases or tools that are made available through the EBI website. These include:

The Systems Biology Graphical Notation (Le Novère *et al.*, 2009) – a community of biochemists, modellers and computer scientists led by Nicholas Le Novère has developed the Systems Biology Graphical Notation (SBGN). SBGN consists of three complementary languages: process descriptions, entity relationships and activity flows. Together they enable scientists to represent networks of biochemical interactions in a standard, unambiguous way. SBGN will foster efficient and accurate representation, visualisation, storage, exchange and reuse of information on all kinds of biological knowledge, from gene regulation to metabolism and cellular signalling.

A comprehensive analysis toolkit, visualisation platform and reference set of genome-wide human DNA methylation profiles (Rakyan *et al.*, 2008) – DNA methylation is a tissue-specific epigenetic modification required for genome function and a key regulator of gene expression. Comprehensive maps of genome-wide reference DNA methylation profiles and the identification of tissue-specific differentially methylated regions (tDMRs) are critical to the understanding of the roles of DNA methylation in cellular identity and normal function. In collaboration with others, Paul Flicek's group created a reference set of DNA methylation profiles in 16 human tissues including an open source analysis and visualisation infrastructure that permits the comparison of our reference profiles with newly generated data.

The text-mining platform Whatizit (www.ebi.ac.uk/webservices/whatizit/; Rebholz-Schuhmann *et al.*, 2008) – a powerful tool to extract information from the literature. It enables integration of literature with bioinformatics data resources. A large research community uses Whatizit solutions through CiteXplore, EBIMed, UKPMC and PubMed (from Nature Publishing Group).

Details of all research progress are given in the individual reports.

ELIXIR

In October 2006, the European Strategy Forum on Research Infrastructures (ESFRI), a body set up at the initiative of the European Council by 33 EU Member States and associated countries, published its first roadmap. This presented 35 pan-European research infrastructure projects, identified by ESFRI as being of key importance for the development of science and innovation in Europe. One of the projects included on the roadmap was an upgrade of European bioinformatics infrastructure. The European Commission then made funds available through its Seventh Framework Programme for 'preparatory phase projects' to pave the way towards the construction of the ESFRI infrastructures. The EBI coordinated an application for ELIXIR (European Life Sciences Infrastructure for Biological Information; www.elixir-europe.org) – a preparatory phase project to create a sustainable infrastructure for biological information in Europe. The project, which runs from 2008 to 2011, was funded via a €4.5 million grant and involves a consortium of 32 partner organisations from 13 member states.

ELIXIR will contribute to European science by: optimising access to and exploitation of life-science data; ensuring longevity of the data and protecting investments already made in research that collected the data; increasing the com-

petence and size of the already-large user community by strengthening national efforts in training and outreach; and enhancing the global success and influence of Europe in life science research and industry. Over the first 18 months, ELIXIR has engaged with its numerous stakeholder groups to define ELIXIR's scope and remit. These are summarised in a series of reports available at www.elixir-europe.org.

ELIXIR's structure will be based on a hub and nodes model. The hub will be located at the EMBL-EBI and its purpose will be to: provide central coordination of ELIXIR's activities; provide core data resources; host registries of bioinformatics tools; host the main data centre; ensure backup of core data resources; and coordinate user training and information dissemination. The initial governance structure for ELIXIR is likely to be a 'special', ring-fenced project of the European Molecular Biology Laboratory (EMBL), which is an intergovernmental organisation supported by 20 member states. The ELIXIR nodes will be located throughout Europe. Each node will provide an internationally competitive capability at the European level. A node could provide data resources, computational and storage resources or any of the other capabilities that are envisaged as being provided by ELIXIR – on a pan-European level.

In the absence of central European funding, it is likely that ELIXIR will be supported through several funding streams: at the international level by a consortium of ELIXIR member states and possibly the European Commission, and at the national level by, for example, national governments, charities and other funding bodies. ELIXIR has begun this process by securing funds from the Swedish Government (€1.7M) and the UK Government (€12M). The UK funds will be used to build a new compute infrastructure for Europe's core biomedical data resources, with the capacity to deal with the enormous scale-up in storage now necessary to cope with the output of next-generation sequencing projects.

ELIXIR holds a special place among the ESFRI biomedical science infrastructures because it will provide an integrative informatics infrastructure for all of them. ELIXIR will interact with the other research infrastructures in many different ways, for example: by accepting large datasets and making them available to the community; by linking ELIXIR's core data resources to large datasets that are out of scope for ELIXIR; by providing access to large numbers of observations on the behaviour of biological entities under different conditions, meticulously cross-linked to the data collections for which ELIXIR is responsible, and by supporting the development of standards and ontologies. EMBL-EBI is already a partner in seven out of ten ESFRI biomedical science research infrastructures and is also collaborating with the environmental science infrastructure Lifewatch. ELIXIR has by far the largest user base of any of the research infrastructures, with several million users across all of Europe.

BioSapiens

2009 saw the successful completion of the EU-funded BioSapiens Network of Excellence (www.biosapiens.info). BioSapiens supported a large-scale, concerted effort to annotate genome data by laboratories distributed around Europe, using both informatics tools and input from experimentalists. It had 25 partners throughout Europe, and was coordinated at the EMBL-EBI. BioSapiens provided servers and clients for annotating genomes with a range of different types of biological data, and a large number of gene and protein annotations. Through the 'European School of Bioinformatics', it trained 349 young scientists from over 20 countries. Over 400 publications, with over 5,000 citations, are directly associated with BioSapiens. Perhaps most importantly, BioSapiens contributed significantly towards building a European Research Area for bioinformatics by: contributing to the organisation of major European and international bioinformatics conferences and organising public meetings; encouraging cooperation between laboratories spread around Europe, and being a major influence in the development of ELIXIR.

Referenced publications

Le Novère N. *et al.*, (2009). The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27, 735-741

Rakyan, V.K. *et al.* (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, 18, 1518-29

Rebholz-Schuhmann, D. *et al.*, (2008). Text processing through web services: calling Whatizit. *Bioinformatics*, 24, 296-298

Aligianni, S. *et al.*, (2009). The fission yeast homeodomain protein Yox1p binds to MBF and confines MBF-dependent cell-cycle transcription to G1-S via negative feedback. *PLoS Genet.*, 5, 1-12

Selman, C. *et al.*, (2009). Ribosomal protein S6 kinase 1 signaling regulates mammalian life span. *Science*, 326, 140-144

Vaquerizas J.M. *et al.*, (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, 10, 252-263

Washietl, S. *et al.*, (2008). Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.*, 24, 583-587

Section 2

Services in 2009

15

The Activities of the PANDA Group	17
The European Nucleotide Archive Team	31
Vertebrate Genomics	37
The Ensembl Genomes Team	45
The Proteomics Services Team	51
The InterPro Team	57
Computational Chemical Biology: the ChEMBL Team	61
Chemoinformatics and Metabolism	65
Database Research and Development	71
The GO Editorial Office	75
The Microarray Informatics Team	79
The Microarray Software Development Team	85
The Protein Data Bank in Europe (PDBe) Team	89
Developing and Integrating Tools for Biologists	93
Literature Resource Development	97

Rolf Apweiler
*PhD 1994, University of
Heidelberg.
At EMBL since 1987;
At EMBL-EBI since 1994.*



Ewan Birney
*PhD 2000, Sanger
Institute.
At EMBL-EBI
since 2000.*



The Activities of the PANDA Group

17

INTRODUCTION

The PANDA (Protein and Nucleotide Data) group was created in June 2007 by merging the former Ensembl (Birney) and Sequence Database (Apweiler) groups.

The activities of the PANDA group are focused on the production of protein sequence, protein family and nucleotide sequence databases at EMBL-EBI. We maintain and host the European Nucleotide Archive, the Ensembl and Ensembl Genomes resources, the UniProt protein resource, the InterPro domain resource and a range of other biomolecular databases. These efforts can be divided into three major groups: nucleotides, proteins, and chemoinformatics and metabolism. In addition to PANDA activities, both the Apweiler and Birney groups have complementary research components (presented on pages 25 and 26 respectively). Substantial training and outreach efforts are also part of the PANDA group's activities (see page 23).

Various external service aspects of the PANDA group's activities are described in the report by Rodrigo Lopez, team leader of the EBI External Services on page 165. The activities of the European Nucleotide Archive, Vertebrate Genomics, Ensembl Genomes, Proteomics Services, InterPro, and the Chemoinformatics and Metabolism groups are described in separate reports by their team leaders Guy Cochrane, Paul Flicek, Paul Kersey, Henning Hermjakob, Sarah Hunter and Christoph Steinbeck, respectively.

The main achievements of the PANDA group in 2009 have been:

- handling an ever-growing amount of nucleotide and protein data;
- launching Ensembl Genomes with the expansion of the Ensembl concept across all high-value genomes;
- managing the data flow for the 1000 Genomes Project of human variation;
- retrofitting the majority of UniProt annotation with evidence tags.

PANDA NUCLEOTIDES

The PANDA Nucleotides activities are overseen by Ewan Birney.

PANDA NUCLEOTIDES STRATEGY

DNA sequence remains at the heart of molecular biology and hence bioinformatics and its use has grown significantly with the recent advent of ultra-high throughput DNA sequencing machines. In 2009 we have seen a striking growth in four areas – first, the use of these new machines for surveying natural variation in populations, in particular the human population (see the report from Paul Flicek, page 37); and second, the more routine determination of genotypes from large disease cohorts, leading to association between genetics and disease (also presented in Paul Flicek's report). Third, we launched Ensembl Genomes, a high-quality, community-led genomic information resource for non-vertebrate species (presented in Paul Kersey's report, page 45). Finally, we launched the European component of the Short Read Archive (SRA). The shift in technology and the repositioning of genomic information as a key organisational principal has meant that there have been significant changes to the way our DNA archival services operate and more focus on coordinating with genomic resources, as described in this report.

The PANDA Nucleotides group has three main sections with Ewan Birney providing strategic oversight across all branches. The sections are: the European Nucleotide Archive (under the leadership of Guy Cochrane, see page 31); Vertebrate Genomics, which includes Ensembl (under the leadership of Paul Flicek, see page 37); and Ensembl Genomes (under the leadership of Paul Kersey, see page 45). In addition, the HUGO Gene Nomenclature Committee (HGNC), a smaller group coordinated by Elspeth Bruford, is part of PANDA Nucleotides and presented below. The key organising principal across all these groups is to best coordinate resources for each genome sequence of a species. In contrast, a key difference between the groups is the provenance of the data. In the case of the European Nucleotide Archive, the data content is determined by the submitter, and any added value information is provided as additional resources on this submitted set. This data can be redundant and conflicting, but represents the foundational DNA dataset on which all genomic and nearly all protein sequence is based. This dataset is coordinated worldwide by virtue of the International Nucleotide Sequence Database Collaboration (INSDC), forming a single coordinated set of information with partner groups at NCBI and DDBJ. In contrast, for Ensembl and Ensembl Genomes, these resources are community led and we aim to present a single, non-redundant view of a species' DNA information organised around its genomic sequence. In this case, decisions are made, via interactions with the community, on the best representation of information to provide most utility to users. In the human genome, an important component of this community is the unambiguous assignment of gene symbols, allowing researchers to use memorable names for genes in scientific communication. This is provided by the HGNC group.

HUGO GENE NOMENCLATURE COMMITTEE

Elspeth Bruford, Susan Gordon, Michael Lush, Ruth Seal, Matthew Wright

The HUGO Gene Nomenclature Committee (HGNC) is the only worldwide authority that assigns standardised human gene nomenclature, and remains an essential component of human gene and genome management. The HGNC has two overriding goals: providing a unique name and symbol (abbreviation of the name) for every human gene, and ensuring this information is freely available, widely disseminated and universally used. Achieving these goals involves three key components:

- bioinformatic analysis of nucleotide and protein sequences;
- curation of online resources, particularly a database comprising individual gene records containing the gene name, symbol and relevant information (cDNA sequence, chromosomal location, key publications, links to other databases, etc.);
- constant communication including consultation with researchers, coordinated naming of orthologous genes with nomenclature groups in other species, exchanging data with numerous databases, and raising awareness of the resource within the scientific community, both electronically and through publications and attendance at conferences and meetings.

This year gene naming has continued to focus on the increasing number of genes identified by the consensus coding sequence (CCDS) project, with less than 130 of the current total of 18,177 genes in the CCDS set not yet having an HGNC approved gene symbol. The total of approved gene symbols presently stands at 28,481 (as of 22 October 2009), a sizeable increase of over 2,300 in the past year. A large part of this is due to the increased assignment of names for non-protein-coding RNA genes and the systematic naming of ribosomal protein pseudogenes based on a genome-wide analysis from Mark Gerstein's group at Yale.

In the past year HGNC staff have attended six international conferences, including the 14th annual meeting of the RNA Society held in Madison, Wisconsin; ISMB in Stockholm; the Third International Biocurator Conference in Berlin; and the American Society of Human Genetics meeting in Hawaii where we shared a booth with the Human Genome Organisation (HUGO). Elspeth Bruford and Matt Wright also organised a Gene Nomenclature Across Species meeting in Cambridge (12–13 October) which brought together invited representatives from gene nomenclature and annotation groups, vertebrate genomes, genome databases and orthology resources, and gene family experts. The aim of the meeting was to discuss consistent gene naming across vertebrate species, and following very productive discussions during the meeting, we are hopeful that we can produce some common guidelines and proposals for unifying vertebrate gene naming.

Further developments this year have included:

- updating the HGNC Comparison of Orthology Predictions search tool, HCOP, to enable users to compare orthologues predicted for human genes in a further four genomes (cow, *C. elegans*, *S. cerevisiae* and duck-billed platypus, in addition to chimp, mouse, rat, dog, chicken, zebrafish and fruit fly) and including new orthology assertions from the OPTIC (orthologous and paralogous transcripts in clades) database produced by Chris Ponting's group in Oxford;
- developing an online HGNC forum, which enables us to host nomenclature discussions between many different researchers;

- inclusion of a statistics page notifying users of the numbers of entries in our database grouped by locus type. This also includes links for quick and easy downloading of each of these sets of data;
- addition of a list search facility which allows multiple gene symbols to be searched for in one step;
- a new ncRNA web page that presents our work to date on naming non-protein-coding RNA genes;
- HGNC authorship on two gene family publications and on a paper describing the IUPHAR database of G protein-coupled receptors and ion channels.

PANDA PROTEINS

The PANDA Proteins activities are overseen by Rolf Apweiler.

PANDA PROTEINS STRATEGY

The activities of the PANDA proteins teams are centred on the mission of providing public access to all known protein sequences and functional information about these proteins. The UniProt resource provides the centrepiece for these activities. Most of the UniProt sequence data is derived from translation of nucleotide sequences provided by the European Nucleotide Archive and Ensembl. All UniProt data undergoes classification provided by InterPro (see the report from Sarah Hunter, page 57). In addition, we add information extracted from the scientific literature and curator-evaluated computational analysis whenever possible. The combined InterPro literature annotation forms the basis for automatic annotation approaches to annotate all the sequence data without experimental functional data. Protein interaction and identification data is or will be provided to UniProt by the IntAct protein-protein interaction database and by the Protein Identification (PRIDE) database. The progress of these resources is presented in Henning Hermjakob's report, page 51).

The rest of this section details the three areas (UniProt, Gene Ontology Annotation and RESID) which directly report to Rolf Apweiler whereas the two other areas (InterPro and Proteomics Services) are described by their own team leaders.

THE UNIVERSAL PROTEIN RESOURCE

Yasmin Alam-Faruque, Ricardo Antunes, Daniel Barrell, Benoit Bely, Mark Bingley, Paul Browne, Wei Mun Chan, Emily Dimmer, Ruth Eberhardt, Alexander Fedotov, Rebecca Foulger, John S. Garavelli, Renato Golin, Rachael Huntley, Julius Jacobsen, Michael Kleen, Kati Laiho, Duncan Legge, Quan Lin, Wudong Liu, Jie Luo, Michele Magrane, Maria-Jesus Martin, Claire O'Donovan, Sandra Orchard, Samuel Patient, Diego Poggioli, Tony Sawford, Eleanor Stanley

The Universal Protein Resource (UniProt; www.ebi.uniprot.org) is a collaboration of EMBL-EBI, the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource group at Georgetown University Medical Center (The UniProt Consortium, 2009). Its purpose is to provide the scientific community with a single, centralised, authoritative resource for protein sequences and functional information.

The primary mission of the consortium is to support biological research by maintaining a freely accessible, high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross references and querying interfaces. In addition, UniProt also provides several non-redundant sequence databases suitable for efficient searching and a comprehensive collection of all publicly available protein sequences.

The UniProt databases consist of four database layers optimised for different purposes:

- UniProt Knowledgebase (UniProtKB) provides the central database of protein sequences with accurate, consistent, and rich sequence and functional annotation;
- UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for the newly expanding area of metagenomic and environmental data;
- **UniProt Archive** (UniParc) provides a stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data;
- **UniProt Reference Clusters** (UniRef) provide non-redundant data collections based on the UniProt Knowledgebase and UniParc in order to obtain complete coverage of sequence space at several resolutions.

The UniProt Knowledgebase

The UniProt Knowledgebase consists of two parts: 1) UniProtKB/Swiss-Prot, which contains manually annotated records from literature-derived information and curator-evaluated computational analysis, and 2) UniProtKB/TrEMBL, which contains computationally analysed records enriched with automatic annotation and classification. The UniProt Knowledgebase release 15.9 (October 2009) consists of 10,011,983 entries (510,076 UniProtKB/Swiss-

Prot entries and 9,501,907 UniProtKB/TrEMBL entries). A specimen UniProt report can be found at www.uniprot.org/uniprot/P57727.

The main principles of the UniProtKB are high-quality annotation, integration with external databases, minimal redundancy and evidence-tagged annotation where experimental information is available.

Annotation: in addition to capturing the core data mandatory to each UniProtKB entry (consisting principally of the amino acid sequence, the protein name or description, taxonomic data and citation information), we strive to attach as much annotation information as possible to the protein. This is achieved in two ways: manually and automatically.

Manual annotation by curators is based on literature and sequence analysis. Sequences for which novel functional, structural, and/or biochemical data have been published are assigned high manual annotation priority. In UniProtKB, annotation consists of the description of the following items:

- function(s) of the protein;
- enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms);
- biologically relevant domains and sites;
- post-translational modifications (PTMs);
- molecular weight determined by mass spectrometry;
- subcellular location(s) of the protein;
- tissue-specific expression of the protein;
- developmental-specific expression of the protein;
- secondary structure;
- quaternary structure;
- interactions;
- splice isoform(s);
- mature protein products;
- polymorphism(s);
- similarities to other proteins;
- use of the protein in a biotechnological process;
- diseases associated with deficiencies or abnormalities of the protein;
- use of the protein as a pharmaceutical drug;
- sequence conflicts, etc.

The annotation is found in the comment (CC), feature table (FT) and keyword lines (KW). Comments are classified according to topics to allow easy retrieval of specific categories of data from the database. The UniProt curators also contribute to the work of the Gene Ontology Annotation (GOA) project (Barrell *et al.*, 2009) by assigning GO terms to information extracted during the annotation process, i.e. the function of a protein, what processes it is involved in and cellular localisation (see project description on page 22). To maintain the most accurate and complete protein data, information is not only obtained from publications reporting new sequence data, but also from review articles to facilitate the periodic revision of protein families or groups of proteins. Furthermore, we have enlisted external experts to send us comments and updates concerning specific groups of proteins.

Automatic annotation of UniProtKB/Swiss-Prot depends upon strict quality control and the work of an experienced team of biologists in keeping the data as consistent, complete and up-to-date as possible. Consequently, this information is of such high quality that it can serve as very clean input for data mining routines for automatic annotation. Several tools exist for automated annotation, some of which might provide more reliable data, while others might produce a larger quantity. The implementation and unification of many of these approaches in the production of UniProtKB/TrEMBL annotation and the provision of evidence tagging gives the scientific community a much more useful source of information. One system, RuleBase, uses a semi-automatic approach, while Spearmin is completely automated and is based on decision trees (Kretschmann, Fleischmann & Apweiler, 2001). Both systems use UniProtKB/Swiss-Prot as the source to generate the annotation rules, which are then stored and managed in RuleBase or Spearmin. InterPro (Hunter *et al.*, 2009) is used to recognise domains and to classify all UniProtKB entries into families and superfamilies. The annotation shared by the functionally characterised UniProtKB/Swiss-Prot proteins of a particular group is then extracted and assigned to the non-annotated UniProtKB/TrEMBL entries of the same group. These systems have been

used to improve the annotation of 30% of UniProtKB/TrEMBL entries. In the last year, extensive work has been done to integrate RuleBase, PIRSF (Natale *et al.*, 2004) and HAMAP (Gattiker *et al.*, 2003) rule systems into a prototype extension to our existing production pipeline to further improve the annotation of UniProtKB/TrEMBL entries. This is expected to be in full operation in the next reporting period and have a significant impact on UniProtKB/TrEMBL coverage. There is now a central UniProt repository of rules and we have begun development on a central UniProt curation tool to allow curators at EMBL-EBI, SIB and PIR to create and update rules in the same system and to have this directly linked to the UniProtKB/TrEMBL production pipeline at the EBI. These are currently part of the curation pipeline for UniProtKB/Swiss-Prot but will switch to only being applied to UniProtKB/TrEMBL in the near future, facilitating a further increase in coverage and curator resources. The provision of evidence tags in UniProtKB XML has also allowed the visualisation of the rules for the predicted automatic annotation present in UniProtKB/TrEMBL in the new unified www.uniprot.org and this has been extended to existing UniProtKB/Swiss-Prot entries for HAMAP.

Automatic classification and annotation, whereby faster and more effective means of large-scale protein sequence characterisation are generated with limited human interaction, offers a mechanism to handle large data volumes. This is an essential component of UniProt activities going forward.

Integration with other databases: UniProtKB provides cross references to external data collections such as the underlying DNA sequence entries in the DDBJ/EMBL-Bank/GenBank nucleotide sequence databases, 2D PAGE and 3D protein structure databases, various protein domain and family characterisation databases, post-translational modification databases, species-specific data collections, variant databases and disease databases. Accordingly, UniProtKB acts as a central hub for biomolecular information with cross references to 115 external databases. A document listing all databases cross-referenced in UniProtKB is available (www.uniprot.org/support/docs/dbxref.shtml) and contains a short description and the server URL for each database. This interconnectivity is achieved almost exclusively via DR (Database cross-Reference) lines. In addition, links from sub-sequences or particular sites to databases specialising in certain types of post-translational modifications or mutations are provided. Unique and stable feature identifiers (FTId) allow reference to a position-specific annotation item in the feature table. Currently, these are systematically attributed to FT VARIANT lines of human sequence entries, to alternative splicing events (VARSPPLIC), to all processed protein sequence (CHAIN, PROPEP, PEPTIDE) and to certain glycosylation sites (CARBOHYD), but will ultimately be assigned to all types of FT lines.

Minimal redundancy: for a given protein sequence, many sequence databases contain separate entries that correspond to different literature reports. In UniProtKB, we strive to merge such data to minimise redundancy. Differences between sequencing reports, caused by splice variants, polymorphisms, disease-causing mutations, experimental sequence modifications or sequencing errors, are indicated in the feature table of the corresponding UniProtKB/Swiss-Prot entry. At the level of UniProtKB/TrEMBL, all reports for the same organism that are identical over the full length of the protein are automatically merged.

The UniProtKB aims to describe all protein products derived from one gene from a specific species in a single record. In addition to assigning an accession number to each record, UniProtKB also assigns isoform identifiers (accession numbers for isoforms) to each protein form derived by alternative splicing, proteolytic cleavage, and post-translational modification. This is because different isoforms derived from the same gene can have different functions or biological roles, or might exist only during specific developmental stages or under certain environmental conditions. The freely available tool VARSPPLIC enables the re-creation of all annotated splice variants from the feature table of a UniProtKB entry, or for the complete database. A FASTA-formatted file containing all splice variants annotated in UniProtKB can be downloaded for use with similarity search programs.

Evidence attribution: the UniProt Consortium emphasises the use of an evidence attribution mechanism for protein annotation that will include, for all data, the data source, the types of evidence and methods for annotation. This is essential as UniProtKB contains data automatically imported from the underlying nucleotide sequence databases, data imported from other databases, data from specific programs, the results of automatic annotation systems and, most importantly, expert manual curation. The implementation of evidence tags allows the user to distinguish between these data sources and to easily identify classes of data of particular interest, such as experimentally proven protein annotation. Evidence tags for the annotation present in all UniProtKB/TrEMBL records and a proportion of UniProtKB/Swiss-Prot entries are available in the UniProtKB XML distribution and are also available from the UniProt website (www.uniprot.org). The full retrofit of the UniProtKB/Swiss-Prot entries is ongoing.

The UniProt Metagenomic and Environmental Sequences database

UniProtKB contains entries with a known taxonomic source so the availability of metagenomic data has necessitated the creation of a separate database, the UniProt Metagenomic and Environmental Sequences database (UniMES). Metagenomics is the large-scale genomic analysis of microbes recovered from environmental samples as opposed to laboratory-grown organisms which represent only a small proportion of the microbial world. UniMES currently contains data from the Global Ocean Sampling Expedition (GOS) which was originally submitted to the DDBJ/EMBL-Bank/GenBank databases. The initial GOS dataset is composed of 25 million DNA sequences primarily from oceanic

microbes and predicts nearly six million proteins. By combining the predicted protein sequences with automatic classification by InterPro, UniMES uniquely provides free access to the array of genomic information gathered from the sampling expeditions, enhanced by links to further analytical resources. The environmental sample data contained within this database is not present in the UniProtKB or the UniProt Reference Clusters but is integrated into UniParc. UniMES is available on the FTP site in FASTA format with UniMES matches to InterPro methods file (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/unimes/).

The UniProt Archive (UniParc)

While most protein sequence data are derived from the translation of DDBJ/EMBL-Bank/GenBank sequences, a significant amount of primary protein sequence data resulting from direct sequencing is submitted directly to UniProtKB. In addition, a large number of protein sequences are found in patent applications, as well as in entries from the Protein Data Bank (PDB). Given the wide variety of primary sources, the UniProt Archive (UniParc; Leinonen *et al.*, 2004) was created. UniParc is designed to capture all available protein sequence data – not just from the aforementioned databases, but also from sources such as Ensembl, the International Protein Index (IPI; Kersey *et al.*, 2004), RefSeq, FlyBase and WormBase. This combination of sources makes UniParc the most comprehensive publicly accessible, non-redundant protein sequence database available.

Although a protein sequence may exist in multiple databases and more than once in a given database, UniParc represents each protein sequence only once, assigning it a unique UniParc identifier. UniParc release 15.9 (October 2009) contained 18,394,349 unique sequences from 69,569,858 original source records. If a UniParc entry does not have a cross reference to a UniProtKB entry, the reason for the exclusion of that sequence from UniProtKB is provided (e.g. pseudogene). In addition, each source database accession number is tagged with its status in that database, indicating if the sequence still exists or has been deleted in the source database, with cross references to NCBI GI and TaxId if appropriate. From the total number of source records in the October release, 19,143,143 records were labelled as obsolete, indicating that the entry no longer exists in the source database with that sequence. A UniParc sequence version is incremented each time the underlying sequence changes, making it possible to observe sequence changes in all source databases. A specimen UniParc report can be found at www.uniprot.org/uniparc/UPI0000000C37. UniParc records carry no annotation, but this information can be found in the UniProtKB or other underlying databases.

The UniProt Reference Clusters (UniRef)

Automatic procedures have been developed to create three UniProt Reference Clusters, UniRef100, UniRef90 and UniRef50, from UniProtKB and selected UniParc entries as representative protein sequence databases with high information content. The databases provide complete coverage of sequence space while hiding redundant sequences from view. The non-redundancy allows faster sequence similarity searches (by using UniRef90 and UniRef50). Identical sequences and subfragments are presented as a single UniRef100 entry, containing the accession numbers of all merged entries and the protein sequence. UniRef90 and UniRef50 are built from UniRef100 to provide non-redundant sequence collections for the scientific user community to perform faster homology searches. Records from all source organisms with mutual sequence identity of >90% or >50%, respectively, are merged into a single record that links to the corresponding UniProtKB records. UniRef90 and UniRef50 yield a size reduction of approximately 40% and 70%, respectively. A specimen UniRef90 report can be found at www.uniprot.org/uniref/UniRef90_P57727.

GO ANNOTATION (GOA)

Yasmin Alam-Faruque, Daniel Barrell, David Binns, Emily Dimmer, Rachael Huntley, Tony Sawford

The Gene Ontology (GO) is a well-established structured vocabulary that has been successfully used in gene product functional annotation (The Gene Ontology Consortium, 2004). GO now contains over 28,400 terms, distributed over three ontologies that describe the molecular functions, biological processes and locations of action of gene products in a generic cell. The Gene Ontology Annotation (GOA) database (Barrell *et al.*, 2009; www.ebi.ac.uk/GOA) was created at the EBI in 2001. GOA's aim is to provide high-quality manual and electronic annotations to the proteins stored in UniProtKB using the GO vocabulary.

Over the last year GOA has provided another twelve file releases, which have included non-redundant sets of GO annotations to the human, mouse, rat, chicken, cow, zebrafish and *Arabidopsis* proteomes as well as data releases to all species (GOA-UniProtKB). GOA now provides more than 51 million GO annotations to over 6.3 million UniProtKB entries, covering more than 205,000 taxonomic groups. This represents a yearly increase of over 11.8 million GO annotations (a 53% increase) and a 17% increase in taxonomic coverage.

Manual GO annotations created by UniProtKB curators continue to be supplemented with the latest data from 22 external GO Consortium and specialist databases. By integrating GO annotations from in-house and external groups, GOA consolidates specialised knowledge to ensure that the database remains a key up-to-date reference for all species.

The GOA group is also the primary supplier of electronic GO annotations to the GO Consortium. In UniProtKB there are currently 204,395 taxonomic groups (6,295,938 proteins) for which electronic annotation pipelines, provided by GOA, are the only source of GO annotation. The group is responsible for providing four GO mapping files (Swiss-Prot keyword to GO, UniProtKB Subcellular Location to GO, Enzyme Commission numbers to GO and HAMAP to GO) which ‘translate’ external vocabularies to equivalent GO terms to create annotations for many species. Along with the InterPro2GO mapping, provided by the InterPro group, such mapping files are continually updated so that currently over 14,000 external terms are mapped to GO, providing more than 50 million annotations from entries in UniProtKB. In addition, the electronic annotation collaboration with Ensembl Compara has continued successfully; now providing over 409,671 annotations (a 177% increase over the year) for 47 species by transferring high-quality, experimentally-evidenced manual annotations to one:one orthologues in closely related species.

In 2009 GOA curators have been responsible for providing manual GO annotation training to Swiss-Prot curators at SIB, Geneva. Over the course of 2009 GOA has held five annotation training sessions at SIB, which were followed by intensive annotation checking. Currently 25 SIB curators have been trained and they have generated 14,567 manual GO annotations. GOA curators additionally continue to be key members of the GO Consortium Reference Genomes Initiative for the human proteome, working to provide detailed, high-quality annotations to human proteins. The group also supports the GO annotation activities of the British Heart Foundation initiative at UCL in London.

Two new staff members have joined the GOA team over the last year. In April in Yasmin Alam-Farouque joined to head the GOA renal annotation initiative (www.ebi.ac.uk/GOA/kidney), a curation project funded by Kidney Research UK, and in June Tony Sawford joined as the team’s main programmer. The GOA renal project intends to provide a valuable community resource for renal researchers by improving annotations and GO terms for mammalian proteins implicated in kidney development and disease.

RESID

John S. Garavelli

The RESID Database of Protein Modifications (Garavelli, 2004; www.ebi.ac.uk/RESID/) is a comprehensive collection of annotations and structures for protein modifications and cross-links including pre-, co-, and post-translational modifications. The database provides systematic and alternate names, atomic formulae and masses, and enzymatic activities that generate the modifications, keywords, literature citations and cross references to the Gene Ontology (GO), ChEBI, PSI-MOD, and PDB, structure diagrams and molecular models. Quarterly Release 59 (September 2009) contained 495 entries for chemically unique protein modifications.

The RESID Database documents the controlled vocabulary for natural protein modifications in the feature table annotations of UniProtKB. It was used during the first phase of the UniProt project in merging the feature annotations of Swiss-Prot and the PIR, and in designing new standard annotations. In ongoing work, the RESID Database is used to enhance the modification descriptions in the feature tables of UniProt entries. Information retrieval projects for the database uncover original reports for new types of modification and for modifications newly found in additional proteins. This information contributes to the annotation of UniProtKB by describing the newly discovered modifications, producing standard feature annotations for them, and predicting their occurrence in other entries through automated annotation. As an internet resource, the RESID Database assists researchers in high-throughput proteomics to search monoisotopic masses and mass differences, to identify known and predicted protein modifications, and to suggest the modified sequences from alternative isobaric peptides that are the most consistent with current knowledge of natural modifications. It is used as a contributing component of the Proteomics Standards Initiative ontology of protein modifications (PSI-MOD), which is maintained by John Garavelli and Luisa Montecchi-Palazzi for the Proteomics Standards Initiative and PRIDE.

PANDA OUTREACH AND TRAINING

The PANDA Outreach and Training activities are overseen by Rolf Apweiler and Ewan Birney.

OUTREACH

Jeff Almeida-King, Xosé M. Fernández, Bert Overduin, Michael Schuster, Giulietta Spudich

The PANDA group continues its effort to offer training worldwide, helping users to make the most of our resources. The group provides extensive support to the Outreach and Training team at the EBI as a core component of the courses based in Hinxton (such as the ‘A dip into EBI’s data resources’ overview courses) and Roadshows. Furthermore, beyond the outreach representatives, other PANDA developers provide training support in meetings.

- As part of the deliverables within SLING (Serving Life-science Information for the Next Generation, FP7) a number of workshops have taken place in Eastern European countries, and overall more than 24 countries have hosted PANDA training events. Figure 1 displays a world map where countries are coloured according to access to Ensembl and locations hosting PANDA training activities in the last year have been pinpointed.

- Ensembl (VectorBase, Ensembl Genomes) 92 courses
- Proteomics (UniProt, InterPro, IntAct, PRIDE) 25 courses
- Sequence databases (EMBL-Bank, EGA, GOA) 7 courses
- Pathways (Reactome, ChEBI) 9 courses
- EBI (Roadshows and Hinxton-based workshops) 16 courses

The group has attended large scientific meetings such as the 59th Annual Meeting of the American Society of Human Genetics, 17th Annual International Conference on Intelligent Systems for Molecular Biology, amongst others. Moreover, visibility of PANDA resources was increased by targeting specialised meetings such as the European Atherosclerosis Society's 'Genetics of Complex Diseases' organised by the British Atherosclerosis Society or the 3rd Biocuration Conference. The Swiss Institute of Bioinformatics hosted a training course for Swiss-Prot curators on GO annotation.

European Genome-phenome Archive (EGA)

Jeff Almeida-King joined the group in June 2009 to provide helpdesk support to Ensembl Genomes and the European Genome-phenome Archive. To illustrate the success of these projects, EGA has handled 442 user requests since then. A glance at our request system provides a picture of what users contact us for, from access to the different collections (account request, encryption key), to queries relating to the different datasets available. It is not surprising that the most frequent requests relate to the data generated by the Wellcome Trust Case Control Consortium (WTCCC), followed by WTCCC2 and MalariaGEN.

TRAINEE PROGRAMME

As part of EMBL-EBI's training mission, the PANDA group runs an active trainee programme. Undergraduates and PhD students (usually Marie Curie fellows) join PANDA for a period of three to twelve months, applying their theoretical knowledge to practical problems. In 2009, the group has hosted seven trainees and nine visitors, working on a broad variety of projects. The trainees were:

- Jigisha Anupama (VIT University, India), Chemogenomics team;
- Christoph Bueschl (FH Hagenberg, Austria), IntAct, Proteomics Services team;
- Gavin Ha (University of British Columbia, Canada), Discovery of copy number variable regions, Vertebrate Genomics team;

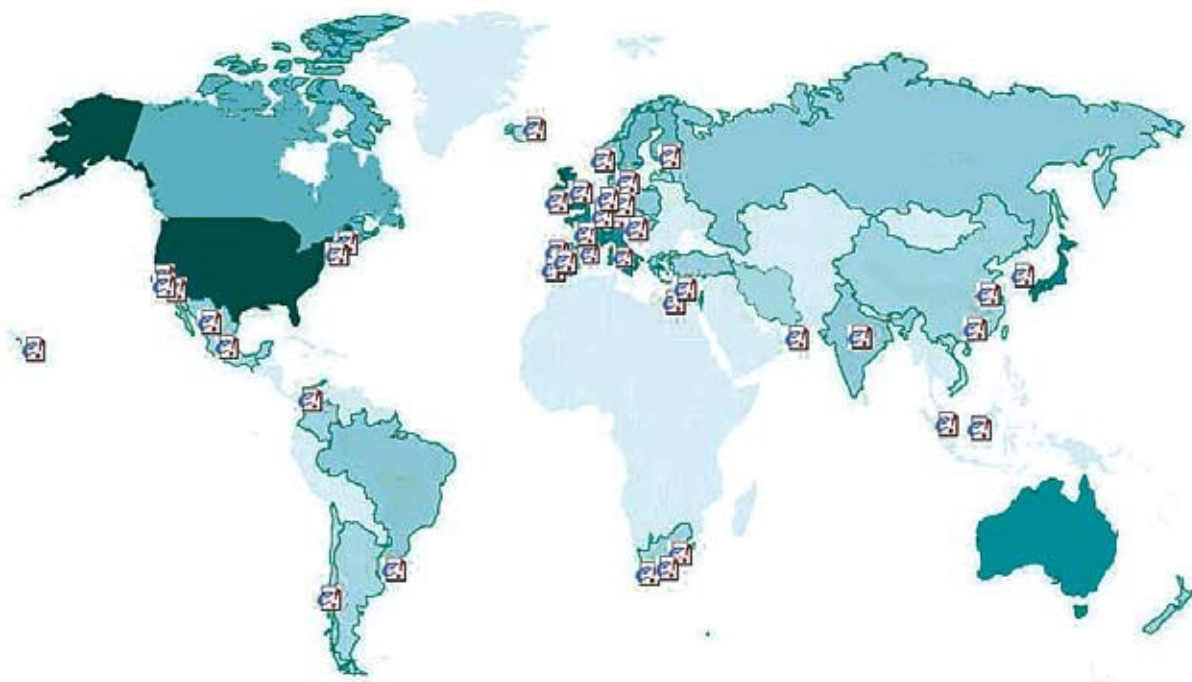


Figure 1. Geographical distribution of Ensembl access and locations that hosted PANDA training activities in 2008–2009. The degree of shading is representative of Ensembl access as recorded over a period of one month in 2009, where darker areas represent the regions where use of Ensembl was highest. Ensembl logos pinpoint the locations of Ensembl training events.

- Jules Kerssemakers (UMC Nijmegen, Netherlands), Proteomics services/structures, Proteomics Services team;
- Marcel Schulz (MPI Berlin, Germany), Compara, Ensembl team;
- Sander Timmer (University of Amsterdam, Netherlands), Analysis of 1000 Genomes data, Vertebrate Genomics team;
- Kristel van Eijk (University of Utrecht, Netherlands), Proteomics Services team.

Several project outputs have become part of EMBL-EBI's production process and resulted in co-authorships in publications and conference contributions.

APWEILER RESEARCH

Two PhD students (Joe Foster and Garth Ilsley) are currently working under the supervision of Rolf Apweiler. One PhD student (Michael Mueller) and one postdoc (Kai Ye) have successfully concluded their work during the reporting period. Another ongoing research activity is the automatic annotation of proteins, which has been described in the UniProt Knowledgebase section; see page 19.

Investigating the application of peptide retention time for improved transition selection in single reaction monitoring (SRM)

Joe Foster

Single reaction monitoring (SRM) has been utilised in the small molecules field for over 30 years, and has more recently been used for proteomics. While the majority of high abundance proteins are readily characterised by SRM, the difficulties lie with selecting likely transitions for proteins of low abundance. It is fairly well understood in the field that detection of peptides of low abundance proteins by mass spectrometry is thwarted by peptides of high abundance proteins that can competitively interfere with the ionisation and detection of less abundant peptides. By using predicted retention time information of peptides derived from low abundance proteins of interest, and by comparing these predictions to the actual Total Ion Chromatograms of the sample being measured, transition selection can be performed in context of the actual sample background. As a result, peptide candidates can be selected that represent the best chance for detection in the mass spectrometer, giving the experimentalist optimal conditions to successfully identify and quantify proteins of interest.

Improved exploration of large biological datasets

Garth Ilsley

Most studies of the dynamics of gene expression have used tissue samples hybridised to gene expression microarrays. Unfortunately, their usefulness is limited when studying gene expression during development since spatial information is lost. More recently, however, *in situ* hybridisation methods have been improved to the point of providing quantitative data. One example is the Berkeley Drosophila Transcription Network Project, which includes the gene expression levels of almost 100 genes in approximately 6,000 nuclei over a key period of development. Various statistical models have been applied to this dataset, recovering known regulatory relationships and suggesting new ones. These predictions are currently being tested by our collaborators. Additionally, the form of the best-fitting models is suggestive of an underlying cis-regulatory module structure. This is informing further work, both computational and experimental.

Estimating the scope and selectivity of a targeted proteomics approach based on combinatorial proteolysis

Michael Mueller

Dynamic range and complexity of the proteome result in limitations of shotgun proteomics affecting sensitivity and confidence of protein identification. This has led to the recent emergence of more targeted approaches in mass spectrometry based proteomics. These are based on the targeted detection of proteotypic peptides by single reaction monitoring (SRM), a highly sensitive and selective peptide identification method.

An in-depth *in silico* analysis of proteolytic digests of human protein sequences suggests that targeted proteomics approaches might be partially compromised in most experimental protocols. This is due to the absence of candidate proteotypic peptides for a significant proportion of the proteome if samples are digested with trypsin. Furthermore, the analysis shows that this shortcoming can be overcome to a significant extent through the diversification of the peptide population by a combinatorial proteolytic digest. An estimation of the detectability of candidate proteotypic peptides in the generated peptide mixture by SRM suggests that the majority of target peptides can be detected by monitoring a small number of peaks in the product ion spectrum.

Detecting breakpoints of large deletions and medium sized insertions on the low coverage samples of 1000 Genomes Project and high coverage samples of Cancer Genome Project from paired-end short reads

Kai Ye

In order to investigate disease-related variants in genetic surveys of large populations, we developed a so-called Pindel method to identify breakpoints of large deletions (1bp-10kb) and medium sized insertion (1-20bp) at base level precision from 36bp paired-end short reads. In the pre-processing step, all reads are mapped to the reference genome. Then the mapping results are examined to select the paired reads where only one end can be mapped. For each of those remaining pairs, the mapped reads must be uniquely located in the genome while their mates cannot be mapped to anywhere in the genome under a given threshold alignment score. Our Pindel program uses the mapped read to determine the anchor point on the reference genome and the direction of the unmapped read. Knowing the anchor point, the direction to search for the unmapped read and the user-defined Maximum Deletion Size, Pindel will break the unmapped reads into two (deletion) or three (short insertion) fragments and map the two terminal fragments separately.

We also adapted the Pindel algorithm to process multiple samples. We added tags to the reads to indicate their sources and then ran Pindel using the entire pool of reads as the input. We modified the Pindel program to report the sample sources of the supporting reads for each identified indel event. We also adapted the algorithm to include detection of deletions with small insertions of non-templated sequences at the breakpoint, and to report a confidence score that is monotonically related to false discovery rate. We demonstrated our method with the low coverage data of 170 individuals in the 1000 Genomes Project and high coverage data in the Cancer Genome Project.

BIRNEY RESEARCH

Ewan Birney's research group focuses on algorithmic methods for genome analysis. Two PhD students are currently working under the supervision of Ewan Birney (Markus Hsi-Yang Fritz and Dace Ruklisa). Two PhD students (Alison Meynert and Daniel Zerbino) have successfully concluded their work during the reporting period. In addition, one joint EIPOD postdoc works in the group (Mikhail Spivakov).

Hominid segmental duplications and repeat evolution

Markus Hsi-Yang Fritz

Using the recent sequencing of the extinct hominid, Neanderthal, we have been probing the evolution of human segmental duplications and transposons. Traditionally this analysis has difficult due to low coverage of a whole genome shotgun approach, but by focusing on specific aspects of these processes we can probe for the presence or absence of certain segmental duplications and transposon insertions, placing them before or after the modern human to Neanderthal split.

Investigations into weak binding motifs and the broader scale evolution of regulatory regions

Alison Meynert

We have used suffix array based methods to discover potential motifs that work in a weak binding manner, in particular in ultra conserved regions. Using these motifs we have extended the steric and binding energy model to include cooperativity effects. Weak binding motifs are clearly critical in transcriptional control from this analysis. We have explored the potential way that such weak binding models might work, and we have developed a logic gate-like model of transcriptional control. In collaboration with Laurence Ettwiller from the University of Heidelberg, we have shown that this model can predict important motifs in Medaka fish. We also developed a method of identifying transcription factor binding sites affected by single nucleotide changes in regulatory DNA. The method is based on simple probabilistic models of transcription factor binding. It was applied to single base changes observed in promoter regions of two cancer genomes from the Cancer Genome Project, and identified a number of mutations that caused statistically significant changes to potential binding sites of particular transcription factors.

Large-scale association studies: from inference framework to effects

Dace Ruklisa

Genome wide studies are very important when trying to understand genetic contribution to complex phenotypes and traits. However, such studies require proper solutions to two problems: the first is scale of analysis, and the second is choice of adequate models. We propose an inference framework that is suitable for a wide variety of experimental settings and model types while mitigating some of the technical problems. The principal part of our modelling approach is stepwise model selection, where a set of genetically/statistically significant loci is augmented by one locus at a time. Such incremental building of a model seems to be more correct than just a single scan of the genome in search of interesting regions, because all genetic effects are estimated jointly. We have adapted our model selection method for parallel computing via the Map/Reduce paradigm. Using this method, we were able to discover loci associated with

various diseases in the human case/control data from WTCCC involving more than 5,000 individuals and covering seven diseases (Crohn's disease, coronary artery disease, hypertension, bipolar disorder, type I diabetes, type II diabetes and rheumatoid arthritis). We plan to perform similar types of association studies for loci that are crucial for *Drosophila* oocyte development.

De Bruijn graph representation of DNA sequence

Daniel Zerbino

De Bruijn graphs show great promise for DNA analysis as a de Bruijn graph simultaneously represents valid assemblies of fragmented data, compression of DNA sequence and multiple alignments between different species. De Bruijn graphs can be created with read lengths as low as 30bp and this has allowed us to develop an accurate *de novo* assembler from short read pairs (30bp) which provides N50s up to 150kb in bacterial sequences and up to 100kb in larger genomes including large sections of human. This technology has already revolutionised the ability to generate genome assemblies with the next-generation sequencing machines, with a number of smaller genomes of pathogens sequenced at very small cost. We developed a tool called Velvet, which now counts at least 400 registered users and which will be extended in future to handle structural variation.

EIPOD project: *Drosophila* mesoderm development

Mikhail Spivakov

Using ChIP-chip and more recently ChIP-seq from *Drosophila* mesoderm at a variety of time points we have been studying how the *Drosophila* developmental system specifies organogenesis in the fly. We have also recently started analysing the effects of individual genetic variation on developmental cis-regulatory networks in *Drosophila*. The distribution of variation in *Drosophila* is varied, as expected, and includes variation in fundamental genetic components, such as development. Overall there is clear evidence of negative selection on developmentally important motifs, but also intriguing signals of potential positive or balancing selection. This is a joint project with Eileen Furlong's group at EMBL Heidelberg.

FUTURE PROJECTS AND GOALS

It is our intention to work on improved integration and synchronisation of all PANDA resources. In addition to major improvements of our current systems, we intend to add mining of high-throughput genomics and proteomics datasets to our automatic annotation toolset. Despite the abundance of data from large-scale experimentation on a genome-wide level, such as expression profiling, protein-protein interaction screens or protein localisation, the systematic and integrated use of this type of information for high-throughput annotation of proteins remains largely unexplored. We therefore intend to build on ongoing research activities at EMBL-EBI to develop and assess new protocols to integrate and analyse functional genomics datasets for the purpose of high-throughput annotation of uncharacterised proteins. This will include the analysis of different data types regarding their suitability for the approach, development of data structures that allow the efficient integration and mining of data of different types and quality, as well as benchmarking of the obtained results and the application of the new methodologies to UniProtKB/TrEMBL annotation.

Team Members

Joint Team Leader PANDA Group (Proteins)

Rolf Apweiler

Joint Team Leader PANDA Group (Nucleotides)

Ewan Birney

Team Leaders

Guy Cochrane*
Paul Flicek
Henning Hermjakob
Sarah Hunter
Paul Kersey
John Overington
Christoph Steinbeck

Group Coordinators

Elsbeth Bruford
Paula de Matos*
Lennart Martens
Maria-Jesus Martin
Claire O'Donovan
Glenn Proctor

Project Coordinators

Laura Clarke*
Fiona Cunningham
Emily Dimmer
Xosé M. Fernández
Javier Herrero-Sanchez
Anne Hersey*
Pascal Kahlem
Ilkka Lappalainen*
Daniel Lawson

Rasko Leinonen
David Lonsdale
Michele Magrane
Sandra Orchard*
Manuela Pruess
Esther Schmidt
Damian Smedley*
Dan Staines*
Robert Vaughan

Senior Scientific Database Curators

Paul Browne
Nadeem Faruque
John Stephen Garavelli
Kati Laiho*
Jennifer McDowall

cont.

Scientific Database Curators

Ruth Akhtar
Yasmin Alam-Faruque
Louisa Bellis*
Patricia Bento*
Mark Bingley*
Wei Mun Chan
Louise Daugherty
Ruth Eberhardt
Marcus Ennis
Rebecca Foulger
Phani Garapati
Richard Gibson
Susan Gordon*
Christopher Hunter
Rachael Huntley
Julius Jacobsen
Jyoti Khadake
Bijay Jassal
Steven Jupe*
Duncan Legge
Yvonne Light*
Gaurab Mukherjee
Petra ten Hoopen
Ruth Seal*
Inma Spiteri*
David Thorneycroft
Stephen Turner*
Matthew Wright
Siew-Yit Yong*

Bioinformaticians

Mark Bingley*
Yuan Chen
David Croft
Martin Hammond*
Gemma Hoad
Michael Maguire*
Craig McAnulla
Diego Poggioli*

Senior Software Engineers

Ricardo Antunes*
Daniel Barrell*
Richard Côté
Alexander Fedotov
Alan Horne
Phil Jones
Samuel Kerrien
Michael Kleen*
Jie Luo
John Maslen
Samuel Patient
Antony Quinn
Mark Rijnbeek
Peter Sterk*
Manjula Thimma

Software Engineers

Premanand Achuthan*
Rafael Alcantara Martin
Bruno Aranda
Benoit Ballester
Kathryn Beal
Benoit Bely*
David Binns
Lawrence Bower
Mario Caccamo*
Chao-Kung Chen*
Ying Cheng*
Manuel Corpas*
Ujjwal Das
Mark Davies*
Bernard de Bono
Adriano Dekker*
Fehmi Demiralp

Paul Derwent*
Stephen Fitzgerald
Anna Gaulton*
Renato Golin*
Neil Goodgame
Leo Gordon
Matthias Haimel
Janna Hastings
Kenneth Haug
Jonathan Hinton
Zamin Iqbal*
Mikyung Jang
Andrew Jenkinson
Nathan Johnson
Szilveszter Juhos
Andreas Kahari
Ikeda Kazuyoshi*
Damian Keefe
Stephen Keenan
Arnaud Kerhornou
Rhoda Kinsella
Gautier Koscielny
Stefan Kuhn
Eugene Kulesha
Vasudev Kumanduri
Davang Lakhani*
Wudong Liu
Quan Lin
Ian Longden
Michael Lush
Uma Mahashwari*
Edoardo Marcora*
Shaun McGlinchey*
Karyn Megy
Gavin O'Kelly*
Rajesh Radhakrishnan
Florian Reisinger
Daniel Rios
Tony Sawford*
Andrey Sitnov*
Guy Slater
Richard Smith*
Siamak Sobhany
Gilleain Torrance*
Albert Vilella
Juan A. Vizcaino
Steven Wilder
Phil Wilkinson
Andy Yates
Vadim Zalunin
Holly Zheng-Bradley*

Helpdesk Officers

Jeff Almeida-King*
Bert Overduin
Michael Schuster
Giulietta Spudich

Database Administrators

Matt Corbett*
Giuseppe di Martino
Mike Donnelly
Pieter van Rensburg

Postdocs

Mikhail Spivakov
Kai Ye*

Group Secretaries

Shelley Goddard
Tracy Mumford

Administrative Assistant

Kerry Smith

Data Assistant

Sheila Plaister

PhD Students

Joe Foster*
Andre Faure*
Markus Fritz*
Garth Ilsley
Alison Meynert*
Pablo Moreno*
Michael Mueller*
Dace Ruklisa
Petra Schwalie*
Daniel Zerbino*

Students

Jigisha Anupama*
Christoph Bueschl*
Gavin Ha*
Jules Kerssemakers*
Marcel Schulz*
Sander Timmer*
Kristel van Eijk*

Visitors

Cele Abad-Zapatero
Kirill Degtyarenko
Ian Dunham
Alex Mitchell
Marvin Mundry
Augusto Rendon
Will Spooner
Eleanor Stanley
Matthieu Visser
Valerie Wood

* Indicates part of the year only

Publications**Apweiler****2008**

Barrell, D., *et al.* (2008). The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, D396-403

Chatr-Aryamontri, A., *et al.* (2008). MINT and IntAct contribute to the Second BioCreative challenge: Serving the text-mining community with high quality molecular interaction data. *Genome Biol.*, 9, Suppl 2, article S5

Eisenacher, M., *et al.* (2008). Proteomics data collection – 3rd ProDaC Workshop: April 22 nd 2008, Toledo, Spain. *Proteomics*, 8, 4163-4167

Helsens, K., *et al.* (2008). Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol. Cell. Proteomics*, 7, 2364-2372

Jimenez, R.C., *et al.* (2008). Dasty2, an Ajax protein DAS client. *Bioinformatics*, 24, 2119-2121

Jones, A.R. & Orchard, S. (2008). Minimum reporting guidelines for proteomics released by the proteomics standards initiative. *Mol. Cell. Proteomics*, 7, 2067-2068

Koscielny, G., *et al.* (2008). ASTD: The Alternative Splicing and

cont.

- Transcript Diversity database. *Genomics*, 93, 213-220
- Kuhn, S., *et al.* (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, 9, 400
- O'Neill, K., *et al.* (2008). OntoDas – A tool for facilitating the construction of complex queries to the Gene Ontology. *BMC Bioinformatics*, 9, 437
- Orchard, S., *et al.* (2008). Annual Spring Meeting of the Proteomics Standards Initiative 23-25 April 2008, Toledo, Spain. *Proteomics*, 8, 4168-4172
- Reeves, G.A., *et al.* (2008). The protein feature ontology: A tool for the unification of protein feature annotations. *Bioinformatics*, 24, 2767-2772
- The Uniprot Consortium. (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 36, D190-D195
- 2009**
- Apweiler, R., *et al.* (2009). Approaching clinical proteomics: Current state and future fields of application in fluid proteomics. *Clin. Chem. Lab. Med.*, 47, 724-744
- Apweiler, R., *et al.* (2009). The Universal Protein resource (UniProt) 2009. *Nucleic Acids Res.*, 37, D169-D174
- Barrell, D., *et al.* (2009). The GOA database in 2009 – An integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, 37, D396-D403
- Barsnes, H., *et al.* (2009). OMS SA Parser: An open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics*, 9, 3772-3774
- Barsnes, H., *et al.* (2009). PRIDE converter: Making proteomics data-sharing easy. *Nat. Biotechnol.*, 27, 598-599
- Bell, A.W., *et al.* (2009). A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods*, 6, 423-430
- Berriman, M., *et al.* (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature*, 460, 352-358
- Blankenburg, H., *et al.* (2009). DASMI: Exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25, 1321-1328
- Cao, H., *et al.* (2009). Comparative genomics indicates the mammalian CD33rSiglec locus evolved by an ancient large-scale inverse duplication and suggests all Siglecs share a common ancestral region. *Immunogenetics*, 61, 1-17
- Couto, F., *et al.* (2009). Verification of Uncurated Protein Annotations. In 'Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration', IGI 311-325, Global Publishing
- Degtyarenko, K., *et al.* (2009). ChEBI: An open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics*, Suppl 26, chapter 14, unit 14.9, 1-20
- Eisenacher, M., *et al.* (2009). Proteomics Data Collection – 4th ProDaC workshop, 15 August 2008, Amsterdam, the Netherlands. *Proteomics*, 9, 218-222
- Eisenacher, M., *et al.* (2009). Proteomics Data Collection – 5th ProDaC Workshop 4 March 2009, Kolymari, Crete, Greece. *Proteomics*, 9, 3626-3629
- Eisenacher, M., *et al.* (2009). Getting a grip on proteomics data – Proteomics Data Collection (ProDaC). *Proteomics*, 9, 3928-3933
- Furnham, N., *et al.* (2009). Missing in action: Enzyme functional annotations in biological databases. *Nat. Chem. Biol.*, 5, 521-525
- Harland, L. & Gaulton, A. (2009). Drug target central. *Expert Opin. Drug Discov.*, 4, 857-872
- Hunter, S., *et al.* (2009). InterPro: The integrative protein signature database. *Nucleic Acids Res.*, 37, D211-D215
- Illesley, G.R., *et al.* (2009). Know your limits: Assumptions, constraints and interpretation in systems biology. *Biochim. Biophys. Acta*, 1794, 1280-1287
- Jain, E., *et al.* (2009). Infrastructure for the life sciences: Design and implementation of the UniProt website. *BMC Bioinformatics*, 10, 136
- Kathiresan, T., *et al.* (2009). A protein interaction network for the large conductance Ca²⁺-activated K⁺ channel in the mouse cochlea. *Mol. Cell. Proteomics*, 8, 1972-1987
- Koscielny, G., *et al.* (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93, 213-220
- Kuhn, S., *et al.* (2009). Components for computer-assisted structure elucidation. *Chem. Cent. J.*, 3, 62
- Kuhn, T., *et al.* (2009). Creating chemo- and bioinformatics workflows, further developments within the CDK-Taverna Project. *Chem. Cent. J.*, 3, 42
- Martens, L. & Apweiler, R. (2009). Algorithms and databases. *Methods Mol. Biol.*, 564, 245-259
- Matthews, L., *et al.* (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37, D619-D622
- Mehta, A. & Orchard, S. (2009). Nucleoside diphosphate kinase (NDPK, NM23, AWD): recent regulatory advances in endocytosis, metastasis, psoriasis, insulin release, fetal erythroid lineage and heart failure; translational medicine exemplified. *Mol. Cell. Biochem.*, 329, 1-13
- Montecchi-Palazzi, L., *et al.* (2009). The PSI semantic validator: a framework to check minimum information about a proteomics experiment compliance of proteomics data. *Proteomics*, in press
- O'Connor, M.N., *et al.* (2009). Functional genomics in zebrafish permits rapid characterization of novel platelet membrane proteins. *Blood*, 113, 4754-4762
- Orchard, S. (2009). Ending the "publish and vanish" culture: How the data standardization process will assist in data harvesting. *J. Proteome Res.*, 8, 3219
- Orchard, S., *et al.* (2009a). Second Joint HUPO publication and PSI Workshop 24th April 2009, Turku, Finland. *Proteomics*, 9, 4426-4428
- Orchard, S., *et al.* (2009b). Annual Spring Meeting of the Proteomics Standards Initiative, 27-29 April 2009, Turku, Finland. *Proteomics*, 9, 4429-4432
- Orchard, S., *et al.* (2009). Managing the data explosion: A report on the HUPO-PSI workshop August 2008, Amsterdam, the Netherlands. *Proteomics*, 9, 499-501
- Orchard, S. & Ping, P. (2009). HUPO world congress publication committee meeting. *Proteomics*, 9, 502-503
- Orchard, S. & Taylor, C.F. (2009). Debunking minimum information myths: one hat need not fit all. *N. Biotechnol.*, 25, 171-172
- Overington, J. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput. Aided Mol. Des.*, 23, 195-198
- Persson, B., *et al.* (2009). The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.*, 178, 94-98
- Rodriguez, H., *et al.* (2009). Recommendations from the 2008

International Summit on proteomics data release and sharing policy: The Amsterdam principles. *J. Proteome Res.*, 8, 3689-3692

Steinbeck, C., *et al.* (2009). New open drug activity data at EBI. *Chem. Cent. J.*, 3, 62

The Uniprot Consortium. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, 37, D169-D174

Vizcaino, J.A., *et al.* (2009). A guide to the PRIDE proteomics data repository. *Proteomics*, 9, 4276-4283

Vizcaino, J.A., *et al.* (2009). Charting online OMICS resources: A navigational chart for clinical researchers. *Proteom. Clin. Appl.*, 3, 18-29

Watkins, N.A., *et al.* (2009). A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, 113, e1-9

Yan, W.H., *et al.* (2009). Systematic comparison of the human saliva and plasma proteomes. *Proteom. Clin. Appl.*, 3, 116-124

Ye, K., *et al.* (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25, 2865-2871

Zhenyu, J., *et al.* (2009). Association study between gene expression and multiple relevant phenotypes with cluster analysis. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Pizzuti, C., Richie, M.D. & Giacobini, M. (eds), 5483, p1-12, Springer-Verlag

Other EMBL publications

Garavelli, J.S. (2004). The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics*, 4, 1527-1533

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32, D258-D261

Kersey, P.J., *et al.*, (2004). The International Protein Index: An integrated database for proteomics experiments. *Proteomics*, 4, 1985-1988

Leinonen, R., *et al.* (2004). UniProt Archive. *Bioinformatics*, 20, 3236-3237

Kretschmann, E., Fleischmann, W. & Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17, 920-926

Other publications

Gattiker, A., *et al.*, (2003). Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.* 27, 49-58

Natale, D.A., Vinayaka, C.R. & Wu, C.H. (2004). Large-scale, classification-driven, rule-based functional annotation of proteins. In 'Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics', Bioinformatics Volume, Subramaniam, S. (ed), John Wiley & Sons Ltd

Birney

2008

Fernandez-Suarez, X.M. & Birney, E. (2008). Advanced genomic data mining. *PLoS Comput. Biol.*, 4, e1000121

Holland, R.C.G., *et al.* (2008). BioJava: An open-source framework for bioinformatics. *Bioinformatics*, 24, 2096-2097

Paten, B., *et al.* (2008). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, 18, 1814-1828

Paten, B., *et al.* (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18, 1829-1843

Rakyan, V.K., *et al.* (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, 18, 1518-1529

Smedley, D., *et al.* (2008). Solutions for data integration in functional genomics: A critical assessment and case study. *Brief. Bioinform.*, 9, 532-544

2009

Cochrane, G., *et al.* (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37, D19-25

Durinck, S., *et al.* (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4, 1184-1191

Galperin, M.Y. & Cochrane, G.R. (2009). Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Res.*, 37, D1-D4

Gnad, F., *et al.* (2009). MAPU 2.0: High-accuracy proteomes mapped to genomes. *Nucleic Acids Res.*, 37, D902-D906

Haider, S., *et al.* (2009). BioMart

central portal – Unified access to biological data. *Nucleic Acids Res.*, 37, W23-W27

Hubbard, T.J.P., *et al.* (2009). Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697

Koscielny, G., *et al.* (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93, 213-220

Krestyaninova, M., *et al.* (2009). A system for information management in BioMedical studies – SIMBioMS. *Bioinformatics*, 25, 2768-2769

Lawson, D., *et al.* (2009). VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res.*, 37, D583-587

Matthews, L., *et al.* (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37, D619-D622

Megy, K., *et al.* (2009). Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. *Infection, Genetics and Evolution*, 9, 308-313

Paten, B., *et al.* (2009). Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, 25, 295-301

Pruitt, K.D., *et al.* (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, 19, 1316-1323

Sieglauff, D.H., *et al.* (2009). Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc. Natl Acad. Sci. USA*, 106, 3053-3058

Smedley, D., *et al.* (2009). BioMart – Biological queries made easy. *BMC Genomics*, 10, 22

Stolovitzky, G., *et al.* (2009). Annals of the New York Academy of Sciences: Preface. *Ann. N. Y. Acad. Sci.*, 1158, 9-12

Vilella, A.J., *et al.* (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19, 327-335

Wilkinson, P., *et al.* (2009). EMMA – mouse mutant resources for the international scientific community. *Nucleic Acids Res.*, in press



Guy Cochrane

*PhD 1999, University of East
Anglia.
At EMBL-EBI since 2002.
Team Leader since 2009.*

The European Nucleotide Archive Team

31

INTRODUCTION

The European Nucleotide Archive (ENA) provides a comprehensive repository for public nucleotide sequence data, attracting external users from a multitude of research disciplines and serving as underlying data infrastructure for PANDA services such as Ensembl, Ensembl Genomes and UniProt, and broader services such as ArrayExpress. The foundation for the ENA was the EMBL Data Library (latterly known as EMBL-Bank), which was established in EMBL Heidelberg in the early 1980s. While this component continues to be operated to this day, the mandate of the ENA has expanded enormously as sequencing technology has advanced and the breadth of applications to which sequencing can now be applied has grown. Broadly, ENA captures and presents the whole scale of sequencing information from raw data, through assembly and mapping information that relates very fragmented raw sequence reads into contigs and higher order structures, through to high-level interpretations of the function of parts of nucleic acid molecules, in the form of functional annotation. The ENA achieves comprehensive coverage through partnerships with other global bioinformatics service providers, namely NCBI in the US and DDBJ in Japan. The longest running ENA collaboration, the International Nucleotide Sequence Database Collaboration (INSDC; www.insdc.org/), has been underway for over a quarter of a century and now serves as a model for data sharing in the life sciences.

CURRENT STATUS OF THE ENA

A three-tiered structure

Three conceptual tiers within ENA provide abstraction from the underlying legacy infrastructure which has resulted from the integration of three databases: the original EMBL Data Library, the Trace Archive that was established at the Wellcome Trust Sanger Institute in the early 2000s and the newly established Sequence Read Archive (SRA). The three tiers are defined as follows:

- **'reads'** – sequencing machine output, base calls and quality scores;
- **'assembly'** – information relating overlapping fragmented sequence reads to contigs and covering higher order structures where contigs are structured into representations of complete biological molecules, such as chromosomes;
- **'annotation'** – where interpretations of biological function are projected onto coordinate-defined regions of assembled sequence in the form of annotation.
- Associated with these core tiers are a number of auxiliary databases that provide integration across ENA and serve to expand the information content of particular parts of ENA. These include:
- **'sample'** – information relating to the biological sample studied in the sequencing experiment;
- **'project'** – high level records that serve to unite content otherwise dispersed across the three ENA tiers.

A public repository

Presentation of data in ENA and its INSDC partner databases has become the globally accepted means of public sequence data dissemination. INSDC database records serve as a complement to traditional literature publications. All major journals require the availability of sequence data in INSDC databases and established practice in large-scale public genomics studies is for pre-publication data to be made available shortly after its generation. While ENA does

much to improve integration of records into the broader services offered at EMBL-EBI and elsewhere, in contrast to most other data resources within the PANDA team, ownership of ENA records, and hence editorial control, remains with the original submitters of the data.

Access to ENA services

The range of submission, update and retrieval services, along with complete documentation of the ENA project are available from our website at www.ebi.ac.uk/ena/. Queries, collaboration proposals and feedback are all very welcome at datasubs@ebi.ac.uk.

Growth and the future

ENA continues to grow rapidly and currently consumes almost half a petabyte of disk space and comprises 164 million assembled sequences (just under seven million of which are supported by at least some assembly information), over 40 million functional annotations, 110 million cross references to records in external databases, 56 million links to over 200,000 literature publications, 3,676 complete genomes (of which 69 are eukaryotic, 979 prokaryotic and 2,628 non-cellular), 1.94 billion capillary traces and the output of 6,419 next-generation sequencer production runs. While the increasing rates of growth in assembled and annotated sequences have continued (see figure 1), we expect that over the next few years, the impact of next-generation sequencing platforms will bring shorter doubling times to ENA. Not only will such growth challenge our submissions services and processing pipelines through sheer data throughput, but as a result of the cost reductions that the new technologies bring, we expect the use of sequencing as a core platform for general genomic and transcriptomic assays (such as ChIP-seq and sequence-based expression studies) to increase on a massive scale. Although the traditional expertise of the ENA team in the treatment of sequencing data largely intended for *de novo* assembly and annotation pipelines provides a healthy starting position to rise to these challenges, we recognise the need to develop innovative approaches to the handling of incoming data into ENA that allow us to scale our operations usefully. Strategies under development include partnering with expert resources already established at EMBL-EBI to offer combined services (such as ArrayExpress and Ensembl Variation), the development of rule-based curation procedures that focus the efforts of biologists away from individual records and onto whole classes of information, integration efforts to provide context for ENA records, leverage of community efforts to standardise information through minimal reporting standards and ontology development and the curation of project records to provide systematic access to large datasets.

KEY DEVELOPMENTS

Next-generation sequence data and the Sequence Read Archive

The Sequence Read Archive (SRA), the newest and fastest growing component of the ENA, provides the setting for the major content and technology developments of 2009. This year, we have seen a substantial growth in content of 3,670

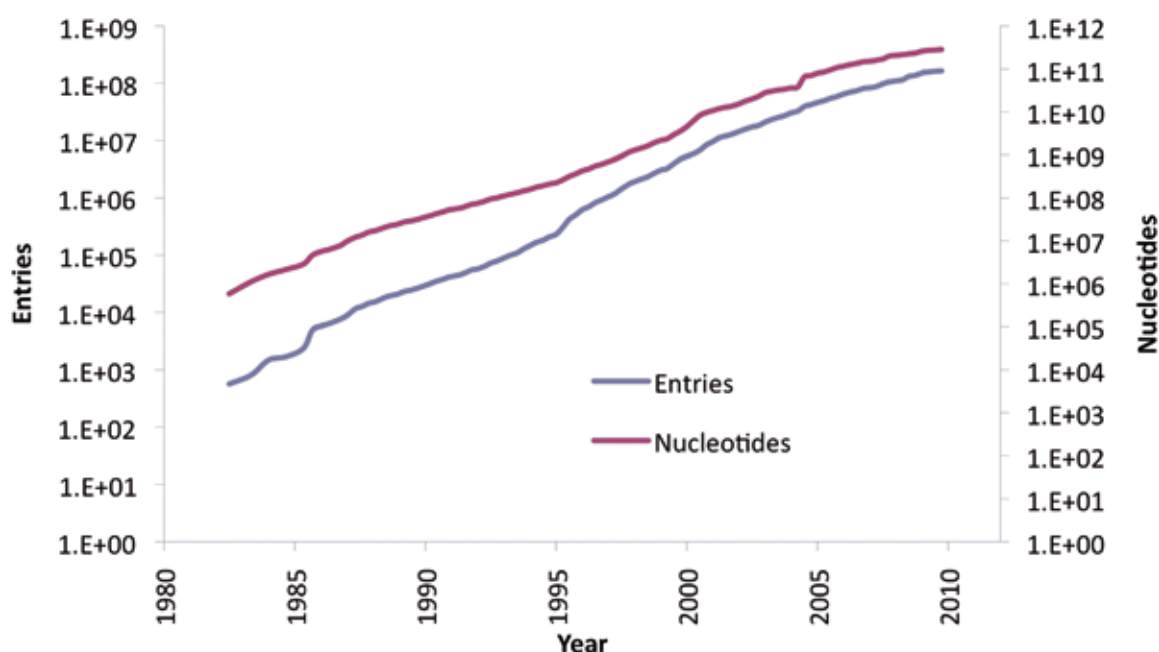


Figure 1. Data growth for assembled/annotated sequences in ENA-Annotation.

production runs on next-generation machines, representing an increase of 134% over last year's figure. As expected, SRA already contributes the major sequence volume to ENA (see figure 2). This year, we have further developed our data model for the robust persistence of next-generation read metadata and data and have introduced complete submission pipelines for both laboratory submitters and sequencing centres, data mirroring technology to capture daily submissions to NCBI and data presentation tools, albeit preliminary, for the entirety of SRA contents.

Next-generation sequencing technologies share a massively increased parallel throughput with respect to capillary sequencing machines, the previous dominant platform for sequencing. As a result, the information relating to a production run of a sequencing machine differs enormously in its nature; per unit of information that describes the sample and experimental configuration, the number of data points is vastly increased with respect to previous technologies. The data model developed for SRA takes advantage of this property of next-generation data by optimising treatment of metadata (information relating to the sequenced sample and experimental configuration) and data (intensity, read and quality information emitted by the sequencing machine) in isolation from each other. For the comparatively low volume metadata, we have developed a relational representation that relates sample, experiment, run, study and submission objects and provides a high level of support for queries relating to these relationships. For the high volume data component we have deployed a custom file system populated by data files that can be accessed through a number of routes. Our overall principle is that the selection of datasets of interest will be performed as far as possible at the level of metadata (for which we are currently developing search and browse functions) and retrieval of these limited datasets is a final step in the user's interaction with SRA.

At the 2009 collaborative meeting of the INSDC, it was decided that the SRA lay clearly within the mandate of the INSDC and that it would henceforth be considered an INSDC activity. The NCBI launched their service prior to EMBL-EBI and the Japanese partner, DDBJ, is developing its service initially for local data providers. While the traditional INSDC collaboration mode has been one of data sharing using an agreed exchange format, it was clear that a further level of collaboration around technologies that are developed and adapted for the purposes of handling next-generation sequencing data was required for SRA.

We have adopted technology from Aspera (www.asperasoft.com/) that serves to optimise access to existing network bandwidth for SRA data transfers, both for data exchange with NCBI and for submission and retrieval of data by users. In addition, we have contributed to the `io_lib` set of open source tools for the handling of high volume sequence data in Sequence Read Format (which we support as a submission, archiving and retrieval format).

Recognising that we must develop an archive that remains sustainable in the long term, we have explored data reduction and cross-INSDC redundancy removal options. We have established, not least through our outreach activities, a community discussion on the long-term value of the more raw forms of data that are produced by next-generation machines (such as Illumina intensity data). These discussions are progressing towards a set of standard recommendations as to the level to which raw data should be captured for a given sequencing application and sample type. In relation to the reduction of redundancy across INSDC, we have explored options to provide catastrophe data backup services between NCBI and EMBL-EBI, through either the persistence of only a single copy of any one high volume dataset at both of the sites, or the persistence of the dataset along with a backup copy at only one of the sites, according to the route through which data were submitted. A continued exchange between NCBI and EMBL-EBI of the comparatively low volumes of metadata, and perhaps lightweight components of data, such as sequence and quality

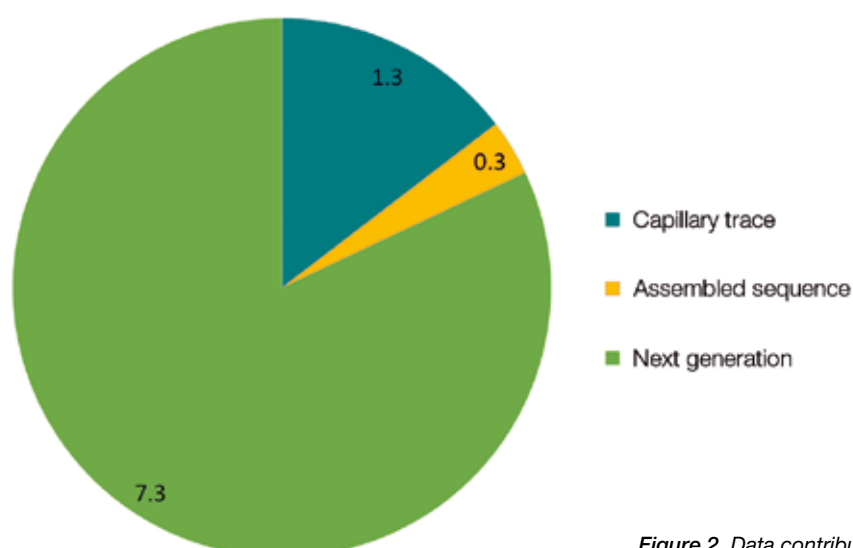


Figure 2. Data contribution to ENA.

series, and the development of a common API to provide users with access to remote data on demand would ensure the continuation of a comprehensive SRA service.

In delivering the SRA to users, we recognise that many of the long-established working practices and procedures associated with sequencing studies must change to reflect new features of the technologies. As such, we launched the Next Generation Sequencing Wellcome Trust Hinxton Retreat annual workshop series in December 2008, in partnership with the Wellcome Trust. This invitation-only workshop drew together manufacturers, tools developers, SRA colleagues from the NCBI and sequencing application experts, and attracted sponsorship from next-generation platform manufacturers. The aim of the first workshop was to facilitate discussion and collaborative working across the sequencing disciplines in order to optimise the utility of next-generation technologies, and this aim will be carried through into a second workshop, which we are currently organising. In addition, the ENA team organised a further workshop in collaboration with the PANDA outreach team, focusing on user training around SRA and analysis tools developed at EMBL-EBI and elsewhere on campus.

We have developed two pipelines for the submission of sequence data from next-generation sequencing platforms. Just as the SRA data model provides isolation between metadata and data, our submission procedures allow for separate treatment of the two components. Data files, prepared in Sequence Read Format (<http://srf.sourceforge.net>) or native machine formats, are submitted with limited need for input from the submitter and metadata are submitted with greater interaction between EMBL-EBI and the submitter, reflecting the high degree of validation that can be achieved. Laboratory submitters currently alert the ENA team that they have a submission pending, a private FTP drop box is created and they are invited to upload data files and metadata prepared using a *pro forma* spreadsheet or XML files that we have helped them to create. Once files have been received, the ENA team validates, allocates accessions and uploads the data into the production database. In due course, we will integrate the spreadsheet functionality into our interactive Webin submission tool. Sequencing centre submitters, in contrast, typically run LIMS systems, in which many pieces of information that are ultimately required as metadata for a submission to ENA have already been stored for tracking purposes. For these submissions, we receive metadata in XML format through a web service that was designed in collaboration with the European sequencing centres. The web service provides a management tool for the upload and tracking of both XML metadata files and data files.

Interactive submissions of annotated sequence through Webin

In mid-February 2009, we launched public beta testing of a new Webin submission service to trial functions specifically for the handling of small sets of annotated sequences. Interactive web applications are well established as the methods of choice for the submission of small-scale datasets, particularly from infrequent and non-expert submitters who have limited knowledge of ENA data structures and limited bioinformatics expertise. With a view to retiring the ten-year old Perl-CGI Webin application that has supported the submission of over a million annotated sequences to date, we embarked some time ago on the replacement of our underlying core submissions infrastructure. Our goal for this technology is to provide facilities such as robust rule-based validation (see below), extensibility and support for large-scale genome submissions and we are expecting to launch further services in November 2009. We expect that such improvements will provide us with the capacity to continue to maintain and improve upon our traditional quality standards. Included in the 2009 beta release are an integrated Lucene-based taxonomy browser and search facility (capable of resolving taxonomic names and their synonyms and of providing visualisation of taxonomic classifications), an integrated Feature Table Definitions browser (that allows context-dependent user-reporting of the latest definitions, value formats, usage examples and comments for features and qualifiers) and improved grouping of features and qualifiers in annotation pages.

Rule-based validator

The pressure of increased data volumes and increased breadth in sequencing applications is most notable in ENA data input workflows, where biological curation is applied to incoming data to ensure consistency and optimal downstream utility. In line with our strategy for continual review and automation of workflows when new technologies and knowledge become available, we have developed a rule-based data validation system, primarily targeting sample information and functional annotation. Using this tool, an ENA curator maintains a set of rules, each of which can be applied to a class of data. An example might be a rule constraining the use of a particular annotation structure to a given evolutionary clade. With such a rule system, a single application of biological knowledge impacts upon multiple records from multiple submissions. We have launched this system in the first instance as part of the beta Webin submission application release. While this initial launch is limited in its coverage, further development of the validator will allow its application in larger-scale data submission pipelines and data exchange pipelines from INSDC, with ultimate aim that it will be applied to all ENA sample and annotation content.

FUTURE PROJECTS AND GOALS

The focus of current development work is the presentation of data. A comprehensive browser for all ENA data types is under development and is expected to be launched in 2010. Accompanying this will be a data retrieval web service and text-based search functionality using the EB-eye search engine. We are currently working on sequence similarity search tools optimised for unassembled next-generation sequence reads, based on de Bruijn graph representations that have been successfully used at EMBL-EBI and elsewhere in sequence assembly work. In addition, we will deliver further submission components as part of our ongoing submissions infrastructure replacement programme and will continue to develop our scalable approach to face the challenging growth in global nucleotide sequencing activity.

Team Members

Coordinator

Nadeem Faruque
Rasko Leinonen
Robert Vaughan

Data Assistant

Sheila Plaister

Scientific Database Curator

Petra ten Hoopen
Christopher Hunter
Gaurab Mukherjee*

Senior Scientific Database Curator

Ruth Akhtar
Richard Gibson

Senior Software Engineer

Fehmi Demiralp

Siamak Sobhany

Software Engineer

Lawrence Bower
Ying Cheng*
Neil Googdame*
Rajesh Radhakrishnan
Vadim Zalunin*

Bioinformatician

Gemma Hoad
Mikyung Jang
Szilveszter Juhos*
Quan Lin
Michael Maguire*

Visitor

Steven Leonard

* Indicates part of the year only

Publications

2009

Cochrane, G., *et al.* (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37, D19-25

Galperin, M.Y. & Cochrane, G.R. (2009). Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Res.*, 37, D1-D4



Paul Flicek

DSc 2004, Washington University.

At EMBL-EBI since 2005.

Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008.

Team Leader since 2008.

Vertebrate Genomics

37

INTRODUCTION

The Vertebrate Genomics team is a combined service and research group that creates and manages data resources focusing on genome annotation and variation. The team's research is on computational epigenomics with a particular focus on the integration of diverse data types, such as DNA-protein interactions, epigenetic modifications, and the DNA sequence itself, in the context of comparative genomics.

The major service projects of the Vertebrate Genomics team are Ensembl, the European Genome-phenome Archive, the data coordination centre for the 1000 Genomes Project and the mouse informatics team. In support of these projects, we develop large-scale and novel bioinformatics infrastructure aimed at integrated data analysis and provision of data to the scientific community.

COMPUTATIONAL EPIGENOMICS

Mapping human methylomes

In collaboration with Vardhman Ramanan, Barts and the London; Stephan Beck, University College London; Thomas Down, Wellcome Trust Sanger Institute; Natalie Thorne and Simon Tavaré, Cancer Research UK

DNA methylation is required for genome function; it is a key regulator of gene expression in normal tissue and aberrant DNA methylation is a hallmark of certain cancers. Genome-wide reference DNA methylation profiles in multiple tissues and the means to compare these profiles is critical to understanding the role of DNA methylation, as is the identification of tissue-specific differentially methylated regions (tDMRs) that are thought to play a role in cellular identity. Using a custom designed microarray and the MeDIP (methylated DNA immunoprecipitation) technique, the team and our collaborators presented the most comprehensive set of DNA methylation profiles, consisting of 13 normal human somatic tissues in addition to human placenta, sperm and the GM06990 lymphoblastoid cell line (Raykan *et al.*, 2008). The results suggested that promoters across a wide range of CpG densities are regulated by tissue-specific DNA methylation and demonstrated that exon methylation is a common feature in mammalian genomes. These profiles are currently available in Ensembl (www.ensembl.org), the infrastructure of which can be used by the community to present similar data.

Analysis of DNA-protein interactions

In collaboration with Thomas Down, Ian Dunham and David Vetrie while all were members of the Wellcome Trust Sanger Institute

The genome-wide variability of transcription factor binding in individual cell types and the relationship of this binding to cellular identity is largely unknown. An investigation to map the binding of REST (repressor element 1-silencing transcription factor) across eight human cell lines leveraged analysis methods recently developed in the group and further led to the development of methods to compare positive regions in multiple cell types. The study exposed several interesting characteristics of the transcription factor binding site usage across a single species (Bruce *et al.*, 2009). The experiment was conducted on a PCR tiling array platform across the ENCODE regions and included seven cell lines expressing REST (see figure 1a) and the KELLY cell line which does not express REST. After analysis, a total of 591 positive regions were identified across the expressing cell lines, while the non-expressing KELLY cells were found to have no positive regions, providing confidence in a low false positive rate for the analysis. The positive regions were

further categorised by whether they appeared in a single cell line, all seven cell lines, or multiple but not all cell lines (see figure 1a). These groupings identified a core set of approximately 30 positive binding sites across the ENCODE regions and larger sets with restricted or unique binding patterns, which corresponded to the enrichment values observed on the array and the strength of the motif with respect to the consensus. Unexpectedly, the DNA sequence in the restricted and unique binding sites shows increased evolutionary constraint (at every conservation threshold) compared to the common sites (figure 1b). The genes closest to the binding sites with restricted binding profiles were enriched in tissue-specific genes, and we hypothesise that the higher conservation in these cases is analogous to the observed higher conservation of alternatively spliced exons with tissue or condition-specific expression patterns.

ENSEMBL

Ensembl (Hubbard *et al.*, 2009), a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute, provides an integrated set of tools for genome annotation, data mining and visualisation. Ensembl's mission is to enable genomic science by providing high-quality, integrated annotation on chordate genomes within a consistent and accessible infrastructure. At EMBL-EBI, Ensembl includes members of the Vertebrate Genomics team and components of the PANDA Nucleotides group (see page 17).

The Ensembl genome browser at www.ensembl.org is the primary entry point for most users. In addition to the website, we also provide data access through a number of other routes including an extensively supported Perl API, the Ensembl BioMart (Smedley *et al.*, 2009), direct queries of our publicly available MySQL databases, full download of all resources and the provision of Ensembl data as one of the Public Data Sets available on the Amazon Web Services cloud computing platform (<http://aws.amazon.com/publicdatasets/>). Ensembl places no restriction on the use of the data and provides all of the code through an open source licence that allows it to be used without cost by any interested organisation.

This year, the two most significant project achievements were the launch of the new Ensembl web interface in November 2008 and the release of the annotation set for the updated GRCh37 version of the human genome assembly in July 2009. The new website was the result of approximately one year of development and was designed to enable greater discovery of the numerous data resources provided by Ensembl with easier and more intuitive navigation. These features were implemented such that the overall speed of the website also saw significant improvements. Since the launch of the new web interface, we have concentrated on increasing overall performance and consolidating previously existing features within the new interface. Ensembl's support for the updated GRCh37 human assembly included a new gene set incorporating both automatic Ensembl gene predictions and manually annotated genes from the Havana project. This combined gene set is created within the context of the GENCODE project. In addition to genome annotation, support for the new assembly included the update of all of the pairwise and multi-species whole genome alignments as well as the mapping of genome variation and Ensembl regulatory features.

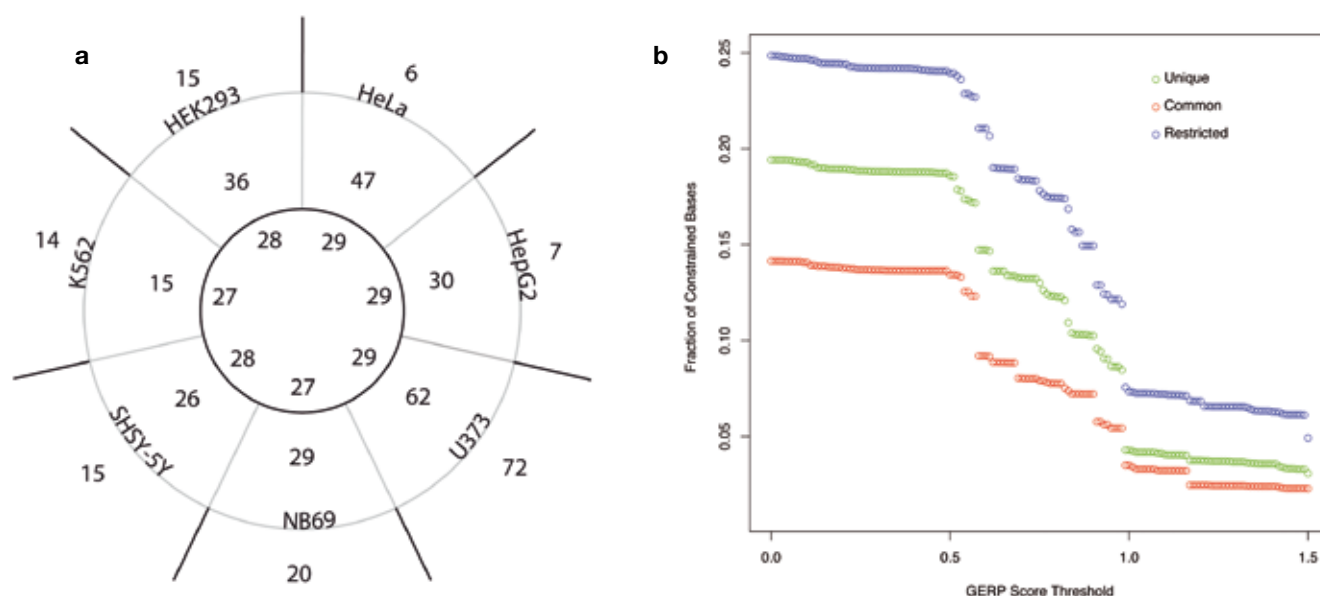


Figure 1. Figures from Bruce *et al.*, (2009). (a) A pinwheel diagram showing the seven REST expressing cell lines and the pattern of overlap of the nestie identified REST binding sites in each of the cell lines. The numbers in the centre of the circle represent the binding sites common to all cell lines and are not equal due to cases in which two regions in one cell line overlapped one region in a second cell line. The outer numbers represent binding sites unique to the given cell line. (b) The amount of evolutionary constraint for GERP score level thresholds and categories of REST binding sites.

Beyond the major efforts detailed above, there were five full Ensembl releases during the period of this report. From the September 2009 release onwards, Ensembl fully supports a total of 24 high coverage chordate genomes and 23 low coverage chordate genomes including the seven new species introduced this year; the anole lizard (*Anolis carolinensis*), the first reptile in Ensembl; the two-toed sloth (*Choloepus hoffmanni*), the white-tufted-ear marmoset (*Callithrix jacchus*), the pig (*Sus scrofa*), the Tamar wallaby (*Macropus eugenii*), the zebra finch (*Taeniopygia guttata*) and the Western lowland gorilla (*Gorilla gorilla*). Of these, the anole lizard, zebra finch, marmoset and pig were high coverage genome assemblies based on approximately 4–6x coverage from Sanger-style sequencing reads and gorilla was the first example of an assembly that combined traditional Sanger-style sequencing at low coverage with high-throughput short read sequencing at high coverage. The lamprey (*Petromyzon marinus*), another high coverage chordate genome, is currently provided with preliminary support only. An additional three non-chordate species (*Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*) are included to facilitate comparative analysis.

In order to increase consistency of the Ensembl resources, we have been steadily increasing our contacts and collaborative activities with similar resources at the University of California Santa Cruz (UCSC) and the NCBI. This year, the first Joint NCBI-EBI Coordination meeting in Washington was attended by all of the Ensembl and Genome Variation project leaders. We also have connections to many model organism-specific database resources such as the Rat Genome Database (RGD). The goal of these connections is to provide the wider research community with data resources that are maximally consistent and interconnected.

Ensembl maintains a significant commitment to user support and training. During the past year, our training team presented nearly 100 training events in over 20 countries. These events range from relatively short presentations as part of larger EMBL-EBI or Wellcome Trust workshops to intensive multiple day courses dedicated to the Ensembl API and those developers maintaining full Ensembl mirror sites. We have also developed a library of video tutorials for users not able to attend a course in person and these are now provided through the Ensembl YouTube channel.

The Ensembl infrastructure is being leveraged by the Ensembl Genome project and this has resulted in the generalisation of key aspects of the Ensembl toolset to support the requirements of the Ensembl Genomes project in novel areas. Additionally, the Ensembl core software team, which is part of the PANDA Nucleotides group (see page 17) have reengineered several key components of the Ensembl infrastructure, especially those supporting the mapping of external database identifiers to the Ensembl identifiers and the management of Gene Ontology (GO) information within the Ensembl databases.

Ensembl comparative genomics

Javier Herrero, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Albert Vilella

Ensembl's comparative genomics resources include pairwise and multi-species whole genome alignments as well as the calculation of homology relationships through Ensembl families and gene trees. As the number of supported species within Ensembl increases, the value of the comparative genomics resources to connect all aspects of the project increases. These resources also provide valuable information about the regions of the well-annotated human and mouse genomes which are subject to evolutionary constraint.

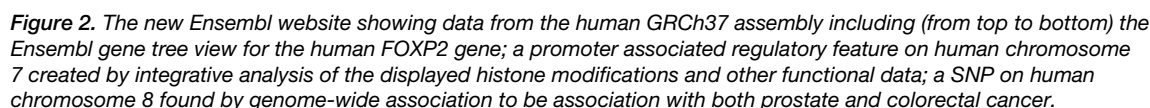
The Ensembl multi-species alignments are produced by the recently published Enredo-Pecan-Ortheus (EPO) pipeline (Paten *et al.*, 2008a; Paten *et al.*, 2008b; Paten *et al.*, 2009) and summarised as follows. In the first step Enredo, a graph-based method that is robust to duplicated regions within the genome, is used to identify orthologous and paralogous collinear genomic regions. Pecan, a consistency-based multiple aligner, is then used to create alignment blocks from the Enredo-identified collinear segments. These alignment blocks are used by Ortheus to infer ancestral sequences using a branch transducer model of sequence evolution that includes insertions and deletions. We extend Ensembl's multiple alignments to low coverage genomes by first constructing the core multiple alignment and then mapping each low coverage genome using pairwise alignments to the human genome. This procedure allows us to better determine sequence constraints through mammalian evolution.

In addition to the alignment resources, Ensembl comparative genomics also provides comprehensive predictions of vertebrate gene phylogeny which result in gene trees that are presented graphically on the Ensembl genome browser (see figure 2) and have recently been described in detail (Vilella *et al.*, 2009). Over the course of this year, we have implemented a number of improvements in the GeneTree pipeline to reflect actual or artefactual gene-split events and improved the GeneTree visualisation to aid in the interpretation of the trees. Phylogenetic predictions are complemented by Ensembl Families which feature alignments of homologous UniProt entries to the Ensembl proteins.

Ensembl functional genomics

Ian Dunham, Stefan Gräf, Nathan Johnson, Damian Keefe, Steven Wilder

The Ensembl functional genomics resources include the Ensembl regulatory build, an integrated analysis of experimental assays designed to create an automatic, evidenced-based annotation of genome function. The regulatory build uses several data types including genome-wide chromatin state maps, experimentally determined locations of



protein–DNA interactions and identified sites of transcriptional complexes such as RNA PolII to annotate regions of the genome with specific function. We also provide comprehensive annotation of gene expression arrays for supported species in the context of the Ensembl gene annotations. Array annotations are displayed within the Ensembl genome browser and are extensively used through the biomaRt package of Bioconductor.

Over the past year we released two updates to the set of human regulatory features, which incorporated newly published datasets, and the first mouse release of the Ensembl regulatory build with a focus on embryonic stem (ES) cells. We have also launched a new visual display to support the Ensembl regulatory features (see figure 2), which provides information about the identity and structure of the supporting histone modifications, sequence specific factors, DNase I hypersensitive sites and other experimental data used in the annotation of the Ensembl regulatory features. The view appears as a 'regulation' tab at the top of the page.

Ensembl functional genomics includes significant participation in the ENCODE project including the project management of the ENCODE Data Analysis Center (DAC), a collaborative effort of data analysis groups in six institutions around the world. This effort requires coordination with the ENCODE Data Collection Center (DCC) at UCSC and includes participation in the development of integrated analysis algorithms for the various functional data types produced within the project. Incorporation of data from the ENCODE project into Ensembl will be an important area of focus over the next twelve months.

Ensembl variation

Fiona Cunningham, Yuan Chen, Will McLaren

Ensembl's variation resources support a subset of the genomes available in Ensembl, mostly defined by those with genome variation data available within the dbSNP archive. For many species, a large fraction of the variation data was originally generated by Ensembl itself, including the new reference SNP sets for the orangutan and zebra finch genomes that were produced last year as part of our participation in those genome projects. These datasets join the reference SNP sets that we previously created for rat, mouse, platypus, tetraodon and other species. All Ensembl-created SNPs are submitted to dbSNP and, due to the dbSNP build cycle, are usually accessible within Ensembl before they appear in a formal dbSNP release. We continue to work closely with dbSNP and import the data for our supported species after each dbSNP release.

This year we assigned annotations to more than 1,100 SNPs that have been found through genome-wide association to be associated with approximately 200 phenotypes. These SNPs are primarily from the curated Catalog of Published Genome-Wide Association Studies (www.genome.gov/26525384), which has been created by the NHGRI's Office of Population Genomics. All Ensembl variation resources are presented in a consistent interface including a dedicated 'variation' tab that was introduced in the new Ensembl web interface in November 2008 (see figure 2). Information provided for each SNP includes ancestral alleles drawn from multiple alignments and OrthoUS ancestral sequence reconstructions (see Ensembl comparative genomics section), allele frequency and individual genotypes assayed in populations such as those used by the HapMap project, and locations of the individual SNPs with respect to Ensembl's gene annotations.

Ensembl variation has worked closely with the 1000 Genomes Project to create supporting databases for the preliminary project results. This resource will be incorporated into Ensembl as the final results become available.

GENOME VARIATION

European Genome-phenome Archive

Ilkka Lappalainen, Jeff Almeida-King, Mario Caccamo, Jonathan Hinton, Vasudev Kumanduri

The European Genome-phenome Archive (EGA) is a permanent repository for all types of potentially identifiable individual-level research data including array-based phenotype information, genotypes, genome sequences and other sequence-based assays. EGA archives and provides access to both individual-level genetic data and certain summary-level data, such as allele frequency data from cohorts. These summary-level data can be used in combination with an individual's genotype data to determine whether that individual is a member of a given cohort following multiple methods published in the last year. In addition to the infrastructure for secure data storage and distribution, we have developed an extensive data consistency suite to ensure that datasets submitted to the archive are internally coherent.

The EGA displayed considerable growth in its second year of existence and currently hosts datasets from nine major consortia, which users may request access to through the appropriate data access committees. The submitted datasets include supporting data for major publications, such as the geographical population structure of Northern Europeans and the molecular basis of breast and ovarian cancers. In total, the EGA now contains individual-level data for more than 45,000 individuals and summary-level data for an additional 20,000 individuals. The number of users approved to access one or more of our data collections has grown by three-fold over the past year to more than 1,200 researchers worldwide.

The EGA is currently working within two major international consortia – the International Human Microbiome Consortium (IHMC) and the International Cancer Genome Consortium (ICGC) – during the organisational and early stages of these projects. The EGA is an appropriate repository for the potentially identifiable human data produced by all non-United States groups in both projects and, in working with the data coordination groups, we will facilitate data sharing of the resources produced.

The 1000 Genomes Project

Laura Clarke, Zamin Iqbal, Richard Smith, Holly Zheng-Bradley

The 1000 Genomes Project will create the most comprehensive public catalogue of human variation in major world populations by leveraging next-generation sequencing technology. The Vertebrate Genomics team, in collaboration with the NCBI, leads the data coordination centre for the project and is responsible for most aspects of project data management and tracking.

Over the past year, the 1000 Genomes Project, which includes sequencing centres in the UK, Germany, China and the United States and numerous data analysis groups worldwide, completed the data production and initial analysis associated with the pilot phase of the project. The pilot phase includes three distinct scientific experiments: 1) the sequencing at approximately 2x coverage of 180 individuals equally divided among members of European, African and East Asian populations; 2) sequencing at greater than 30x coverage of six individuals from European and African trios consisting of both parents and a female child; 3) the targeted capture and sequencing of approximately 1,000 genes from a collection of nearly 800 individuals. Based on the experience of the pilots, the main phase of the 1000 Genomes Project has been defined as the sequencing to 4x coverage of approximately 1,200 individuals from three major geographic locations of significant medical genetics interest. Sequencing dedicated to the main production phase of the project began in the summer of 2009 and is expected to be completed in the first half of 2010.

Throughout the year, the 1000 Genomes Project has released intermediate and pre-publication results to provide immediate benefits to the wider scientific community. The Vertebrate Genomics team launched a 1000 Genomes Project-specific genome browser based on the Ensembl platform at <http://browser.1000genomes.org>. The 1000 Genomes browser allows project partners and any member of the scientific community to access the pre-publication data released by the project in a compact form and through a user-friendly interface. We also manage the 1000 Genomes Project website (www.1000genomes.org) with assistance from the EBI's External Services team.

MOUSE INFORMATICS

Damian Smedley, Chao-Kung Chen, Edoardo Marcora, Phil Wilkinson

The mouse informatics project coordinates EBI's interaction with major mouse genome and biology resources around the world. Regarding informatics resources in particular, we collaborate closely with the Mouse Genome Informatics (MGI) group at the Jackson Laboratory and the NCBI. Within Europe, major partnerships have been established with the EMMA project (European Mouse Mutant Archive; www.emmanet.org), for which we run the database and website, and the CASIMIR project (www.casimir.org.uk) for the coordination and sustainability of international mouse informatics resources. The CASIMIR project is focused on coordinating policy on use cases, interoperability, standards for mouse informatics and identifying technical issues in the field. As part of our lead on interoperability we published a review of the technical approaches, including a proof of principle study, and took part in an international meeting sponsored by CASIMIR to discuss and recommend best practices for data and tools sharing in the context of mouse biology.

This year saw the start of two new projects that expanded the scope and reach of mouse informatics. The first, CREATE, is focused on mice using cre recombinase driver strain technologies and our role will be to develop a database and website of these lines to ensure that the resource is maximally useful to the scientific community. The second project, called the I-DCC (International Data Coordination Centre; www.i-dcc.org), will create a single point of access to information from the International Knockout Mouse Consortium's efforts to produce knockout lines for all protein-coding genes in mouse (www.knockoutmouse.org). Our role in the I-DCC is to utilise and develop the BioMart data management system (www.biomart.org) to support federated queries across the multiple data collections of the IKMC and our primary collaborators in this effort are located at the Wellcome Trust Sanger Institute and Jackson Laboratory.

FUTURE PROJECTS AND GOALS

Our work on the analysis of DNA-protein interactions is continuing with the comprehensive and evolutionary-based analysis of individual transcription factors in matched tissues from several species. We are also exploring the utility of DNA methylation profiles for the prediction of genome function.

Developments in services, including support of a public copy number and structural variation (CNV/SV) database, have been ongoing as part of the larger genome variation effort. These developments will result in the launch of the

CNV/SV database in late 2009/early 2010. This will also provide data to Ensembl to ensure that the CNV/SV data generated over the past several years is made as widely available as possible. The database infrastructure will also be leveraged by the EGA project to provide support for CNV/SV data generated from cohort and disease studies. Ensembl continues to adapt to high-throughput sequencing data and next year we expect to release gene sets and multi-species alignments including genome assemblies created entirely with next-generation sequencing data. More traditional genomes assemblies created from Sanger-style sequencing reads will see enhanced gene sets that are extensively informed by RNA-seq data. Ensembl will also focus on the display and annotation of variation data. This effort is supported by our participation in the Locus Reference Genomic (LRG) consortium (www.lrg-sequence.org), though which we plan to incorporate specific summary data from Locus Specific Databases (LSDBs).

Team Members

Project Leaders

Laura Clarke (Resequencing Informatics)
Fiona Cunningham (Ensembl Variation)
Ian Dunham* (Ensembl Functional Genomics)
Javier Herrero (Ensembl Compara)
Ilkka Lappalainen (Variation Archive)
Damian Smedley (Mouse Informatics)

Scientific Programmers

Mario Caccamo*
Stefan Gräf*
Jonathan Hinton
Zamin Iqbal*
Damian Keefe
Vasudev Kumanduri
Damian Keefe
Richard Smith*
Steven Wilder
Holly Zheng-Bradley*

Ensembl Developers

Kathryn Beal
Yuan Chen
Stephen Fitzgerald
Leo Gordon
Will McLaren*
Albert Vilella

Bioinformaticians

Chao-Kung Chen
Edoardo Marcora
Phil Wilkinson

Postdoctoral Fellow

Benoit Ballester

User Support Officer

Jeff Almeida-King*

PhD Student

Petra Schwalie

Visitors

Andre Faure
Gavin Ha
Hallam Stevens
Sander Timmer

Team Secretary

Kerry Smith*

*Indicates part of the year only

Publications

2008

Coghlan, A., *et al.* (2008). nGASP – The nematode genome annotation assessment project. *BMC Bioinformatics*, 9, 549

Cotton, R.G., *et al.* (2008). GENETICS. The Human Variome Project. *Science*, 322, 861-862

Paten, B., *et al.* (2008a). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, 18, 1814-1828

Paten, B., *et al.* (2008b). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18, 1829-1843

Rakyan, V.K., *et al.* (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, 18, 1518-1529

2009

Bruce, A.W., *et al.* (2009). Functional diversity for REST (NRSF) is defined by in vivo binding affinity hierarchies at the DNA sequence level. *Genome Res.*, 19, 994-1005

Carlile, M., *et al.* (2009). Strand selective generation of endo-siRNAs from the Na/phosphate transporter gene Slc34a1 in murine tissues. *Nucleic Acids Res.*, 37, 2274-2282

Flicek, P. (2009). The need for speed. *Genome Biol.*, 10, 212

Flicek, P. & Birney, E. (2009).

Visualising the Epigenome. In 'Epigenomics', Ferguson-Smith, A.C., Gready, J.M. & Martienssen, R.A. (eds), 55-66, Springer, Netherlands

Haider, S., *et al.* (2009). BioMart central portal – Unified access to biological data. *Nucleic Acids Res.*, 37, W23-W27

Hubbard, T.J.P., *et al.* (2009). Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697

Kaput, J., *et al.* (2009). Planning the human variome project: The Spain report. *Hum. Mutat.*, 30, 496-510

Krestyaninova, M., *et al.* (2009). A system for information management in BioMedical studies – SIMBioMS. *Bioinformatics*, 25, 2768-2769

Morley, R.H., *et al.* (2009). A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. *Proc. Natl Acad. Sci. USA*, 106, 3829-3834

Paten, B., *et al.* (2009). Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, 25, 295-301

Pruitt, K.D., *et al.* (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, 19, 1316-1323

Smedley, D., *et al.* (2009). BioMart – Biological queries made easy. *BMC Genomics*, 10, 22

Vilella, A.J., *et al.* (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19, 327-335



Paul Kersey

*PhD 1995, University of
Edinburgh.
At EMBL since 1999.
Team Leader since 2008.*

The Ensembl Genomes Team

45

INTRODUCTION

The Ensembl Genomes team is responsible for the provision of services based on the genomes of non-vertebrate species. As the name of the team suggests, the team uses the Ensembl software framework, originally developed by EMBL-EBI and the Wellcome Trust Sanger Institute in the context of the human genome project, as its primary vehicle for achieving this. 2009 has seen the initial release of five new portals that provide public access to this genome-scale data under this strategy: Ensembl Bacteria, Ensembl Protists, Ensembl Fungi, Ensembl Plants and Ensembl Metazoa, complementing the coverage of vertebrate genomes in Ensembl.

The launch of these sites is in part a response to the huge explosion in data generation triggered by the development of high-throughput sequencing technologies. The Ensembl software system is a modular platform, with different modules capable of handling the analysis, display and distribution of different data types (e.g. variation data, comparative and functional genomics data). Often the raw data is hard to interpret unless integrated into its genomic context. The use of Ensembl technologies enables this; the data model and browser are centred on standard biological concepts (genes, SNPs, orthologue sets) that can be linked back to the underlying data providing the evidence for their existence and functional annotation.

The data explosion represents an opportunity, but also a challenge, because the success of Ensembl for vertebrates is due in large part to the scientific quality of the data it contains. It is not feasible for EMBL-EBI to become experts in every species whose genome has been sequenced. The solution to this problem is to work collaboratively with communities focused on individual species or collections of species. By combining their biological expertise and our own strength in infrastructure, it becomes possible to envisage an Ensembl for every species where there is sufficient justification – that is, a community generating data and actively interested in maximising its value over the potential lifespan of its usefulness. Through being involved with many communities simultaneously, we have the potential to influence annotation standards, perform comparative analyses of our own and facilitate further analyses by others, in a way that would not be possible if each community developed its own resources in isolation.

We are already working in tight collaboration with other groups producing Ensembl databases including VectorBase, WormBase, the Central Aspergillus Database Repository, and Gramene, and represent gene models produced by model organism databases including TAIR, SGD, FlyBase and GeneDB_Spombe. Additionally, the group is involved in maintaining four legacy databases that represent different aspects of genomic data: Integr8, Genome Reviews, IPI and ASTD. These projects are now in the process of being wound down following the launch of Ensembl Genomes and their residual functionality incorporated within the Ensembl, Ensembl Genomes, and UniProt sites. The reduction of the number of separate interfaces, and the unification of services through a more limited portfolio of sites, will significantly simplify navigation of EMBL-EBI resources, with the genome sequence (accessed through Ensembl or Ensembl Genomes) providing a natural index to many of the other data types handled at the EBI.

The team's other tasks include the provision of data relating to complete proteomes in the UniProt database, and information relating to splice variants in (vertebrate) Ensembl. Substantial training and outreach efforts are also part of the group's activities. Development of the Ensembl web code for the purposes of Ensembl Genomes takes place within the PANDA Nucleotides team led directly by Ewan Birney (see page 17).

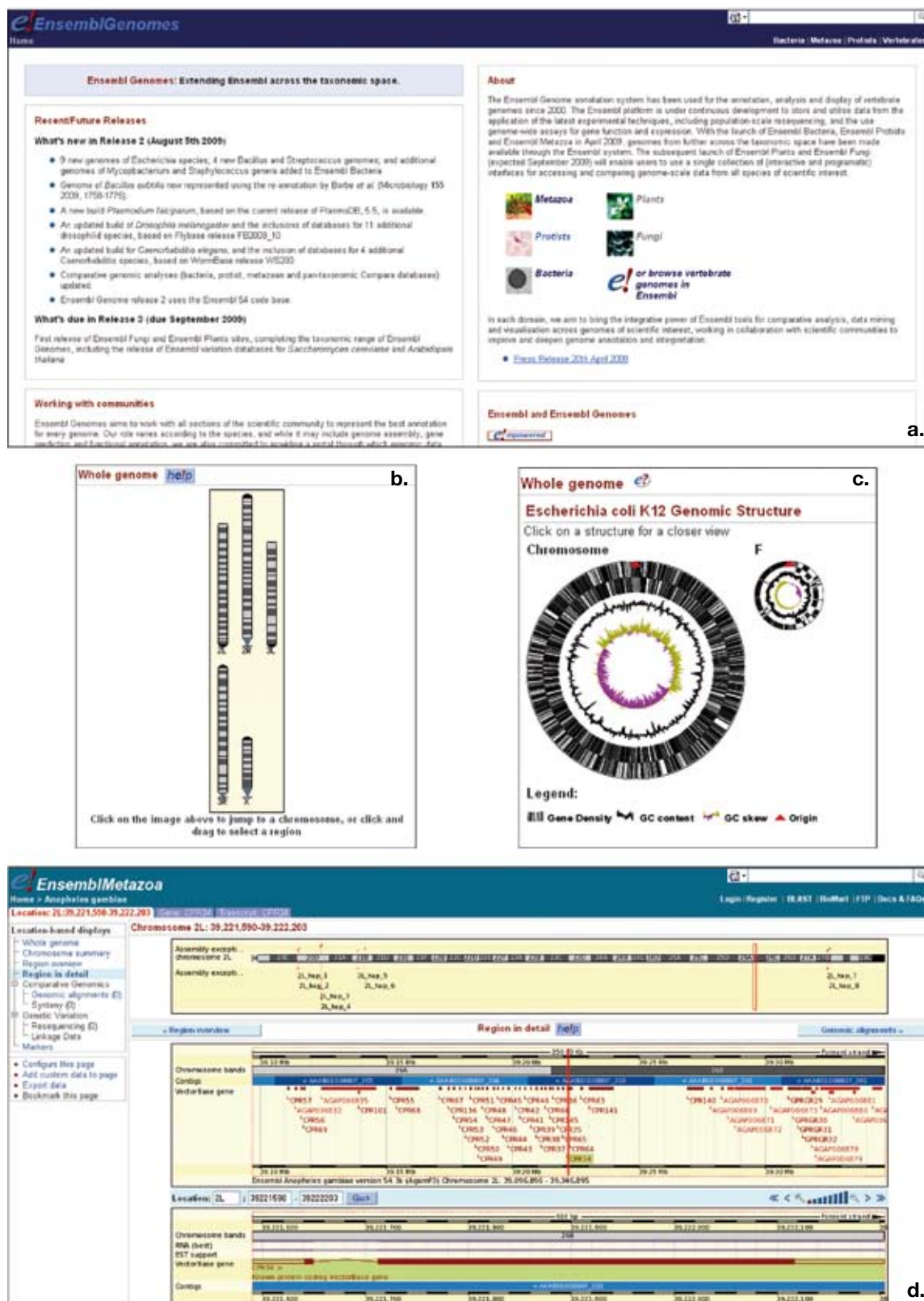


Figure 1. Some views of Ensembl Genomes (a) the homepage, (b) graphical karyotype view for *Anopheles gambiae*, using the same representation as found in vertebrate Ensembl, (c) the equivalent view for a bacterial genome, showing an alternative representation more suitable for circular chromosomes and plasmids. This view has been produced using the program `circular_diagram.pl`, (courtesy of Kim Rutherford), which has been incorporated into the Ensembl browser. (d) location view for a region of the left arm of chromosome 2 of *A. gambiae*, annotated by VectorBase. The image illustrates gene and (at a larger scale) transcript-based views. The views are track based, with features of particular types located in their own horizontal bands (which can be turned on or off using a control panel). On the lower panel, a single track (labelled 'RNA best EST support') provides the evidence for the transcript, which appears in a second track labelled 'VectorBase gene'. A variety of alternative location-based views can be selected in the left hand column; the tabs (at the base of the banner) allow users to switch between location, gene and transcript-centric views.

ENSEMBL GENOMES

Jeff Almeida-King, Paul Derwent, Alan Horne, Matthias Haimel, Martin Hammond, Arnaud Kerhornou, Paul Kersey, Gautier Koscielny, Devang Lakhani, Daniel Lawson, Michael Nuhn, Uma Maheswari, Karyn Megy, Daniel Staines, Andrew Yates

Since 1995, when the genome of a cellular organism was completely sequenced for the first time, genome sequencing has transformed the biological sciences. Today, in excess of 1,600 genomes have been sequenced, assembled, annotated and deposited in the public nucleotide archives; numerous other genomes exist in states of partial assembly and annotation; thousands of viral genomes sequenced have also been generated. Moreover, the increasing use of high-throughput sequencing technologies is rapidly reducing the cost of genome sequencing, leading to an accelerating rate of data production. This not only makes it likely that in the near future the genomes of all species of scientific interest will be sequenced, but also the genomes of many individuals, with the possibility of providing accurate and sophisticated annotation through the similarly low-cost application of functional assays. Many of these trends have first become visible in human genomics, but are spreading to other species as costs continue to fall. For example, the 1000 Genomes Project (to sequence 1,000 human genomes) has quickly been followed by the launch of similar initiatives in *Arabidopsis* (<http://1001genomes.org>), *Plasmodium* (www.genome.gov/26523588), *Drosophila* (www.dpgp.org), and other species.

Solutions for handling these data in vertebrate genomes have been successfully developed in the context of the Ensembl project (Hubbard *et al.*, 2007; a joint project of the EBI and the Wellcome Trust Sanger Institute; see the report by Paul Flicek on page 37). The main focus of the Ensembl Genomes team is to leverage the use of these solutions for non-vertebrate species through the five new Ensembl-based sites for bacteria, protists, fungi, plants and invertebrate metazoa. There are two big advantages of this approach. Firstly, the Ensembl software system has evolved in response to new data types that first appeared in human genomic studies, but which are now increasingly appearing in the context of other species; its re-use provides a cost-effective way of providing sophisticated data analysis and visualisation tools for a wider range of genomes. Secondly, the availability of data from across the taxonomy through a common set of interfaces (which include a graphical genome browser, FTP, BLAST search, a query optimised data warehouse, programmatic access, and a Perl API) greatly decreases the cost of inter-species data analysis such as comparative genomics or the study of pathogenesis. For example, data from malarial parasites (genus *Plasmodium*), the malarial vector (*Anopheles gambiae*) and the human host are now all available through Ensembl Protists, Ensembl Metazoa, and Ensembl. Exploiting this facility, we make a pan-taxonomic comparative analysis available as part of each release of Ensembl Genomes, with sequence alignments and implied evolutionary histories available for both families spanning the individual clades and the entire taxonomic range. The Ensembl comparative analysis pipeline consists of two primary approaches; a protein-centric method (that produces gene trees, illustrating the implied evolutionary history of conserved families); and a DNA-centric method based on the use of Pecan, a high-accuracy multiple aligner; and both methods have been (selectively) applied to the Ensembl Genomes divisions.

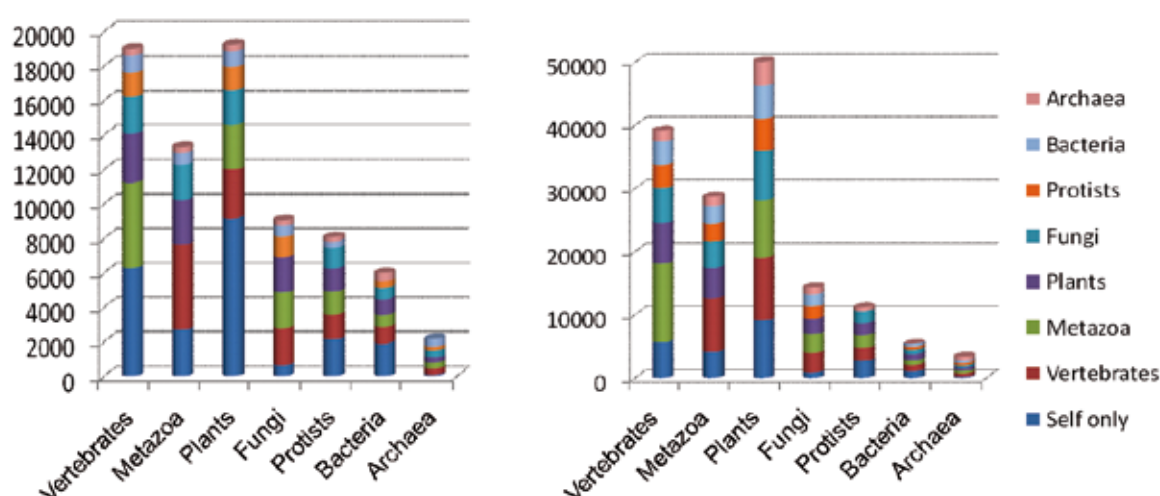


Figure 2. Pan-taxonomic comparative analysis in Ensembl Genomes. The charts show the number of protein clusters (left) and average cluster members per species (right) for seven major taxonomic divisions, such that the cluster contains either only members from the same division, or members from each of the other six divisions (the latter categories are exclusive of the first, but not of each other).

In a wider evolutionary context, all vertebrates are very closely related, and represent a tiny portion of the taxonomy. The launch of Ensembl Genomes has thus greatly widened the diversity of species available through the Ensembl system. At this stage, Ensembl Bacteria contains databases for six bacterial clades (*Bacillus*, *Escherichia/Shigella*, *Mycobacterium*, *Neisseria*, *Streptococcus* and *Staphylococcus*), and one archaeal clade (*Pyrococcus*), with each database containing the sequences of between four and 35 strains. Ensembl Protists currently focuses on the Apicomplexa species of human pathogens, including the causative agents of malaria (*Plasmodium falciparum* and other plasmodia). Ensembl Fungi contains data for the budding yeast, *Saccharomyces cerevisiae*, the fission yeast *Schizosaccharomyces pombe*, and eight species of the genus *Aspergillus*. Ensembl Plants includes the genomes of four dicotyledons: *Vitis vinifera*, *Populus trichocarpa*, and two species of *Arabidopsis*; and four monocotyledons: *Oryza sativa* groups *indica* and *japonica*, *Brachypodium distachyon*, and *Sorghum bicolor*. Ensembl Metazoa contains four vector genomes (*Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus* and *Ixodes scapularis*), five genomes of the *Caenorhabditis* genus, and twelve *Drosophila* genomes. A priority for the next year is the inclusion of additional species in all divisions, aiming at the representation of all well-studied model species within the system.

As an example of the benefits of this approach, the third release of Ensembl genomes contains ‘variation databases’ – databases which capture genome-wide occurrence of single nucleotide polymorphisms and indels from whole-population sampling – for four plant, one insect and one fungal species. The set for the thale cress *Arabidopsis* combines data from 17 recently sequenced strains (produced as part of a collaboration between ourselves, the Wellcome Trust Human Genetics Centre, and the University of Manchester) with earlier chip-based assays applied to a larger set of strains, and contains 2.9 million SNPs and 48 million individual locus-specific genotypes. The Ensembl variation infrastructure comes complete with a powerful interface for visualising these polymorphisms from the perspective of reference and individual genomes and thereby exposing these data to potential users. Other units of the system include modules for functional genomics and EST alignments. As we encounter new issues specific to certain domains of life, we adapt and extend the core Ensembl code base. For example, the internal database structure and some of the visualisation code have both been substantially modified to represent the different challenges of storing bacterial genomes, which (unlike vertebrates) are small, extremely numerous, and possess circular chromosomes.

VECTORBASE

Martin Hammond, Daniel Lawson, Karyn Megy

VectorBase is an NIAID Bioinformatics Resource Center (www.pathogenportal.org) focused on the genomes of invertebrate vectors of human pathogens. The group is responsible for the annotation and ongoing curation of a number of important vectors including the mosquitoes that transmit malaria (*Anopheles gambiae*), arboreal viruses such as Yellow fever and Dengue (*Aedes aegypti*), lymphatic filariasis (*Culex quinquefasciatus*) and the tick that transmits Lyme Disease (*Ixodes scapularis*).

VectorBase uses a modified Ensembl gene prediction pipeline which places greater emphasis on manually appraised- and community-submitted annotations above the computational gene predictions. In 2009 VectorBase has worked with the NIAID Microbial Sequencing Centers to finalise annotations for the body louse *Pediculus humanus* as well as performing updates for both *A. gambiae* and *A. aegypti* to capture community annotations for these genomes.

Currently, VectorBase is developing infrastructure for the capture of high-throughput transcriptomics and population genomic datasets in collaboration with partners in the community.

VectorBase related data is available from the VectorBase website (www.vectorbase.org) and Ensembl Genomes (www.ensemblgenomes.org). A central aim of the project is to ensure that genome annotations are submitted to EMBL-Bank on a regular basis and to maintain links with the UniProt protein database.

INTEGR8 AND GENOME REVIEWS

Matthias Haimel, Arnaud Kerhornou, Paul Kersey, Peter Sterk

Integr8 (Kersey *et al.*, 2005; www.ebi.ac.uk/integr8) is a portal for species with completely deciphered genomes. The portal holds data from numerous underlying resources integrated in a model reflecting the central dogma of biology, capturing the current state of knowledge about the genome of each organism, and providing a single site where resources are available for download and analysis. Genome Reviews is a related database providing standardised annotation of non-vertebrate genomes. Following the launch of Ensembl Genomes this year, we are now in the process of preparing to phase out Integr8 and Genome Reviews as separate brands. Data that currently appears through these projects will be made available through the Ensembl, Ensembl Genomes and UniProt interfaces.

IPI

Matthias Haimel

IPI (Kersey *et al.*, 2004; the International Protein Index, www.ebi.ac.uk/IPI) provides a top-level guide to the main databases that describe the proteomes of selected higher eukaryotic organisms. IPI effectively maintains a database of cross references between the primary data sources, provides minimally redundant yet maximally complete sets of proteins for featured species (one sequence per transcript), and maintains stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between releases. The database has been used in mass spectrometry experiments, where the use of a database with balanced trade-off between completeness and redundancy is essential to produce statistically valid protein identifications. However, the growing standardisation between Ensembl and UniProt for increasingly well-defined standard proteome sets has reduced the need for IPI, which will be phased out shortly (following the availability of complete protein sets in UniProt for mouse and rat).

ALTERNATIVE SPLICING AND TRANSCRIPT DIVERSITY (ASTD)

Gautier Koscielny

ASTD (www.ebi.ac.uk/astd) provides access to a vast collection of alternative transcripts that integrate transcription initiation, polyadenylation and splicing variant data together with extensive biological and expression information. Previous developments were reported in a recent consortium publication (Koscielny *et al.*, 2009). In 2009, developments were concentrated on the incorporation of splicing events and polyadenylation sites into the Ensembl pipeline framework. Intron retention, cassette exons, mutually exclusive, and alternative 5'/3' splicing events are available from the Ensembl core API interface and the Ensembl BioMart for *Homo sapiens* since release 55. Splicing events will be computed for other model organisms in the future.

The polyadenylation site computational pipeline is ready for ESTs/cDNAs datasets but is yet to be incorporated in Ensembl. Further developments are necessary to exploit the potential of next-generation sequencing whereby the transcriptome is sequenced in its entirety using short reads. A collaboration has been initiated with the Wellcome Trust Sanger Institute to develop a scoring system for polyadenylation sites in one model organism based on deep sequencing of 3' ends.

Finally, experimental developments have been carried out to display exon–exon junctions from 454/Solexa mRNASeq as tracks in Ensembl Genomes using the DAS technology.

GENOMICS STANDARDS CONSORTIUM

Peter Sterk

The Genomic Standards Consortium (GSC) is an initiative working toward richer descriptions of our collection of genomes and metagenomes. Established in September 2005, this international community includes representatives from the International Nucleotide Sequence Databases, major genome sequencing centres, bioinformatics centres, and a range of research institutions. The goal of the GSC is to promote mechanisms of standardising the description of (meta)genomes and the exchange and integration of (meta)genomic data. The ready availability of such data is of clear importance to Ensembl Genomes and the team has therefore played an active role within the GSC, organising GSC workshops and contributing to the development of the emerging standards. We organised the sixth GSC workshop in October 2008 and a satellite meeting concerned with the technical development of GCDML (Genomic Context Data Mark-up Language), which is being developed to capture this data.

FUTURE PROJECTS AND GOALS

We are aiming to increase the coverage of Ensembl Genomes to cover all important model species. Species scheduled for inclusion in the first wave include the red bread mould *Neurospora crassa*, the slime mould *Dictyostelium discoideus* and the body louse *Pediculus humanus*. Further species will be added throughout the year.

Two EC funded projects are due to start shortly: INFRAVEC, a 30 partner project to produce a research infrastructure for the genetic control of mosquitoes, and Microme, a 13 partner project to produce a new resource for bacterial metabolic pathways. INFRAVEC is one of many projects (alongside others in every domain of Ensembl Genomes) in which we will be ramping up the use of the Ensembl variation infrastructure as population-wide resequencing becomes increasingly common in all species. Microme will be hosted at the EBI and its development offers the potential to improve the annotation of bacterial genomes in the context of our knowledge of metabolism.

As the number of sequenced genomes increases, we are reworking our comparative genomics pipelines to ensure the scalability of the analysis and to integrate the presentation of analyses with different taxonomic scope. The main Ensembl code base was developed for eukaryotic genomes and we are working on improving the representation of genome, gene structure and variation to provide a better fit to the biology of bacterial species.

Team Members**Ensembl Metazoa Coordinator**

Daniel Lawson

Ensembl Genomes Software Coordinator

Daniel Staines

Senior Software Engineer

Alan Horne*

Gautier Koscielny

Peter Sterk*

Software Engineer

Matthias Haimel

Arnaud Kerhornou

Devang Lakhani*

Bioinformatician

Paul Derwent*

Uma Maheswari*

Michael Nuhn*

Scientific Programmer

Karyn Megy

VectorBase Bioinformatician

Martin Hammond*

User Support Officer

Jeff Almeida-King*

*Indicates part of the year only

Publications**2008**

Kersey, P., *et al.* (2008). Building a biological space based on protein sequence similarities and biological ontologies. *Comb. Chem. High Throughput Screen.*, 11, 653-660

2009

Hubbard, T.J.P., *et al.* (2009). Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697

Koscielny, G., *et al.* (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93, 213-220

Lawson, D., *et al.* (2009). VectorBase: a data resource for invertebrate vector genomics.

Nucleic Acids Res., 37, D583-587

Megy, K., *et al.* (2009). Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. *Infection, Genetics and Evolution*, 9, 308-313

Sieglaff, D.H., *et al.* (2009). Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc. Natl Acad. Sci. USA.*, 106, 3053-3058

Other EMBL publications

Kersey, P., *et al.* (2004). The International Protein Index: An integrated database for proteomics experiments. *Proteomics*, 4, 1985-1988

Kersey, P., *et al.* (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, 33, D297-D302

Henning Hermjakob

*Dipl. Inf (MSc.) in bioinformatics, 1996, University of Bielefeld.
Research assistant at the National Research Centre for Biotechnology (GBF),
Braunschweig, in the Transfac Database team.
At EMBL-EBI since 1997.*



The Proteomics Services Team

51

INTRODUCTION

The Proteomics Services team develops tools and resources for the representation, deposition, distribution and analysis of proteomics and proteomics-related data. The team is a major contributor to the Proteomics Standards Initiative (PSI; www.psidev.info) of the international Human Proteome Organization (HUPO). We provide reference implementations for the PSI community standards, in particular the PRIDE protein identification database (www.ebi.ac.uk/pride) and the IntAct molecular interaction database (www.ebi.ac.uk/intact). On the next level of abstraction, we provide the Reactome database of pathways (www.reactome.org) in collaboration with New York University (NYU) and the Ontario Institute for Cancer Research (OICR).

As a result of long-term engagement with the proteomics community, journal editors, and funding organisations, proteomics data deposition in PSI-compliant data resources such as IntAct and PRIDE is increasingly becoming a strongly recommended part of the publishing process. Accordingly, this has resulted in a rapid increase in the data content of our resources.

The Proteomics curation teams ensure consistency and appropriate annotation of all data; whether from direct depositions or literature curation, to provide the community with high-quality reference datasets.

Across a range of European projects (Apo-Sys, LipidomicNet, SLING, ENFIN, and ProteomeBinders) we contribute to the development of data integration technologies using the Distributed Annotation System (DAS) and web services. In particular, the successful Ontology Lookup Service (OLS; www.ebi.ac.uk/ols), Protein Identifier Cross-Reference Service (PICR; www.ebi.ac.uk/Tools/picr) and the DASTY DAS client (www.ebi.ac.uk/dasty) are under constant evolution and further development.

The Proteomics Services team follows an open source, open data approach; all resources we develop are freely available.

PROTEOMICS STANDARDS INITIATIVE

Bruno Aranda, David Gloriam, Samuel Kerrien, Lennart Martens, Sandra Orchard, Juan Antonio Vizcaino

Proteomics data are still highly fragmented; many datasets are not available in the public domain, or are only available in different and largely incompatible formats spread over database, author and journal websites. The PSI, a HUPO work group, aims to standardise the representation and annotation of proteomics data and to promote the systematic collection of proteomics data in publicly accessible databases (Orchard *et al.*, 2008). The PSI has several work groups, currently focusing on molecular interactions, mass spectrometry, protein modifications and protein separations. The deliverables of each work group are

- minimum information guidelines: in analogy to the MIAME guidelines for DNA microarray experiments, 'Minimum Information About a Proteomics Experiment' (MIAPE) documents were developed to define the data items that should be minimally reported about a proteomics experiment in order to allow independent critical assessment. The MIAPE guidelines consist of a general 'parent document' (Taylor *et al.*, 2007) and work group-specific modules. So far, modules for molecular interactions (MIMIX; Orchard *et al.*, 2007), mass spectrometry (Taylor *et al.*, 2008a), mass spectrometry informatics (Binz *et al.*, 2008), and gel electrophoresis (Gibson *et al.*, 2008) have been released;

- data exchange formats: to facilitate data management and exchange, the PSI develops data exchange formats for proteomics. For each work group/domain, these should minimally represent the data items specified in the MIAPE guidelines, but also allow a much more detailed representation. In August 2009, PSI-PAR, a community standard format for the representation of protein affinity reagents was released (Gloriam *et al.*, 2009);
- controlled vocabularies: while XML schemata provide a syntax for data exchange, they do not specify the semantics of data elements exchanged. As an example, the yeast two-hybrid technology might be designated by many different terms, most of which are sufficiently distinct to make automatic recognition impossible. Thus, the PSI either references external controlled vocabularies and ontologies such as the Gene Ontology where possible, or develops its own controlled vocabulary where necessary. The combination of reasonably stable XML schemata and regularly maintained controlled vocabularies allows quick adaptation to new terms and technologies, while providing the stability required for database and software development;
- databases and tools: while the PSI develops community standards for proteomics, their implementation is usually promoted by the individual member organisations, for example, applying the standards to data submitted and contained in the IntAct and PRIDE databases;
- data capture and data exchange: the ultimate aim of the PSI is to make proteomics data more easily accessible in the public domain. To this end, PSI initiates regular data exchange between major databases, similar to the established mechanisms for nucleotide sequence data and macromolecular structures. Initiatives for regular exchange of molecular interaction data (IMEx; imex.sf.net) and protein identification data (ProteomExchange) are currently in the implementation phase. In March 2009, the EU PSIMEx grant started. This grant, co-ordinated by the Proteomics Services team, unites 14 partners from Europe, North America, and Asia, with a common interest in molecular interaction data annotation, dissemination and analysis. A broad spectrum of participants, from instrument providers (BiaCore [GE Healthcare] to journals (Proteomics, Nature Biotechnology) and repository providers (IntAct, MINT, DIP, BioGrid, MatrixDB) is expected to ensure strong community interaction and high consortium impact.

The major PSI event is the annual PSI spring meeting which was held in Turku, Finland, in 2009 (Orchard *et al.*, 2009a, Orchard *et al.*, 2009b), and will take place in Seoul, Korea, in March 2010. The PSI is an open, collaborative initiative.

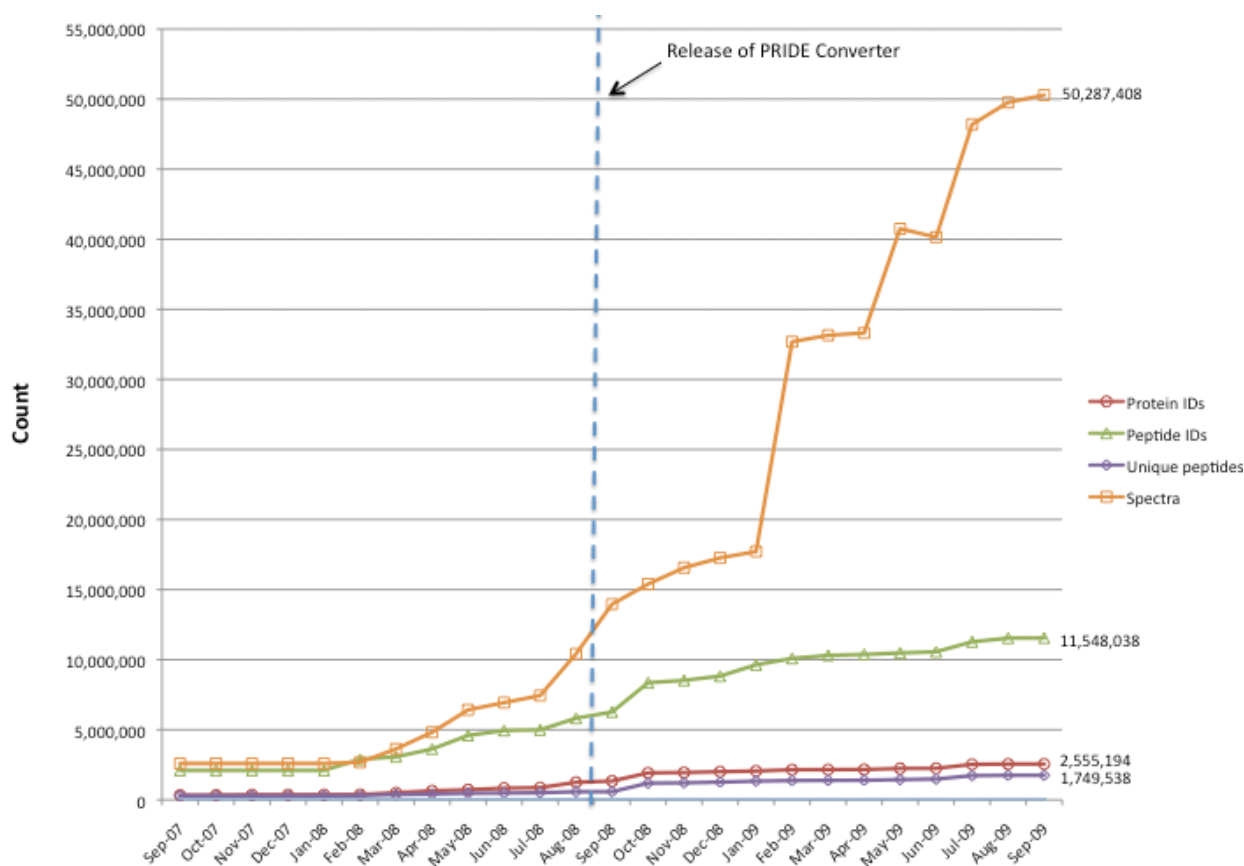


Figure 1. PRIDE data content September 2007 – September 2009.

We invite comments and participation in new and existing work groups. Full project information is available from the PSI website www.psidev.info.

In the context of the HUPO PSI, we also contribute to related international efforts, such as ‘Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles’ (Rodriguez *et al.*, 2008).

MOLECULAR INTERACTIONS

Premanand Achuthan, Bruno Aranda, Samuel Kerrien, Jules Kersemakers, Jyoti Khadake, Sandra Orchard, Avazeh Taskakkori, David Thorneycroft, Kristel van Eijk

As a framework for the formal representation of molecular interaction data, the PSI MI 2.5 format has been published (Kerrien *et al.*, 2007), extending the scope of the format from protein–protein interactions to general molecular interactions, for example between proteins and ligands. These updates have been successfully implemented in the IntAct database, extending the scope of the IntAct platform to new domains, for example drug–target interactions.

A major redevelopment of the IntAct web interface (www.ebi.ac.uk/intact) has provided fast access and complex search options to identify protein interactions of interest, for example based on annotation with controlled vocabularies.

As the result of an ESF-funded workshop ‘Development of standards-compliant tools for molecular interaction data management’ held on 16–19 November 2008, we have now defined PSICQUIC, the PSI Common Query Interface. This web service interface is already implemented by BioGrid, IntAct, MINT, MPIDB, MatrixDB and iRefIndex, providing a common interface to multiple independent molecular interaction data resources.

As journals increasingly encourage authors to deposit data in public databases, but also as a result of communication with key experimentalists, the IntAct database has seen a marked increase in direct data depositions. Overall, IntAct now contains more than 200,000 molecular interactions from direct data depositions, literature curation focusing on specific journals and literature curation focusing on specific topics such as cancer, chromatin or *Arabidopsis* (Morsy *et al.*, 2008).

PROTEIN IDENTIFICATIONS

Richard Côté, Lennart Martens, Jonathan Rameseder, Florian Reisinger, Juan Antonio Vizcaíno

The PRIDE database (www.ebi.ac.uk/pride) has strengthened its position as one of the major global repositories for proteomics data. The *Proteomics* journal’s instructions to authors now mandate deposition of proteomics datasets in PRIDE or a comparable database. PRIDE data content has increased five-fold, from 10.5 million spectra in August 2008 to almost 50 million spectra in August 2009 (figure 1). Data content is not only increasing in quantity, but also in terms of observations of low abundance proteins, as exemplified in figure 2. A major milestone in our capacity to handle data depositions was the introduction of the PRIDE converter (Barsnes *et al.*, 2009), a user-friendly tool for

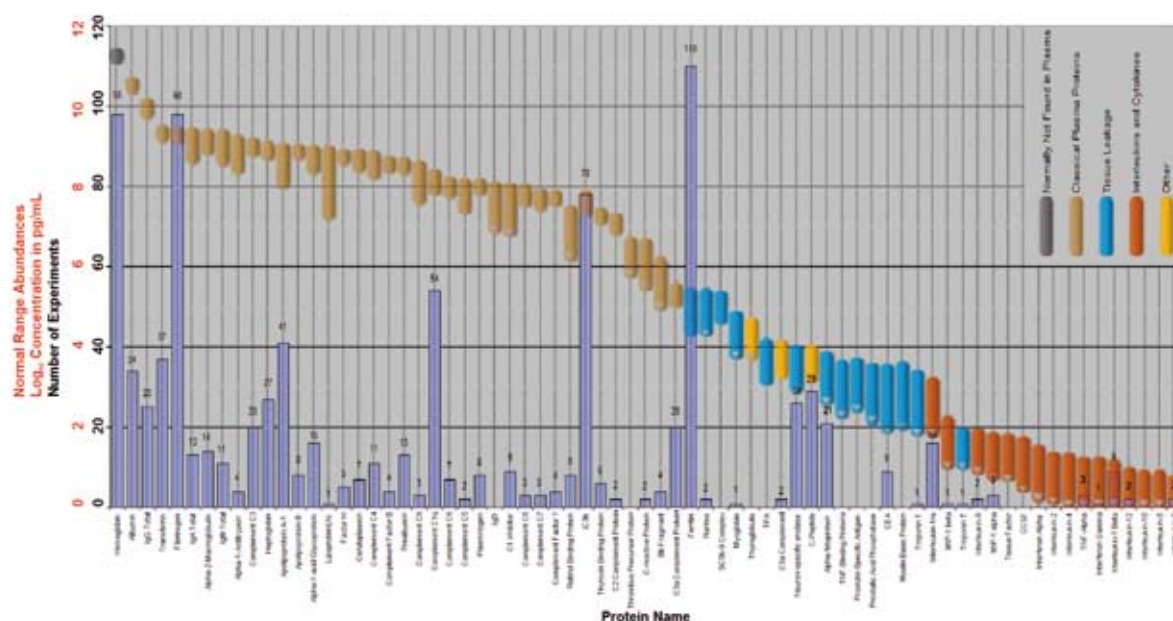


Figure 2. Adapted from Anderson & Anderson, 2002. The figure shows the abundance of 70 protein analytes in human plasma, their abundance spanning twelve orders of magnitude (original Anderson & Anderson data), and the number of PRIDE experiments in which the corresponding protein has been observed.

the preparation of PRIDE data depositions. For further facilitation of data deposition in PRIDE, we have published a tutorial-style guide to using PRIDE (Vizcaino *et al.*, 2009).

The masters' thesis 'Computational Interpretation of Tandem MS Spectra Including Fragment Ions with Post-Translational Modifications' of Jonathan Rameseder, written during a traineeship in the PRIDE team in 2008, was selected by the Austrian Computer Society (Oesterreichische Computergesellschaft, OCG) for the 'best master thesis award' (OCG Foerderpreis FH 2009).

PATHWAYS

David Croft, Bernard de Bono, Phani Garapati, Bijay Jassal, Steven Jupe, Gavin O'Kelly, Esther Schmidt

Reactome (www.reactome.org) is an expert-authored, peer-reviewed knowledgebase of human reactions and pathways that functions as a data mining resource and electronic textbook. The basic information in Reactome is provided by bench biologists who are experts on a particular pathway. The information is then managed by groups of curators at EMBL-EBI, NYU and OICR, peer-reviewed by other researchers and published on the web.

Reactome coverage ranges from the basic processes of metabolism to complex regulatory pathways such as haemostasis. The current release includes 3,916 human proteins, 3,541 reactions and 5,571 literature citations. This represents a coverage of approximately 19.3% of 20,334 curated UniProtKB human proteins, a 2.6-fold increase over the last three years. Linking to external datasets has been improved for better maintainability, and Reactome-derived protein-protein interaction data are now available in the PSI-MITAB standard format. In response to suggestions by the Reactome Scientific Advisory Board, a new website prototype has been developed, focusing on entity-level visualisation, a pathway analysis portal with functionalities for expression and interaction data overlay, and increased rendering speed. Further customisation will facilitate the creation of new Reactome instances for model organisms, in collaboration with individual communities.

COMPUTATIONAL PHYSIOLOGY

Bernard de Bono

The VPH Network of Excellence (www.vph-noe.eu/) is an FP7 project that started in June 2008. It is currently focusing on building an interoperability framework that links physiology-related data and model resources, particularly by fostering the adoption and demonstration of standards in the areas of mark-up languages, ontologies and minimal information for representation. The successful RICORDO (VPH STREP) proposal (co-ordinated by EMBL-EBI with the project due to start in February 2010), will further develop, implement and demonstrate the effectiveness of such standards for ordinary differential equation (ODE) models of physiology and radiological models of anatomy. Furthermore, it aims to integrate models and ontologies related to medical physiology and human anatomy in collaboration with the OBO Foundry. The RICORDO infrastructure is already attracting collaborative links with modellers in the pharma domain (through the EBI's involvement in the Innovative Medicines Initiative project), and with groups that focus on the development of clinical terminologies.

DATA INTEGRATION

Richard Côté, Rafael Jimenez, Omar Pera Mira, Florian Reisinger

The Proteomics Services team has a strong presence in data integration projects, triggered by the diversity of proteomics data handled within the team. As part of our involvement in the EU-funded ENFIN, Apo-Sys and SLING projects, we contribute to the development and application of data integration infrastructure based on DAS and XML technologies.

Based on the Dasty protein DAS client (www.ebi.ac.uk/dasty; Jimenez *et al.*, 2008), we have developed OntoDas – a tool for facilitating the construction of complex queries to the Gene Ontology (O'Neill, 2008).

The Protein Identifier Cross-Referencing Service (PICR; www.ebi.ac.uk/Tools/picr; Côté *et al.*, 2007), translates between protein identifier namespaces and thus facilitates the joint analysis of protein datasets from multiple sources, for example protein identifications performed against different databases.

The Ontology Lookup Service (OLS; Côté *et al.*, 2008) provides a unified interface to currently 61 ontologies in OBO format, facilitating management of ontology data across multiple projects within and beyond the Proteomics Services team.

FUTURE PROJECTS AND GOALS

In 2007, our molecular interactions activities resulted in a substantial set of published manuscripts, from the MIMIx guidelines via the PSI MI 2.5 format to the standard implementation in the IntAct database. In 2008, a similar breakthrough was achieved in the domain of protein identifications, with three published MIAPE modules and the release

of the mzML format for mass spectrometry data representation. In 2009, this prior work was recognised by the award of the PSIMEx grant, which for the first time specifically funds our international integration activities. For 2010, we plan to build on these strengths, and expect to initiate production mode for regular international molecular interaction data exchange.

We also plan to intensify data integration within and beyond the projects of the Proteomics Services team, in particular in the context of the EnVision platform and DAS. We will also provide closer integration between Reactome pathways and IntAct molecular interactions, by moving current prototypes into production.

Finally, we will continue our successful collaboration with all PSI partners, in particular with journals and editors, to encourage data producers to make their data available to the community through public databases by utilising community-supported standards.

Team Members

Coordinator

Lennart Martens
Sandra Orchard
Esther Schmidt

Senior Software Engineer

Richard Côté
Phil Jones
Samuel Kerrien

Scientific Database Curator

Bernard de Bono
Phani Garapati
Bijay Jassal
Jyoti Khadake
David Thorneycroft
Steven Jupe*

Software Engineer

Premanad Achuthan
Bruno Aranda
Antony Quinn*
Florian Reisinger
Rafael Jimenez

Bioinformatician

David Croft
Juan Antonio Vizcaino
Gavin O'Kelly*

Visitor

Matthieu Visser

Trainees

Omar Pera Mira*
Kieran O'Neill*
Jonathan Rameseder*
Laurence Newman*
Jules Kersemakers*
Avazeh Taskakori*
Kristel van Eijk*

* Indicates part of the year only.

Publications

2008

Chatr-Aryamontri, A., *et al.* (2008). MINT and IntAct contribute to the Second BioCreative challenge: Serving the text-mining community with high quality molecular interaction data. *Genome Biol.*, 9, Suppl 2, article S5

Eisenacher, M., *et al.* (2008). Proteomics data collection – 3rd ProDaC Workshop: April 22nd 2008, Toledo, Spain. *Proteomics*, 8, 4163-4167

Helsens, K., *et al.* (2008). Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol. Cell. Proteomics*, 7, 2364-2372

Jimenez, R.C., *et al.* (2008). Dasty2, an Ajax protein DAS client. *Bioinformatics*, 24, 2119-2121

O'Neill, K., *et al.* (2008). OntoDas – A tool for facilitating the construction of complex queries to the Gene Ontology. *BMC Bioinformatics*, 9, 437

Orchard, S., *et al.* (2008). Annual Spring Meeting of the Proteomics Standards Initiative 23-25 April 2008, Toledo, Spain. *Proteomics*, 8, 4168-4172

Reeves, G.A., *et al.* (2008). The protein feature ontology: A tool for the unification of protein feature annotations. *Bioinformatics*, 24, 2767-2772

2009

Barsnes, H., *et al.* (2009). OMSSA Parser: An open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics*, 9, 3772-3774

Eisenacher, M., *et al.* (2009). Proteomics Data Collection -4th ProDaC workshop 15 August 2008, Amsterdam, the Netherlands. *Proteomics*, 9, 218-222

Eisenacher, M., *et al.* (2009). Proteomics Data Collection – 5th ProDaC Workshop 4 March 2009, Kolymari, Crete, Greece. *Proteomics*, 9, 3626-3629

Eisenacher, M., *et al.* (2009). Getting a grip on proteomics data – Proteomics Data Collection (ProDaC). *Proteomics*, 9, 3928-3933

Kathiresan, T., *et al.* (2009). A protein interaction network for the large conductance Ca²⁺-activated K⁺ channel in the mouse cochlea. *Mol. Cell. Proteomics*, 8, 1972-1987.

Martens, L. & Apweiler, R. (2009). Algorithms and databases. *Methods Mol. Biol.*, 564, 245-259

Matthews, L., *et al.* (2009). Reactome

knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37, D619-D622

Montecchi-Palazzi, L., *et al.* (2009). The PSI semantic validator: a framework to check minimum information about a proteomics experiment compliance of proteomics data. *Proteomics*, in press

O'Connor, M.N., *et al.* (2009). Functional genomics in zebrafish permits rapid characterization of novel platelet membrane proteins. *Blood*, 113, 4754-4762

Orchard, S. (2009). Ending the "publish and vanish" culture: How the data standardization process will assist in data harvesting. *J. Proteome Res.*, 8, 3219

Orchard, S., *et al.* (2009a). Second Joint HUPO publication and PSI Workshop 24th April 2009, Turku, Finland. *Proteomics*, 9, 4426-4428

Orchard, S., *et al.* (2009b). Annual Spring Meeting of the Proteomics Standards Initiative, 27-29 April 2009, Turku, Finland. *Proteomics*, 9, 4429-4432

Orchard, S. & Taylor, C.F. (2009). Debunking minimum information myths: one hat need not fit all. *N. Biotechnol.*, 25, 171-172

Persson, B., *et al.* (2009). The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.*, 178, 94-98

Rodriguez, H., *et al.* (2009). Recommendations from the 2008 International Summit on proteomics data release and sharing policy: The Amsterdam principles. *J. Proteome Res.*, 8, 3689-3692

Steinbeck, C., *et al.* (2009). New open drug activity data at EBI. *Chem. Cent. J.*, 3, 62

Vizcaino, J.A., *et al.* (2009). A guide to the PRIDE proteomics data repository. *Proteomics*, 9 4276-4283

cont.

Other EMBL publications

Côté, R.G., *et al.* (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, 8, 401

Côté, R.G., *et al.* (2008). The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, 36, W372-376

Kerrien, S., *et al.* (2007). Broadening the horizon - Level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, 5, 44

Morsy, M., *et al.* (2008). Charting plant interactomes: possibilities and challenges. *Trends Plant Sci.*, 13, 183-191

Orchard, S., *et al.* (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, 25, 894-898

Taylor, C.F., *et al.* (2007). The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.*, 25, 887-893

Taylor, C.F., *et al.* (2008). Guidelines for reporting the use of mass spectrometry in proteomics. *Nat. Biotechnol.*, 26, 860-861

Other publications

Binz, P.A., *et al.* (2008). Guidelines for reporting the use of mass spectrometry informatics in proteomics. *Nat. Biotechnol.*, 26, 862

Gibson, F., *et al.* (2008). Guidelines for reporting the use of gel electrophoresis in proteomics. *Nat. Biotechnol.*, 26, 863-864

Gloriam, D.E. *et al.*, (2009). Report: A community standard format for the representation of protein affinity reagents. *Mol. Cell. Proteomics*. [Epub ahead of print] PMID: 19674966

Rodriguez, H. *et al.* (2009). Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles. *J. Proteome Res.*, 8, 3689-369



Sarah Hunter

*MSc. 1999, University of Manchester.
At EMBL-EBI since 2005.*

The InterPro Team

57

INTRODUCTION

The InterPro team coordinates the InterPro and CluSTr projects at EMBL-EBI.

InterPro is an integrated documentation resource for protein families, domains and functional sites. The project integrates signatures from the major protein signature databases into a single resource, and currently includes data from Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, CATH-Gene3D, PANTHER and HAMAP.

During the integration process, InterPro rationalises where more than one protein signature describes the same protein family/domain, and unites these into single InterPro entries, with relationships between them where applicable. Additional biological annotation is included, together with links to external databases such as GO, PDB, SCOP and CATH. InterPro precomputes all matches of its signatures to UniProt Archive (UniParc) proteins using the InterProScan software, and displays the matches to the UniProt KnowledgeBase (UniProtKB) in various formats.

InterPro has a number of important applications, including the automatic annotation of proteins for UniProtKB/TrEMBL and genome annotation projects. InterPro is used by Ensembl and in the GOA project to provide large-scale mapping of proteins to GO terms.

The CluSTr project aims to cluster all UniProtKB proteins and protein sets from complete genomes. The resulting clusters and similarity scores are accessible via a web interface. It also provides best reciprocal hit orthologue data for a number of complete genomes.

INTERPRO

The InterPro project (Hunter *et al.*, 2009; www.ebi.ac.uk/interpro) aims to provide an integrated resource for protein families, domains and functional sites. InterPro includes data from eleven member databases and continues to grow along with its members. The resource continues to provide up-to-date data and new features, and thus increases its use to the scientific community as a powerful protein classification tool. It is not only useful to bench scientists, but also to large genome sequencing projects.

Functional annotation of proteins by automatic means is vital in the post-genomic era because vast quantities of uncharacterised protein sequences are flooding into the protein sequence databases. There are many new protein families to be integrated from the newest member databases, and constant updates from the older members. As the number of protein signatures increases, so does the coverage of UniProtKB and UniParc. Protein signatures are useful tools for prediction of protein function and many important protein signature databases have been developed. However, the diversity of their methods and foci makes it difficult for a user to discern which one to use. InterPro has solved this problem by integrating the signatures from all the well-known databases into a single coherent resource.

During the integration process, signatures from the different databases that describe the same protein family, domain, repeat or functional site are integrated into a single InterPro entry with a unique InterPro accession number. When a signature matches a subset of a larger group of proteins, matched by a different but overlapping signature, it is assigned a unique InterPro accession number and the entries are then related to each other. There are presently three types of relationships in InterPro: 'parent/child', which shows family relationships, 'contains/found in', which displays domain composition and 'overlapping' which is a catch-all for other relationships which do not fall into the former

two categories. New InterPro entries are annotated with a name, short name, abstract, references and cross-links to related databases. Where possible, InterPro entries are mapped to Gene Ontology (GO) terms. They are also populated with all the UniProtKB proteins that have matches to the signature(s) in the entry. These matches can be viewed in a number of different formats, including a table view, a graphical overview and detailed view, and a domain architectures view. Due to the large numbers of signatures now available in the InterPro member databases, it is currently not possible to manually integrate all of them into InterPro entries. To circumvent this problem, pages have been added to the web interface which display these un-integrated signatures and the proteins they match, together with minimal annotation, such as a name.

Protein 3D structure information is integrated into InterPro through two different approaches: 1) links to curated PDB, SCOP and CATH structural classes, and 2) through SUPERFAMILY and Gene3D Hidden Markov Models (protein signatures) to predict which proteins belong to the structural classes. Where solved structures are not available, links to SwissModel and ModBase predicted structures (from homology modelling) are provided.

InterProScan

The protein matches in InterPro are calculated using InterProScan (Quevillon *et al.*, 2005), which integrates the scanning algorithms from the member databases into a single tool. Both DNA and protein sequences can be submitted to InterProScan; DNA sequences are first translated and then all possible open reading frames are scanned. InterProScan results from the web server are displayed graphically and also in a table view, together with additional information such as GO terms and entry-to-entry relationships for matched entries. A stand-alone version of InterProScan is available to download for users who require privacy or bulk searches; users can also submit searches programmatically via web services. The latest version of InterProScan, 4.5, was released in July 2009.

InterPro database

The number of entries and coverage of proteins by InterPro continues to grow (figure 1). The latest release of the database (23.0) contains 19,150 entries. In its infancy, InterPro covered around 66% of all proteins in UniProtKB, and this has increased to 96.2% for UniProtKB/Swiss-Prot, 78.5% for UniProtKB/TrEMBL, and 79.5% overall for UniProtKB (Swiss-Prot and TrEMBL). In the last year InterPro has changed its update frequency and aims to release the updated database every six weeks, in tandem with the UniProtKB release cycle. This has led to an implicit alteration to InterPro's version numbering, with a major release (e.g. '23.0') occurring when a member database version has been updated and a minor release (e.g. '23.1') occurring when only the underlying proteins in the database have been updated. A consequence of the change in the release policy is that the proteins in InterPro and UniProtKB are much more tightly synchronised than before. InterPro released versions 19.0, 20.0, 21.0, 22.0 and 23.0 during 2009.

In the past year, InterPro has included a new member database – HAMAP (High quality Automated and Manual Annotation of microbial Proteomes), which uses profiles (weight matrices) to characterise mainly bacterial and archaeal proteins over their entire length. HAMAP is used by the UniProtKB team as a method for annotating proteins in a more high-throughput manner than manual inspection alone. The methodology used by the PROSITE database has also been upgraded so that the pattern portion of the database now uses evaluative mini-profiles to check the validity of a match to a pattern. The member databases PRINTS, Pfam, TIGRFAMs, ProDom and SMART have all been updated to their latest versions.

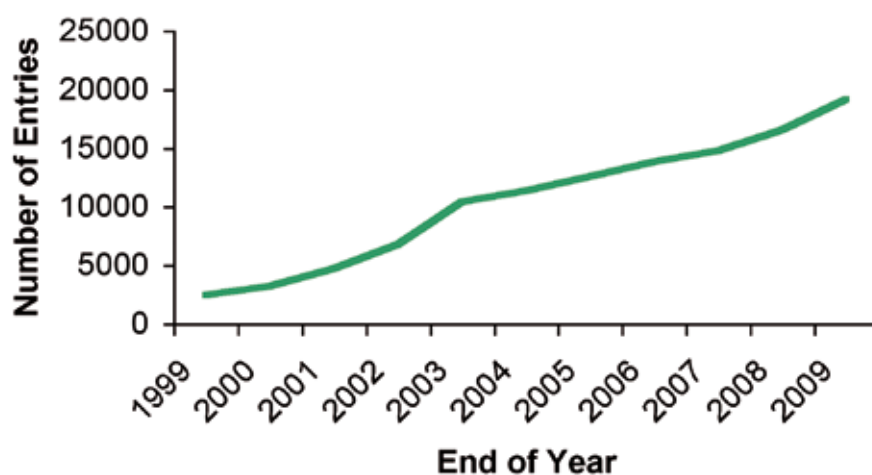


Figure 1. Growth of InterPro in number of entries since 1999.

InterPro data is now also available to users via a BioMart (Smedley *et al.*, 2009) which is linked to the pathway database Reactome and the proteomics database PRIDE. It allows users to create complex queries against InterPro data and download the results in a variety of formats. The BioMart is accessible both via an interface and programmatically.

Training

The InterPro team has been involved in various user training events over the past year, in which lectures and tutorials on InterPro were provided:

- July 2009: Stockholm, Sweden. ISMB/ECCB 2009
- June 2009: VIB Research Institute, Ghent, Belgium. Bioinformatics Roadshow
- May 2009: University of Santiago, Spain. Bioinformatics Roadshow
- April 2009: EBI hands-on training: Programmatic Access to Biological Databases
- April 2009: Poitiers, France. Bioinformatics Roadshow – EBI tools
- February 2009: EBI hands-on training: EMBRACE/EBI workshop: Protein Structure
- February 2009: Boston, USA. Patent Information Users Group Biotechnology Meeting
- February 2009: Harvard, Boston. USA. Bioinformatics Roadshow
- January 2009: Brussels, Belgium. BioSapiens European School of Bioinformatics
- December 2008: EBI hands-on training: Joint EBI-Wellcome Trust Proteomics workshop
- November 2008: EBI hands-on training: Web Service Access using Java
- November 2008: EBI. Masters Open Day
- October 2008: EBI hands-on training: A Dip into EBI Resources

CLUSTR

The CluSTr database (Petryszak *et al.*, 2005; www.ebi.ac.uk/clustr) offers an automatic classification of UniProtKB proteins into groups of related proteins. The clustering is based on analysis of all pairwise comparisons between protein sequences using the Smith–Waterman algorithm. Statistical significance of each similarity is then estimated by Z-value derived from its Smith–Waterman score, as well as an arithmetic mean and standard deviation of Smith–Waterman scores of similarities between the two proteins in question and all the proteins encountered so far.

Analysis carried out at different levels of protein similarity yields a hierarchical organisation of clusters. Working with clusters at different levels of similarity allows biologically meaningful clusters to be selected for different groups of proteins, which greatly increases the flexibility of the database.

A classification of CluSTr-derived protein families using GO terms (via InterPro-to-GO Mapping, <ftp://ftp.ebi.ac.uk/pub/databases/clustr/clustr2go/clustr2go.gz>) is regularly produced. Mapping to GO is now served as part of the CluSTr web service and currently maps just over 1.5 million clusters to GO terms.

CluSTr data and its derivations are available not only through the CluSTr web service, but also through a number of other EMBL-EBI services:

- UniProt (www.uniprot.org/): mapping UniProtKB proteins to clusters to which they belong;
- CluSTr search facility (www.ebi.ac.uk/clustr);
- InterPro (www.ebi.ac.uk/interpro/).

Links from CluSTr to the InterPro detailed graphical interface allow users to see whether proteins from a cluster share the same protein matches. Analysis of a cluster's domain composition is even more apparent with the InterPro Architectures view, which shows a single representative for proteins with exactly the same domain architecture. InterPro now provides reciprocal links from InterPro entries back to those CluSTr clusters that overlap these entries to a sufficient degree.

Currently CluSTr contains the following information:

- 9,450,285 sequences from UniProt Knowledgebase release 15.6;
- 303,139 sequences from IPI;
- 3.6 billion similarities, with pairwise alignments generated on-the-fly;
- 17,616,060 clusters;

- clustering for 972 organisms with completely sequenced genomes, with putative homologue predictions for these species.

The CluSTr web interface includes a visualisation tool, which facilitates the traversal of clustering hierarchies and graphically represents InterPro and GO annotation of individual clusters. CluSTr continues to deliver best reciprocal hit orthology predictions for all species in CluSTr.

FUTURE PROJECTS AND GOALS

We are currently planning an overhaul of the InterPro web interface and web services so that more users will be able to easily access and interpret our data. Our intention is to allow for more complex querying and more navigable web pages; we also intend to provide more data via REST and SOAP-based web services. We are rewriting the InterProScan software package to improve its flexibility and modularity and bring it in line with our internal production pipelines.

Team Members

Coordinator

David Lonsdale (Annotation Coordinator)

Senior Scientific Database Curator

Jennifer McDowall

Scientific Database Curator

Louise Daugherty

Bioinformaticians

Siew-Yit Yong
Craig McAnulla

Senior Software Engineers

Anthony Quinn
John Maslen
Manjula Thimma
Phil Jones

Software Engineers

David Binns
Ujjwal Das

Team Secretary

Kerry Smith

Publications

2008

Jimenez, R.C., *et al.* (2008). Dasty2, an Ajax protein DAS client. *Bioinformatics*, 24, 2119-2121

2009

Barrell, D., *et al.* (2009). The GOA database in 2009 – An integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, 37, D396-D403

Hunter, S., *et al.* (2009). InterPro: The integrative protein signature database. *Nucleic Acids Res.*, 37, D211-D215

Smedley, D., *et al.*, (2009). BioMart - biological queries made easy. *BMC Genomics*, 10, article 22

Other EMBL publications

Petryszak, R., *et al.* (2005). The predictive power of the CluSTr database. *Bioinformatics*, 21, 3604-3609

Quevillon, E., *et al.* (2005). InterProScan: protein domains identifier. *Nucleic Acids Res.*, 33, W116-W120

John Overington

PhD Crystallography, Birkbeck College, London.
Postdoctoral research, ICRF 1990-1992.
Pfizer 1992-2000.
Inpharmatica 2000-2008.
At EMBL-EBI since 2008.



Computational Chemical Biology: the ChEMBL Team

61

INTRODUCTION

This has been our first year as a new group at EMBL-EBI, and as such, much of the year has been spent recruiting, training and reengineering systems and data acquired from Inpharmatica/Galapagos. The opportunity to reassess some design issues of *ad hoc* developed legacy systems, has allowed the streamlining of several key steps, and also has resulted in significant efficiencies to be gained through working with key partner groups (in particular the Chemoinformatics and Metabolism team, see page 65).

The service and research activities of the group centre around the area of computational chemical biology (or chemogenomics), with specific focus on the discovery of chemical probes for biological systems, leading to an understanding of normal and disease biology (figure 1). A major application of such approaches is in the prioritisation and lead discovery of drug-like molecules for use as innovative therapies.

The group is funded by a five-year Wellcome Trust Strategic Award, and tasked to develop publicly available databases relevant to drug discovery, specifically large-scale structure activity relationship (SAR) data. Ongoing data entry and curation has been maintained during the transfer of the resources, and the total size of the ChEMBL database, (chem-

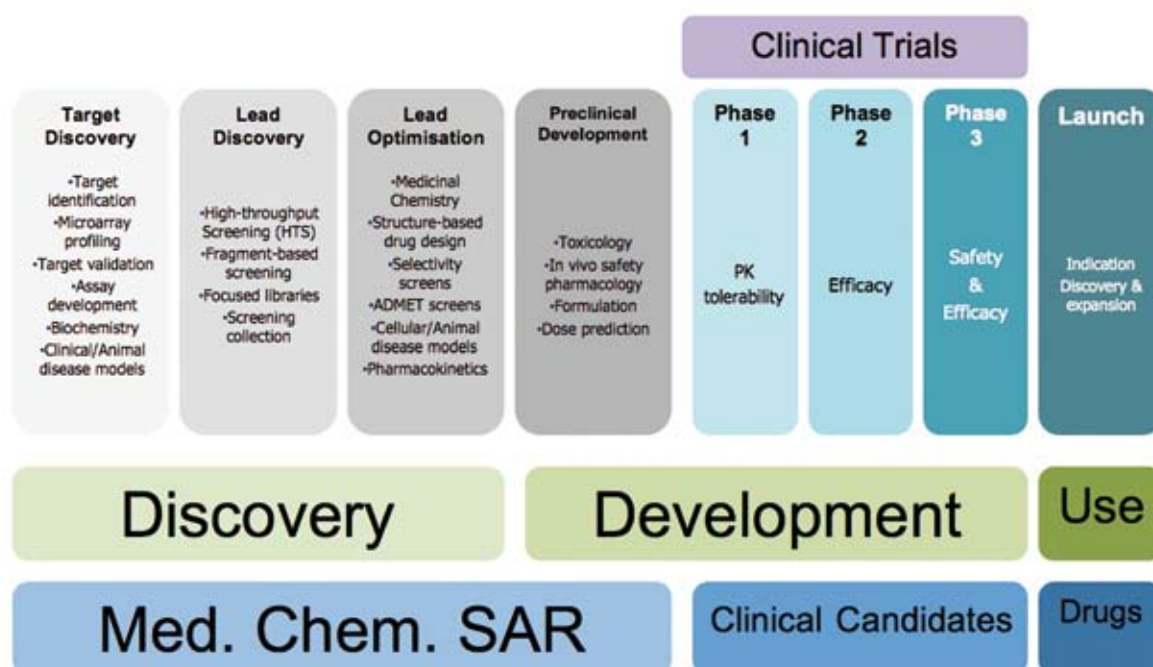


Figure 1. Logical architecture and mapping of ChEMBL resources onto the drug discovery process.

bldb) is currently in excess of 620,000 compound records, covering over 3,600 distinct protein targets. We have also established a network of home-based specialist curators with expertise in particular areas of biology.

The core SAR data is now freely available from www.ebi.ac.uk/chembldb. Uptake of this has been strong and vigorous, and follows on from an active pre-release data sharing programme with over sixty leading academic and industrial groups, in which we gathered feedback on data organisation, quality and community needs. This has already led to the use of the data in several high impact publications (for example Keiser *et al.*, 2009).

We have also had the opportunity to think about new ways of addressing our user community, and one aspect of this has been the establishment of a group blog 'the chembl-og' (at www.chemblog.org), which has had in excess of 29,000 visits this year from over 110 different countries. Series of posts on the chembl-og include a monograph on each newly launched drug and these have been cross-linked from several resources.

CHEMBL DATABASE

Anne Hersey, Anna Gaulton, Louisa Bellis, Yvonne Light, Shaun McGlinchey

The ChEMBL database is the group's primary repository for the bioactivity and drug data. Data is entered manually from the literature using outsourcing contracts – this task is not amenable to automation due to the way in which SAR data is presented in the medicinal chemistry and pharmacological literature. Data is extracted from approximately ten journals covering primarily medicinal chemistry, drug metabolism and pharmacokinetics. During the year, we took the decision to significantly extend our coverage of natural product research, with an additional 24,000 new natural product entries. The data curation and integration is performed on-site at EMBL-EBI using a mixture of expert manual and automated curation/integration approaches.

Historically, the customers of Inpharmatica licensed the data for incorporation into their internal information systems. Consequently, when the group started work at EMBL-EBI there was no 'user-friendly' front-end access point available for the large-scale SAR data. A web front-end has been developed to allow easy searching by compound name, compound structure (with a zero install Java molecular sketcher), target name, target sequence, bioactivity and target class. This provides flexible and easy access to the core data. Future interface developments include enhanced browsing of compound ontologies (for example pharmacological/structural classes of molecules), organism search capabilities (for example the ability to extract all primate data), and cross-indexing of targets with disease associations. At all times these will contain links to other EMBL-EBI portals (such as ArrayExpress, UniProt, Reactome, ChEBI and PDBe etc.).

The database is available for download in a number of standard formats (Oracle and MySQL) from <ftp://ftp.ebi.ac.uk/chembl>.

SARFARI

Mark Davies, Kazuyoshi Ikeda, Yvonne Light

SARfari is a portal for chemical tool/lead discovery built around a gene family paradigm – similar targets often bind structurally related ligands and so integration of data within a gene family is valuable to lead discovery progression. We have recently released Kinase SARfari (www.sarfari.org/kinasesarfari) which is centred on the modulation of the protein kinase family. Protein kinases are often key modulators of extra- and intra-cellular signalling, and are the subject of intense academic and industrial research. Currently there are ten drugs targeting this family, all with a unique profile of specificity across the family.

Kinase SARfari contains a reference alignment of the family combined with known three-dimensional structures, bound ligand conformations, binding sites, SAR data (extracted from chembldb) and clinical candidate data. Additionally the system allows local installation and provides the ability to register private data. This then allows a unique fusion of public and private data in a fully integrated manner which is accessible from a single web interface.

DRUGGABILITY PORTAL

Anna Gaulton, Mark Davies, Kazuyoshi Ikeda

Inpharmatica developed a series of approaches to predict the tractability/likelihood of success for a drug discovery programme addressing a particular target. This concept is of critical importance to the development of novel therapeutics, where empirical experience has established that the vast majority of screening programmes fail to deliver a progressible lead compound. Therefore the prediction of tractable (or druggable) targets is an important general problem. Previously, Inpharmatica had developed a sequence-based, ligand-based, and structure-based approach to target assessment - these approaches are now built on top of the pilot Druggability Portal, funded previously by the Industry Programme. The software components have been re-established, and work is underway to build on the pilot, incorporating features such as therapeutic strategy (recombinant protein targets are very different in properties to small molecule targets).

RESEARCH

Patricia Bento

Our research activities currently centre around two areas; 1) data mining of chemblpdb for rules for chemical tool optimisation, and 2) target/drug identification for neglected diseases.

For the data mining project we have focused on the analysis of the bioactivity properties of peptides, with a view to developing a combined computational and synthesis platform for automated chemical tool discovery. Initial work has been to annotate the peptides contained within chemblpdb with their component amino acids and other structural features (for example cyclisation) and then to analyse the binding affinities of these peptides (figure 2). A novel component in this annotation is the indexing of the data with novel Ligand Efficiency (LE) measures, leading to a publication currently in press. Given that there are currently in excess of 40,000 distinct peptides in chemblpdb, we can potentially build general 'rules' for amino acid exchanges of direct application in peptide tool/drug design, and also in applications such as protein engineering.

There exists much opportunity to apply (or 're-use') current therapeutics against indications other than those they were previously developed for, and this concept can be readily extended to their application in the treatment of infectious neglected diseases (e.g. malaria, leprosy, Chaga's disease, etc.). Chemblpdb provides a unique platform to rapidly annotate sequences with available clinical agents or other chemical tools. A pipeline has been developed to perform this at genome scale, and these data then provide the opportunity to rapidly prioritise therapeutically tractable genes. One example of this is the 'chemical annotation' of the *Schistosoma mansoni* genome with collaborators from the Wellcome Trust Sanger Institute.

FUTURE PROJECTS AND GOALS

Toxicity prediction: given the large volume of binding and functional data within chemblpdb, we are starting to address a new and complementary area of drug discovery and development – toxicology. Increasingly, as chemists are able to make compounds with acceptable affinity and pharmacokinetic properties, compounds are now often failing later (and more expensively) with toxicity liabilities. The source of these failures is complex, and can potentially be related to on-target pharmacology, off-target pharmacology, the activity of various metabolite species, or through unanticipated drug–drug interactions. Surprisingly, little well organised, publicly accessible data exists on the toxicity properties of clinical development candidates and drugs. We are investigating a number of collaborative approaches to populate new resources with toxicity related data, both for human therapeutics, but also for other life science areas, such as cosmetics, herbicides and so forth.

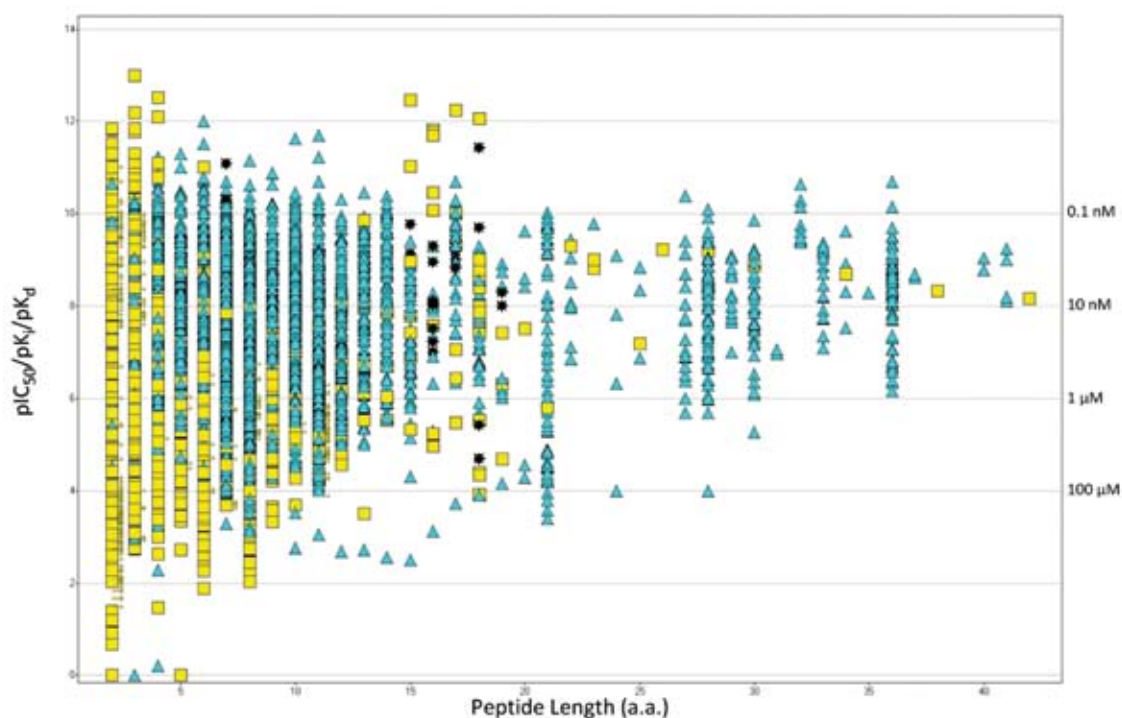


Figure 2. Peptide ligand affinity for enzyme (yellow) and receptor (cyan) targets, plotted as a function of peptide length.

Pre-competitive data sharing: the pharmaceutical industry is undergoing a sea change in its attitude to data sharing and pre-competitive collaboration. The core business of pharmaceutical companies is returning to that of compound optimisation and development. This has led to several opportunities to act as a data broker for pharmaceutical companies for some of their historical data, both inspired by reasonably large-scale funding calls, and also by initiatives by the pharmaceutical companies themselves. EMBL-EBI is a core data provider for many resources for industry, and this role is now further extended by the content of chembl.db. We are thus well placed to act as an informatics hub to facilitate pre-competitive data sharing while leveraging our extensive network of existing contacts. This will lead to the development of a generic registration and deposition system for small molecule bioactivity data.

Resource integration: a further area of future work is the establishment of effective data sharing agreements with key resources around the world, ranging from PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) the chemspider resource (www.chemspider.com), TDR targets (www.tdrtargets.org), NC-IUPHAR (www.iuphar-db.org), Binding-DB (www.bindingdb.org), DrugBank (www.drugbank.ca) as well as seamless integration with other EMBL-EBI resources. We will actively explore opportunities for shared data entry and curation. We will also release a GPCR version of SARfari, and investigate the building of a general integration system for bioactivity/sequence/structure data built around both a specific organism and gene family viewpoint.

Outreach and training: we are planning a wide range of community engagement and training for our new resources in 2010, spanning conference presentations, lab visits to geographical clusters of labs interested in the data, and also through a series of training workshops (for example, the Small Molecule Bioactivity Resources course at EMBL-EBI in January 2010). These will be complemented by offering on-site training for industrial partners, and for other groups on an *ad hoc* basis.

Team Members

Group Coordinator
Anne Hersey*

Content Curator
Louisa Bellis*
Yvonne Light*

Web Developer
Mark Davies*
Shaun McGlinchey*

Data Integration
Anna Gaulton*

Scientific Application Developer
Kazuyoshi Ikeda*

Postdoctoral Researcher
Patricia Bento*

Visitors

Bissan Al-Lazikani
Celerino Abad-Zapatero

* Indicates part of the year only

Publications

2009
Berriman, M., *et al.* (2009).
The genome of the blood fluke
Schistosoma mansoni. *Nature*, 460,
352-358

Harland, L. & Gaulton, A. (2009).
Drug target central. *Expert Opin.*
Drug Discov., 4, 857-872

Overington, J. (2009). ChEMBL.

An interview with John Overington,
team leader, chemogenomics
at the European Bioinformatics
Institute Outstation of the European
Molecular Biology Laboratory
(EMBL-EBI). *J. Comput. Aided Mol.*
Des., 23, 195-198

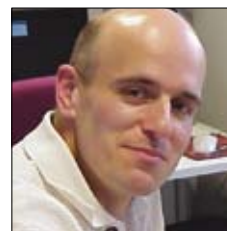
Steinbeck, C., *et al.* (2009). New
open drug activity data at EBI.
Chem. Cent. J., 3, 62

Other references

Keiser M.J., *et al.* (2009). Predicting
new molecular targets for known
drugs. *Nature*, 462, 175-181

Christoph Steinbeck

*PhD 1995, Rheinische Friedrich-Wilhelm-Universität, Bonn.
Postdoctoral research at Tufts University, Boston, USA, 1996-1997.
Head of Research Group for Structural Chemoinformatics, Max Planck
Institute of Chemical Ecology, Jena, 1997-2002.
Habilitation in Organic Chemistry, Friedrich-Schiller-Universität, Jena, 2003.
Head of Research Group for Molecular Informatics, Cologne University
Bioinformatics Center (CUBIC), Cologne, 2002-2007.
Lecturer in Chemoinformatics, University of Tübingen, 2007.
At EMBL-EBI since 2008.*



Chemoinformatics and Metabolism

65

INTRODUCTION

The Chemoinformatics and Metabolism team aims to provide the biomedical community with information on small molecules and their interplay with biological systems. The group develops methods to decipher, organise and publish the small molecule metabolic content of organisms. We develop tools to quickly determine the structure of metabolites by stochastic screening of large candidate spaces and enable the identification of molecules with desired properties. This requires algorithms for the prediction of spectroscopic and other physicochemical properties of chemical graphs based on machine learning and other statistical methods.

We are also investigating the extraction of chemical knowledge from the printed literature by text and graph mining methods, improved dissemination of information in life science publications, as well as open chemoinformatics workflow systems. Together with an international group of collaborators we develop the Chemistry Development Kit (CDK), the leading open source library for structural chemoinformatics as well as the chemoinformatics subsystem of Bioclipse, an award-winning rich client for chemo- and bioinformatics.

CHEBI – SMALL MOLECULE ONTOLOGY AND NOMENCLATURE REFERENCE

Nico Adams, Kirill Degtyarenko, Adriano Dekker, Paula de Matos, Marcus Ennis, Janna Hastings, Kenneth Haug, Duncan Hull, Zara Josephs, Inma Spiteri, Steve Turner

The Chemical Entities of Biological Interest (ChEBI) database is a freely available dictionary of molecular entities focused on 'small' chemical compounds. It was initiated to provide standardised descriptions of molecular entities that enable other databases at EMBL-EBI and worldwide to annotate their entries in a consistent fashion. ChEBI focuses on high-quality manual annotation, non-redundancy and provision of a chemical ontology rather than full coverage of the vast range of chemical entities.

ChEBI systematically combines information on small molecular entities from three main sources, namely the IntEnz database of enzymes (EMBL-EBI), the KEGG COMPOUND database and the PDBeChem database of ligands (EMBL-EBI). A number of subsidiary, freely accessible sources are manually annotated and integrated, such as ChemIDplus (<http://chem.sis.nlm.nih.gov/chemidplus/>) from the NIH, the NIST Chemistry WebBook (<http://webbook.nist.gov/>) and COME and RESID (EMBL-EBI databases). Molecules directly encoded by the genome (such as nucleic acids, proteins and peptides derived from proteins by cleavage) are generally not included in ChEBI.

A major feature of ChEBI is its chemical ontology which makes ChEBI uniquely powerful because it allows relationships between molecular entities (or classes of entities) to be recorded in a defined way. ChEBI has also created its own chemically specific relationships to properly define relationships between entities. The entire dataset is made available in OBO format at <ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology>.

ChEBI uses nomenclature, symbolism and terminology endorsed by the International Union of Pure and Applied Chemistry (IUPAC), the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) and the International Union of Basic and Clinical Pharmacology Committee on Receptor Nomenclature and Drug Classification (NC-IUPHAR). All the data in ChEBI is non-proprietary or derived from a non-proprietary source and is therefore freely available. In addition, each data item is fully traceable and explicitly referenced to the original source. ChEBI is built as a relational database and is available at www.ebi.ac.uk/chebi/, as well as via FTP and the web service interface. ChEBI records are also regularly indexed by PubChem (<http://pubchem.ncbi.nlm.nih.gov/>).

ChEBI contains two- and three-dimensional chemical structures, stored as connectivity tables (MDL molfiles). The corresponding image, SMILES string, IUPAC International Chemical Identifier (InChI; www.iupac.org/inchi/) and InChIKey are automatically generated. ChEBI release 62 contained over 450,000 chemical entities.

Significant developments in 2009

Many, many more compounds: in 2008, EMBL-EBI was awarded a substantial grant to support the transfer of a large collection of information on the properties and activities of drugs and a large set of drug-like small molecules from the publicly listed company Galapagos NV into the public domain. This data, now named ChEMBL (see page 61; www.ebi.ac.uk/chembl/), consisted of the protein targets and their associated bioactive small molecules. The small molecule data consisted of a chemical structure and associated synonyms. These were manually annotated from the original publication. The data were loaded into ChEBI and associated properties such as formulae, mass and charge were automatically generated from the chemical structures. Duplicate entities were found by using the InChI and were merged into a single entity. In order to distinguish the two datasets, a ranking system was devised in which fully annotated ChEBI entities are allocated three stars and partially annotated ChEMBL entities are allocated two.

Submission tool: to alleviate the backlog of user requests for inclusion into ChEBI, and to invite the community to participate more directly in ChEBI's future growth and development, we have developed a web-based software utility to enable direct user submissions. Once a submission has been received, it is annotated by our expert annotators. User submissions are then made publicly available (after the next release cycle) and attributed to the submitter (although the submitter has the option to remain anonymous if he/she wishes). The ChEBI submission tool is available online at www.ebi.ac.uk/chebi/submissions.

Ontology development: changes to the ontology have been made to address user concerns. The relationship 'is part of' has been replaced with 'has part' to remove ambiguous usage. A new relationship 'has role' has been introduced to link chemical entities with their role, making it easier to distinguish between chemical entities and their functions. Furthermore an umbrella term was added called 'role' which now includes 'chemical role', 'biological role' and 'application'.

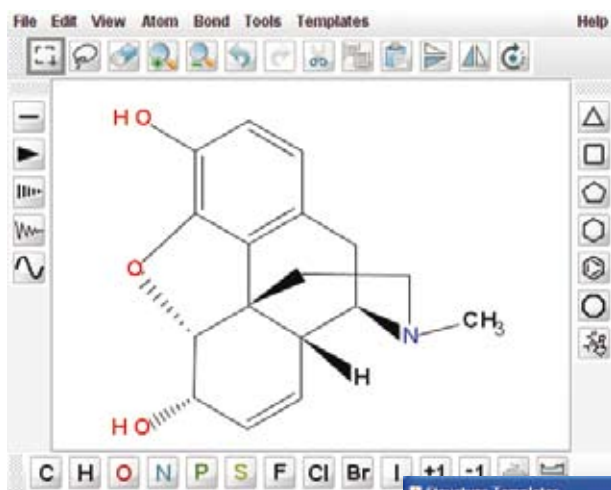
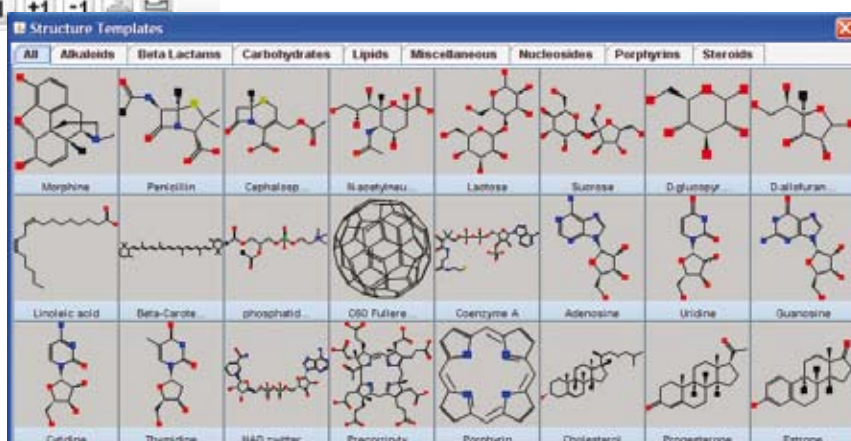


Figure 1. JChemPaint is an easy-to-use open source structure editor and viewer, available as a Java applet for embedding in websites. It has a templates browser in which the user can choose structures to insert.



INTENZ – THE RELATIONAL ENZYME DATABASE AND NOMENCLATURE RESOURCE

Rafael Alcántara, Kirill Degtyarenko, Paula de Matos

The classification and nomenclature of enzymes developed by the NC-IUBMB is based on function; namely the reaction catalysed. Each classified enzyme is assigned a specific numerical identifier known as the EC number. The Integrated relational Enzyme database (IntEnz; www.ebi.ac.uk/intenz/) provides a complete, freely available database focused on enzyme nomenclature approved by the NC-IUBMB, combined with additional information from the ENZYME database (www.expasy.org/enzyme/). Currently, IntEnz contains 4,150 approved entries as well as proposed new entries and revisions of previously published entries.

IntEnz is the master copy for the ENZYME database provided by the Swiss Institute of Bioinformatics (SIB) and is jointly maintained by SIB and EMBL-EBI. The IntEnz entries contain cross references to various resources such as BRENDA, GO, KEGG, MEROPS, PDB, UM-BBD, NIST Thermodynamics of Enzyme-Catalysed Reactions and the Catalytic Site Atlas. EC numbers can also be searched or browsed for via the EC number hierarchy. IntEnz entries have cross references to their corresponding proteins via UniProtKB (www.uniprot.org/). IntEnz also contains literature references to relevant publications. Biochemical compound names from IntEnz are used in the creation of ChEBI.

IntEnz is freely available for download in XML, BioPAX and flat file formats from the EBI FTP site.

Significant developments in 2009

IntEnz BioPAX export: BioPAX is a standard data exchange format for biological pathway data and has been added to the download area, making IntEnz a more useful resource for the systems biology community. The exported file makes use of the existing Rhea BioPAX export, and can be imported with tools such as Protégé.

Rhea reactions available from IntEnz: enzyme entries whose plain text reactions could be parsed and incorporated into Rhea (see below) are shown with links to the Rhea and ChEBI entities. Rhea reactions are much more accurate and informative, as compound names are standardised and the stoichiometry is carefully checked.

IntEnz is open source: all the source code written for IntEnz has been uploaded to the SourceForge project (<http://sourceforge.net/projects/intenz/>). The repository is updated so that the latest production code is available for anyone to use freely.

RHEA: ANNOTATED REACTIONS DATABASE

Rafael Alcántara, Kirill Degtyarenko, Paula de Matos

Rhea is a reaction database where all reaction participants (reactants and products) are linked to ChEBI which provides detailed information about structure, formula and charge. Rhea provides built-in validations that ensure both elemental and charge balance of the reactions. The database has been populated with the reactions found in the EC list (and in the IntEnz and ENZYME databases), extending it with additional known reactions of biological interest. While the main focus of Rhea is enzyme-catalysed processes, its scope is wider and can include reactions not contemplated in IUBMB enzyme nomenclature, or even not catalysed by enzymes.

Rhea is a manually annotated resource providing:

- stable reaction identifiers; these are independent of EC numbers but are linked to them via cross references;
- directionality information if the physiological direction of the reaction is known;
- the possibility to link several reactions together to form complex reactions. This feature can also be used to split reactions into partial reactions;
- extensive cross references to other resources including enzyme-catalysed and other metabolic reactions, such as the EC list (in IntEnz), KEGG, MetaCyc and UniPathway;
- chemical substructure and similarity searches on compounds in Rhea, thereby allowing reactions involving similar compounds to be found.

The reactions contained in the database will be used in IntEnz, in all relevant UniProtKB entries, and also in the UniPathway database to generate pathways and metabolic networks. Rhea is available at www.ebi.ac.uk/rhea, and is updated in monthly releases. Rhea is open source and the source code is available from the SourceForge project (<http://sourceforge.net/projects/rhea-ebi>). Data can be freely downloaded in BioPAX and RXN formats.

INFERRING METABOLOMES

Pablo Moreno

The metabolome refers to the complete set of small molecules (<1500 Da) present in a biological sample or organism. Furthermore, it is also relevant to know the localisation of those molecules in terms of tissue, cell type and subcellular

location. This set of molecules (and their locations) represents a major insight into the phenotype of the organism.

Enzymes, their catalysed reactions and pathways are frequently used to infer a metabolome as these account for most of the metabolite diversity in the cell. Small molecules also play important roles in signalling, gene regulation, biochemical reactions and enzyme regulation. The existing biochemical knowledge on these areas can be used for uncovering the metabolome.

The focus of our current work on metabolomes is to integrate different types and sources of knowledge together in a database in order to contribute to the elucidation of complete metabolomes. These sources of knowledge include present metabolic databases, chemical databases, text mining, expression data and thermodynamic feasibility.

Knowledge of metabolomes would benefit a number of areas in systems biology, biochemical engineering and drug discovery. In areas such as metabolomics, current studies are only able to identify 20–30% of estimated total compounds. Detailed metabolome knowledge, including localisation, expected abundance, chemical structures and simulated spectra for instance, would be a major aid in the metabolomics mass spectral annotation process. In the field of metabolic engineering, the last five years have seen the emergence of a number of whole genome metabolic models in which flux balance calculations are possible. Despite detailed curation and the integration of a massive number of reaction/metabolites, only a few of these models have real chemical structure data or localisation information associated with them. This kind of modelling would benefit from chemical knowledge and metabolite localisation within the cell, as thermodynamic constraints and compartmentalisation could be added to the regularly used mass balances, as has already been undertaken in some works. Also topology of the networks based on chemical substructure linkage rather than reaction edges, would provide a further level of detail for mass balances. Finally, drug discovery research and comparisons of effectiveness of lead compounds could be complemented with information on metabolite similarity/dissimilarity, depending on the target.

COMPUTER-ASSISTED STRUCTURE ELUCIDATION AND PREDICTION OF NMR SPECTRA

Stefan Kuhn, Gilleain Torrance

The understanding and simulation of metabolic networks is currently hindered by a significant lack of information on the structural identity and physical properties of biochemical metabolites in organisms under investigation. Methods developed by our team provide the means to quickly determine the structure of metabolites by stochastic screening of large candidate spaces based on spectroscopic methods. Our so-called SENECA system is based on a stochastic structure generator which is guided by a spectroscopy-based scoring function. Based on our past efforts in the field of structure elucidation, we have started to integrate developed algorithms and methods into the Bioclipse platform in order to make them easily accessible for users. The SENECA software is available as a Bioclipse feature. Currently it contains the following modules:

- stochastic generators for creating candidate structures;
- simulated annealing and genetic algorithm;
- scoring functions using ^{13}C shifts or HMBC 2D spectra.

Since the elucidation process may take up to a couple of hours for very large molecules, the user interface allows the user to monitor the progress and review the results found so far, before using his/her own knowledge to refine these.

SENECA integrates with other spectrum components of Bioclipse, which are also maintained by our group. These allow spectrum simulations of structures elucidated by SENECA, and viewing and manipulation of the spectral data.

JCHEMPAINT – AN OPEN SOURCE CHEMISTRY EDITOR

Stefan Kuhn, Mark Rijnbeek

Based on the CDK, JChemPaint (JCP) is a free and open source chemistry editor. It enables users to edit, save, import and export chemical structures. Users are supported by templates, layout algorithms and a user interface supporting the most common tasks in structure drawing.

JCP has been developed alongside the CDK and in 2009 we undertook a major effort to bring it to a production state. We focused on the applet version and its use as an effective tool for database front-ends on the internet. Speed of loading and usage was a particular focus. We have had several development releases and the final release is planned for November 2009.

ORCHEM – AN OPEN SOURCE CHEMISTRY SEARCH ENGINE FOR ORACLE

Mark Rijnbeek

OrChem was developed to provide a free chemistry extension and search engine to Oracle, the *de facto* database platform in the commercial world. It uses the CDK Java library residing inside the database to provide the required

cheminformatics information/data (figure 2).

OrChem adds registration and indexing of chemical structures to support fast substructure and similarity searching. Searches can be performed with response times of seconds for millions of compounds, depending on query complexity and user search preferences.

A first public release has been made, together with a paper submission to the *Journal of Cheminformatics*. ChEBI will use OrChem as a back-end for its structure search requests.

FUTURE PROJECTS AND GOALS

The integration of major open source software such as OrChem and JChemPaint into our services projects has been a major activity in 2009. We will continue improving these projects, such as adding SMARTS querying and extending the structure editor to deal with all types of structures, such as polymers.

In 2010 we will work closely with our collaborators to produce an enzyme portal involving Rhea and IntEnz databases, the idea being to produce a one-stop shop for enzyme related data. Ensuring a sustainable growth for the ChEBI database and automatic classification of existing entries within the ChEBI ontology will be the focus of our attention. Not only do these tasks require a larger team for data collection and curation but also research into the automated assembly and validation of ChEBI datasets to aid the human curators.

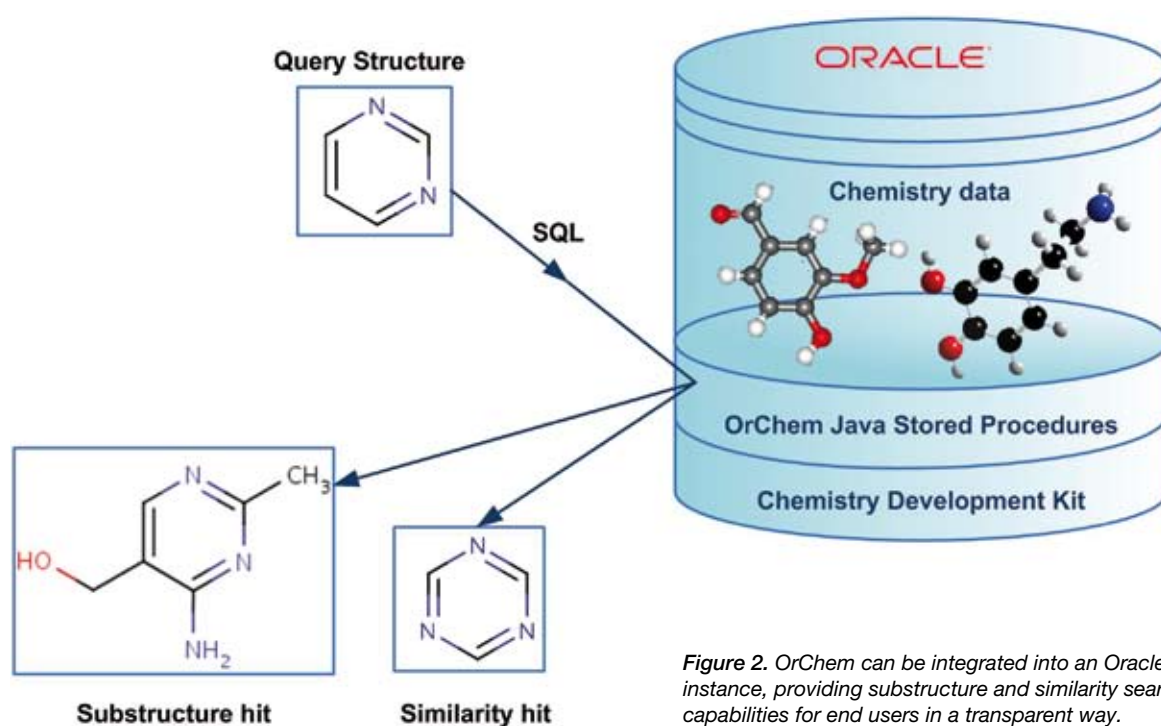


Figure 2. OrChem can be integrated into an Oracle instance, providing substructure and similarity search capabilities for end users in a transparent way.

Team Members**Coordinator**

Paula de Matos

Software Engineers

Rafael Alcántara

Adriano Dekker*

Janna Hastings

Duncan Hull*

Kenneth Haug*

Stefan Kuhn

Senior Software Engineer

Mark Rijnbeek*

Scientific Database Curators

Marcus Ennis

Zara Josephs*

Inma Spiteri*

Steve Turner*

Postdoctoral Fellow

Gilleain Torrance*

Predoctoral Fellow

Pablo Moreno

Visitors

Kirill Degtyarenko

Nico Adams

* Indicates part of the year only

Publications**2008**

Kuhn, S., *et al.* (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, 9, 400

2009

Degtyarenko, K., *et al.* (2009). ChEBI: An open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics*, 26, 14.19.11-14.19.20

Kuhn, S., *et al.* (2009). Components for computer-assisted structure elucidation. *Chem. Cent. J.*, 3, 62

Kuhn, T., *et al.* (2009). Creating chemo- and bioinformatics workflows, further developments within the CDK-Taverna Project. *Chem. Cent. J.*, 3, 42

Steinbeck, C., *et al.* (2009). New open drug activity data at EBI. *Chem. Cent. J.*, 3, 62



Weimin Zhu

*MSc. 1993, University of Toronto.
Project Manager, GDB Project, Toronto, until 2000.
Head of Bioinformatics, Synax Pharmar, Toronto, until 2002.
At EMBL-EBI since 2002.*

Database Research and Development

71

INTRODUCTION

In February 2008, the Database Applications team was reorganised to form the Database Research and Development team due to the creation of the PANDA group.

The team's new mandate is to conduct research and development to find new technologies and solutions to meet challenges related to very large databases (VLDB), which includes data distribution problems when network speed is a bottleneck, and the solutions required to manage and query VLDBs efficiently.

The size of bioinformatics databases has been increasing exponentially over the last ten years. Some core resources are approaching, or have already reached, multi-terabytes in size. This trend of growth has accelerated in recent years by the introduction of new data types and high-throughput data producing technologies. Today, we are facing all the challenges a VLDB brings, such as those in data operational management, data access performance, and data mirroring and distribution. Our current infrastructure in these areas thus requires upgrading in order to realise the full potential of data-rich resources, and optimise the usage of our human and hardware resources.

Data distribution or data synchronisation has been a very active research area in the information technology sector for quite a few years. As a result, some useful tools, such as rsync and its variants, have been developed. The core of the technology, also known as delta compression or delta encoding, is to find the differences (deltas) between two sets of files located remotely from each other, and only those deltas are transferred to the target computer to allow it to rebuild a new version of the files based on the older version. If the deltas are significantly smaller than the full data files, a significant network saving can be achieved. The technology is working very well for some applications, such as internal data synchronisation and remote software distributions. However, it has not been successfully adopted by the database community for the distribution of large datasets, although it has been attempted by many people. When applied to large datasets, the latency of the runtime calculation to identify the deltas between source and target files becomes a new major bottleneck. Also, unclustered changes in files, which are common in bioinformatics databases, can result in full data transfer instead of deltas. A few other technologies, such as peer-to-peer solution and data replication, have also been tried by some bioinformaticians. Although some progress has been made, their application in distributing large-scale biological databases is still very limited.

ACTIVITY UPDATE

This year, our main focus was first to develop a new algorithm, sdeltac, to shorten the network latency for distributing large datasets across the network. The algorithm is a delta encoding-based solution but by taking advantage of some of the unique characteristics of biological databases, the algorithm is also a structure-based differential distribution system that overcomes the limitations of existing delta encoding solutions (see figure 1A). Some key features of biological databases include:

- in a typical database file, an entry is a common atomic data unit separated by well-defined delimiters within a specific database, such as '/' for separating entries in the EMBL-Bank archive and UniProt Knowledgebase, '>' for FASTA sequence data files, and '\n' for records in Ensembl MySQL database files. Entries also exist in XML format which has become a common alternative data format across many biological databases. In XML-based databases, entries are separated by unique opening and closing tags.

- most of the entries in a database release are unchanged. Changed entries, either updates or new insertions, represent only a small percentage of the total release. For instance, by comparing the files in non-WGS (whole genome shotgun) data divisions of EMBL-Bank release 93 and 94, only 14.9% entries have changed. By including most of the WGS projects, the result is similar (14.5%). It means that without further optimisation, only 90GB data need to be transferred for a database with total size of 600GB.
- a significant percentage of changed entries are modified entries, not new ones. The modified entries counted for 12% of the total 15% changes in the above EMBL-Bank example. This means that for this 600GB database, only 18GB has to be transferred, while the remaining 72GB data can potentially be delta compressed. Similar results were obtained by analysing other databases.

These unique features allow us to use entry-level comparison between releases as the higher level of encoding (figure 1B). Huge computation power and time can be saved by removing those unchanged entries from the lower level comparison – by word or chunk comparison process. Source entry data are stored in a hash table for quick target entry lookups. This level of comparison also makes the word-level comparison more accurate and manageable when a data file is very big.

Word-level comparison (figure 1C) starts by seeding the source tokens, a part of the source words, along the target file. This overcomes splitting the target file into an arbitrary size of a set number of words, and enhances the possibility for source strings to be mapped to the target. The token mappings are then extended from both sides of the token, to maximise the mappings and minimise the delta that have to be transferred to the client.

Entry-level and word-level comparisons are the core of the sdeltac algorithm. The delta data, string literal and source offset and length to be copied are pre-computed, and used by client to reconstruct the target file set from its local copy of the source file set.

The sdeltac algorithm is implemented using the C language, and consists of server and client components. The delta data are stored in RDBMS, which further reduces network latency. A functional prototype has been implemented including all the components except the interfaces (figure 2).

A workshop funded by the UK Research Councils (RCUK) has been scheduled for 19–20 October in Beijing, China. The participants of the workshop will exchange ideas and solutions to the challenge of distributing large biological databases, from network and computational angles.

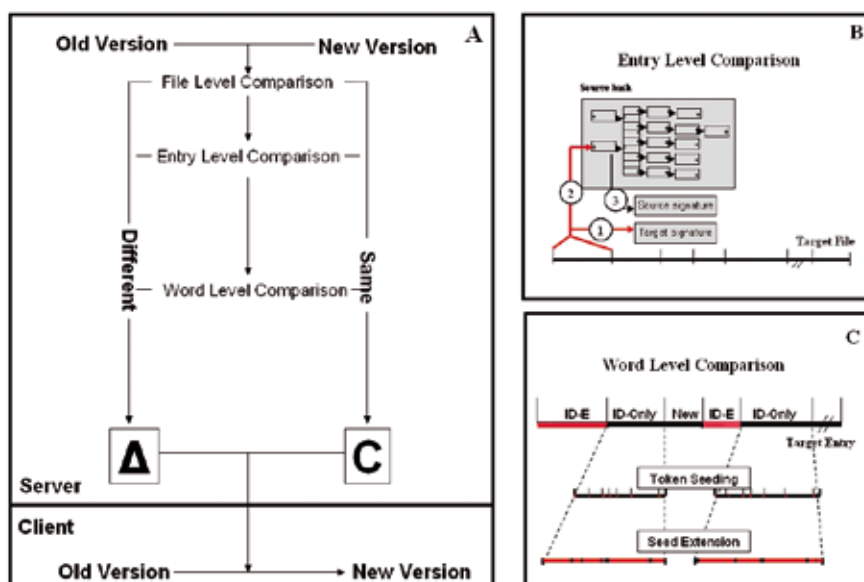


Figure 1. Overview of the sdeltac algorithm. A. The differences and similarities, shown as Δ and C respectively, between old and new sets of files are computed on file, entry and word levels on the server side. The client reconstructs the new file set from its local copy of old files, and the delta data (Δ and C) that are transferred over the network. B. Entry-level comparison. An entry read from a target/new file is calculated for its signature (1), and identifier's hash value. The hash value and entry identifier are used for looking into source/old file hash tables (2). The returned source entry signature (3) is compared with target signatures to determine whether the target entry is changed, unchanged or new. C. Word-level comparison. The changed target entries, with same entry identifiers but different signatures (ID-only) from their source counterparts, are subjected to two tiers of word matching processes. Token seeding is the process of seeding the source tokens (1/10 of predefined word size) to a target entry. The token matching is further extended by token extension process, to maximise the matching between source and target entries and minimise the Δ . Matched areas are in red, and unmapped areas are in black.

FUTURE PROJECTS AND GOALS

The continuing development and optimisation on sdeltac, on both algorithm and implementation, will be our focus for next year. New developments will include:

- algorithm adaptation and implementation work on compressed file formats;
- further development of RDBMS database schema, to allow selection of a subset of a database and version skipping;
- collaborations with database projects and data centres to test the utilisation of the program as a potential data distribution and data mirroring solution;
- commencing work on interfaces for the client to retrieve delta data from the RDBMS database.

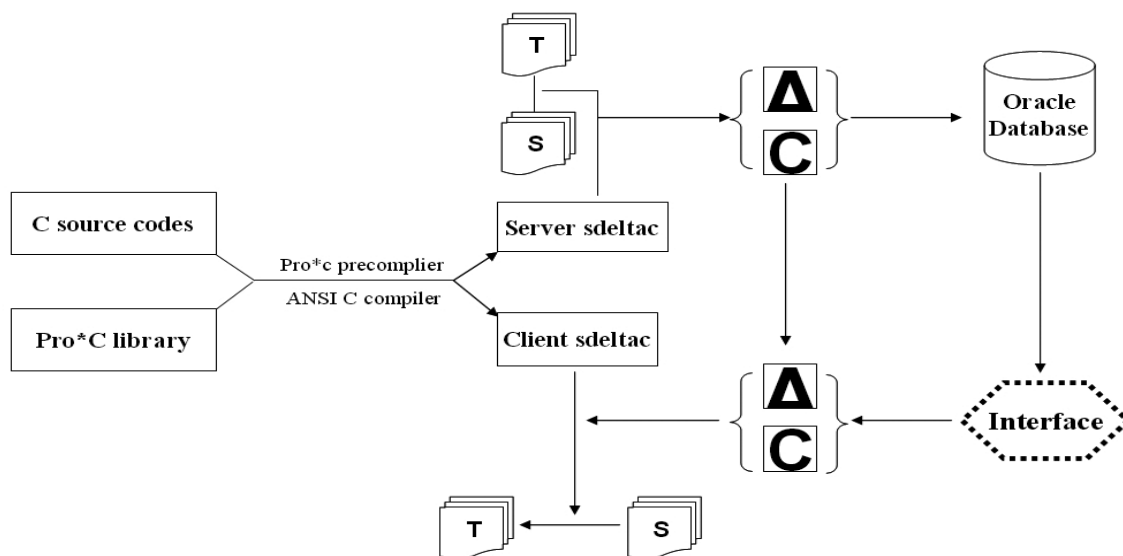


Figure 2. Implementation of sdeltac algorithm. The C program suite has server and client components. The server sdeltac processes source and target file sets. Resulting delta data (Δ and C) are stored in RDBMS, which can be transferred directly to the client, or through client interfaces, such as web services, browser or API. Client reconstructs the target file set from these delta data and the local copy source file set. The future development of the database schema and interfaces will offer the client the flexibility to select subsets of a database, and choose the versions of the source and target file sets.

Midori Harris

GO Editor



The GO Editorial Office

75

INTRODUCTION

The Gene Ontology (GO) project (www.geneontology.org/) is a collaborative effort to construct and use ontologies to facilitate the biologically meaningful annotation of genes and their products in a wide variety of organisms. At EMBL-EBI, the GO Editorial Office plays a key role in managing the distributed task of developing and maintaining the GO vocabularies, and contributes to a number of other GO project efforts, including documentation, web presence, software testing and user support.

THE GENE ONTOLOGY PROJECT

The Gene Ontology Consortium (GOC) provides the scientific community with a consistent and robust infrastructure, in the form of biological ontologies, for describing, integrating, and comparing the structures of genetic elements and the functional roles of gene products within and between organisms. Participating groups include major model organism databases and other bioinformatics resource centres (see Panel 1). The GO ontologies cover three key biological domains that are shared by all organisms (The GO Consortium, 2000, 2001):

- molecular function defines the tasks performed by individual gene products; examples include aminoacyl-tRNA ligase activity and translation elongation factor activity;
- biological process defines broad biological goals, such as signal transduction or ribosome assembly, that are accomplished by ordered assemblies of molecular functions;
- cellular component describes subcellular structures, locations and macromolecular complexes; examples include cytoplasm, ribosome and translation release factor complex.

In addition, sequence features are covered by the Sequence Ontology, which is maintained separately from the three GO ontologies (Eilbeck *et al.*, 2005).

The ontologies in GO are structured as directed acyclic graphs (DAGs), wherein any term may have one or more parents and zero, one, or more children. Within each vocabulary, terms are defined and relationships between terms are specified. The GO vocabularies define several semantic relationships between terms: *is_a*, *part_of*, and three relations representing biological regulation. The *is_a* relationship means that a term is a subclass of another; *part_of* may mean 'physically part of' (as in the cellular component ontology) or 'subprocess of' (as in the biological process ontology). Figure 1 shows a portion of the GO cellular component DAG.

GO terms and gene product annotations are used in a diverse and growing range of applications, including:

- integrating proteomic information from different organisms;
- assigning functions to protein domains;
- finding functional similarities in genes that are overexpressed or underexpressed in diseases;
- predicting the likelihood that a particular gene is involved in causing disease;
- analysing groups of genes that are co-expressed during development;
- developing automated ways of deriving information about gene function from the literature;

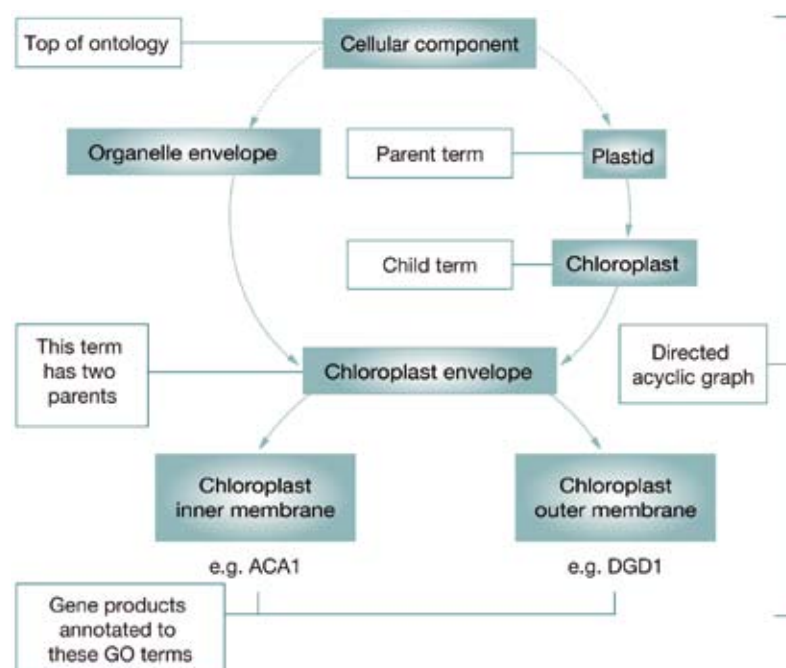


Figure 1. GO terms are organised in directed acyclic graphs (DAGs) – hierarchical structures in which any ‘child’ (more specialised term) can have many ‘parents’ (less specialised terms). For example, the cellular component term chloroplast envelope has two parents, reflecting the fact that it is a part of the chloroplast and a type of membrane. Any gene that is annotated to this term is automatically annotated to both chloroplast and membrane. Some terms and relationships have been omitted for clarity.

- verifying models of genetic, metabolic and product interaction networks.

ACTIVITIES OF THE GO EDITORIAL OFFICE

Ontology development

From its inception, the GO project has developed its ontologies for the purpose of gene product annotation. To this end, the Gene Ontology is dynamic: existing terms and relationships are augmented, refined, and reorganised as biological knowledge advances. Major improvements have been made over the lifetime of the GO project in several areas of the ontology, often in consultation with experts in relevant subject areas. Table 1 shows the current size of each of the four ontologies maintained by the GO Consortium; figure 2 illustrates GO’s growth since 2001.

Alongside the EBI GO Editorial team, curators who use GO terms for gene product annotation play a key role in the development of GO. To complement their input, the GO Consortium strives to involve members of the biological research community in the ontology development process. Curator Interest Groups can be formed of Consortium members and community experts and focus on specific areas within the ontologies. In addition, GO curators and biologists come together to consider specific biological topics at meetings devoted to ontology content. Ontology development can also improve the internal logical consistency of GO, and facilitate quality control procedures.

Significant changes introduced in GO in 2009 affect both biological and logical aspects of the ontologies: new links have been created between the molecular function and biological process ontologies, and a new relationship type, *has_part*, has been introduced. Following the introduction in 2008 of three new relationship types – *regulates*, *negatively_regulates*, and *positively_regulates* – within the biological process ontology, these relationships have now been added within the molecular function ontology and between molecular function and biological process terms. These new links have further improved the representation of biological regulation in GO. Additional links between the two ontologies using *part_of* are now being introduced.

Biological Process terms	17069
Molecular Function terms	8637
Cellular Component terms	2432
Sequence Ontology terms	1620

Table 1. Current status of the GO vocabularies (as of 1 September 2009).

Considerable progress has been made on recasting many complex process terms as explicit cross-products with other ontology terms, improving computability and supporting more sophisticated tool development. The work to date has concentrated on ‘internal’ cross-products, i.e. those that define GO terms by referring to other GO terms. For example, the *regulates* relationships allow GO to make cross-products between regulatory processes and regulated processes or functions.

Biological topics of interest in 2009 include:

- improvements to process and function terms relevant to signal transduction have continued;
- new terms have been added to expand the representation of interactions between organisms;
- process terms for branching organ development have been added;
- function, process and component terms for viral biology are being revised.

Other GO Editorial Office activities

In addition to ontology development, the GO Editor and GO curators contribute to several other GO project efforts:

- **user advocacy:** establishes lines of communication between the scientific community and the GO Consortium to ensure that GO remains useful, relevant, and accessible. This effort encompasses maintaining online project documentation and developing the GO Consortium’s web presence;
- **software development and testing:** within the GO Consortium, a group devoted to software and utilities supports both the GO Consortium and the user community with technical, software, bioinformatics and computer-science related matters. The GO Editorial Office staff participate in coding, interface design, and testing of tools such as AmiGO, a web-based GO browser, and OBO-Edit, a versatile ontology editing application;
- **annotation outreach:** makes contact with potential annotating groups to enable the GO Consortium to obtain annotation of maximally feasible quality across all species;
- **mappings to GO:** GO curators produce and maintain a number of files mapping GO terms to entries in external classification schemes, such as COGs, MetaCyc, Enzyme Commission, TIGR roles, and the MIPS Funcat.

FUTURE PROJECTS AND GOALS

The GO Editorial Office will continue to work closely with the rest of the GO Consortium and with biological experts to ensure that the ontologies are comprehensive, logically rigorous and biologically accurate. Improvements begun or continued in 2009 on signal transduction, viral biology, heart development, and other topics will therefore continue. Additional links between the biological process and molecular function ontologies will be created, using new process-specific function terms. Work on creating cross-products definitions for GO terms will continue, expanding to include orthogonal ontologies such as the ChEBI ontology and the cell ontology.

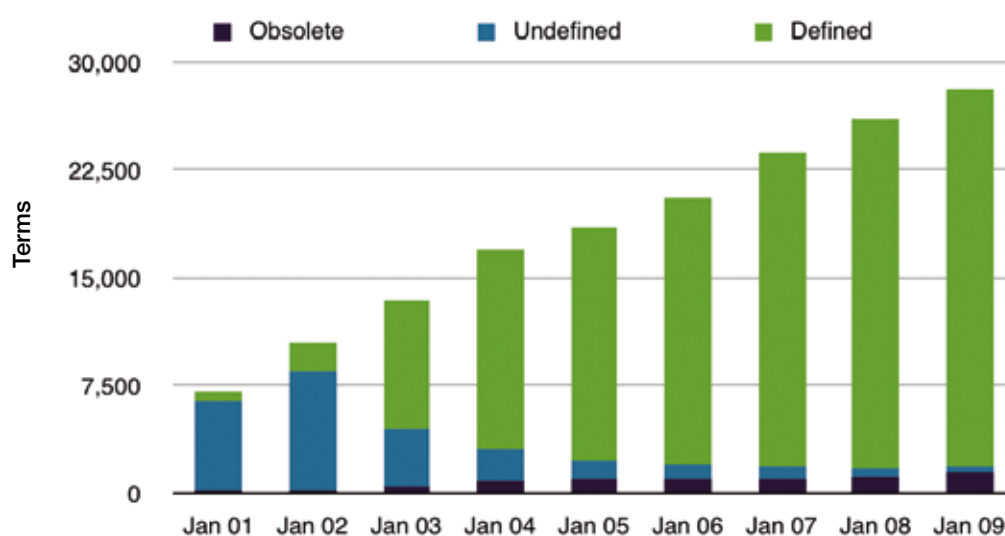


Figure 2. Annual growth of the GO vocabularies, reflecting the addition of new terms and of definitions for existing terms (the latter mainly in 2003). The graph shows the number of terms in the molecular function, biological process, and cellular component ontologies combined. Obsolete terms are those that have been removed from active use.

Team Members

Curation Coordinator

Jane Lomax

Scientific Database Curators

Jennifer Deegan

Amelia Ireland

Panel 1

Gene Ontology Consortium members

AgBase

Berkeley Bioinformatics and

Ontology Project (BBOP)

British Heart Foundation – University

College London (BHF–UCL)

CGD: The Candida Genome

Database

DictyBase (*Dictyostelium discoi-*
deum)

EcoliWiki

FlyBase

GeneDB (Wellcome Trust Sanger

Institute Pathogen Sequencing Unit;

Schizosaccharomyces pombe and

protozoan parasites)

Gene Ontology Annotation (GOA) at

EBI (UniProt annotation)

Gramene

Institute for Genome Sciences

Mouse Genome Informatics

Muscle TRAIT

Plant-Associated Microbe Gene

Ontology (PAMGO) Consortium

Rat Genome Database (RGD)

Reactome

Saccharomyces Genome Database

(SGD)

The *Arabidopsis* Information

Resource (TAIR)

The J. Craig Venter Institute

WormBase

Zebrafish Information Network (ZFIN)

Publications

2009

Carbon, S., *et al.* (2009). AmiGO:
Online Access to Ontology and
Annotation Data. *Bioinformatics*, 25,
288-289

Feltrin, E., *et al.* (2009). Muscle
research and gene ontology: New
standards for improved data integra-
tion. *BMC Medical Genomics*, 2, 6

Giglio, M.G., *et al.* (2009). Applying
the Gene Ontology in microbial
annotation. *Trends Microbiol.*, 17,
262-268

Reference Genome Group of the

Gene Ontology Consortium. (2009).

The gene ontology's reference
genome project: A unified framework
for functional annotation across spe-
cies. *PLoS Comput. Biol.*, 5,

Schober, D., *et al.* (2009). Survey-
based naming conventions for use in
OBO Foundry ontology development.
BMC Bioinformatics, 10, 125

Wortman, J.R., *et al.* (2009). The
2008 update of the *Aspergillus nidu-*
lans genome annotation: a commu-
nity effort. *Fungal Genet. Biol.*, 46,
Suppl 1, S2-13

Other publications

Eilbeck, K., *et al.* (2005). The
Sequence Ontology: a tool for the
unification of genome annotations.
Genome Biol., 6, R44

The Gene Ontology Consortium
(2000). Gene ontology: tool for the
unification of biology. *Nat Genet.*,
25, 25-29

The Gene Ontology Consortium
(2001). Creating the gene ontology
resource: design and implementa-
tion. *Genome Res.*, 11, 1425-1433



Alvis Brazma

*PhD Computer Science, Moscow State University, 1987.
Postdoctoral research in New Mexico State University.
At EMBL-EBI since 1997.*

The Microarray Informatics Team

79

INTRODUCTION

The Microarray Informatics team focuses on functional genomics data services, R&D related to biomedical informatics, and research in functional genomics data analysis, algorithms and methods. We run one of the EBI's core resources, ArrayExpress, which consists of two components:

- **ArrayExpress Archive of Functional Genomics Data:** storing publication related and other data from approximately 10,000 studies based on microarray or next-generation sequencing based experiments;
- **ArrayExpress Gene Expression Atlas:** an added value database providing easy access to information about gene expression in different cell and tissue types, under various disease and other biological conditions.

We have started working on developing the EBI Sample Database, which eventually will hold information about all samples and phenotypes deposited in any of the core databases at EMBL-EBI. We are developing software and providing data management for many biomedically-related collaborative projects. Our team was among the first to use microarray data to study transcription regulation mechanisms on a genomic scale (Brazma *et al.*, 1998). Our PhD students focus mostly on analysing functional genomics data, building models for systems biology (e.g. Rustici *et al.*, 2004, Schlitt & Brazma, 2006) and developing new methods and algorithms. Integration of data across multiple platforms, including genotypes, is among the latest activities of the team. Training provision is also one of our major activities.

SERVICES

ArrayExpress Archive of Functional Genomics Data

Helen Parkinson, Tomasz Adamusiak, Tony Burdett, Anna Farne, Ele Holloway, Natalja Kurbatova, Margus Lukk, James Malone, Gabriella Rustici, Eleanor Williams, Holly Zheng-Bradley, Ugis Sarkans and ArrayExpress Software Development team (separate report, see page 85)

The size of the ArrayExpress Archive (www.ebi.ac.uk/arrayexpress) has been doubling roughly every 15 months since it was established and new technologies for functional genomics, most importantly next-generation sequencing based assays, have been gaining in popularity. Archiving these growing amounts and variety of data has presented us with many challenges and has necessitated a complete redevelopment of the ArrayExpress Archive infrastructure, on which we have been working for the last several years. This has been a major activity of the development and production teams and when completed it will allow us more flexibility and more rational use of resources. Most of the new infrastructure components have been completed in 2009 and we plan for the full data migration in early 2010. Towards this goal we have introduced automated testing methods that will ensure better quality and robustness of software and data integrity after the migration to the new infrastructure (for more see the report by Ugis Sarkans on page 85).

From the user perspective, an important development has been the new powerful functionality of the query interface, including ontology enabled queries. This interface has been developed with the view of porting it to the new infrastructure without any additional effort. An improved REST-style web services for the Archive data has been released. To deal with the sequencing-based assay data in a consistent way, we have developed a pipeline that integrates ArrayExpress with the European Nucleotide Archive (ENA) and the European Genome-phenome Archive (EGA). The raw sequencing data are stored in the ENA and EGA databases while the processed data and metadata are stored

in ArrayExpress. We have imported sequencing data from GEO (NCBI) and reached an agreement with NCBI to exchange all metadata that is related to sequence data in GEO.

Another major direction was the promotion of the MAGE-TAB format, which is the main format for ArrayExpress. Jointly with Stanford University we have developed a stand-alone tool, Annotare, for annotating functional genomics experiments and creating MAGE-TAB documents. We have developed a standard MAGE-TAB parser, which is used in Annotare and various tools outside the EBI, including MeV (Harvard). We have also refined the MAGE-TAB export from ArrayExpress. We have released a tool for MAGE-TAB import into Bioconductor and the paper describing it has been published (Rayner *et al.*, 2006).

ArrayExpress Gene Expression Atlas

Misha Kapushesky, Tony Burdett, Juok Cho, Ibrahim Emam, Ele Holloway, Pavel Kurnosov, Andrew Tikhonov, Andrey Zorin, James Malone, Gabriella Rustici, Eleanor Williams, Helen Parkinson

The Gene Expression Atlas (www.ebi.ac.uk/gxa) is a new database at EMBL-EBI that allows users to query gene expression by gene names or properties, such as Gene Ontology terms, or by tissue types, cell types, disease states or other conditions. It is part of the ArrayExpress infrastructure and adds value to the archived data by curation, reannotation and statistical computations to enable gene or biological condition-based queries. A simple interface allows the user to query for differential gene expression and the results list the conditions where expression has been reported, while condition queries return the genes reported to be expressed under these conditions. A combination of both query types is also possible (figure 1). The query results are ranked using various statistical measures and by the number of independent studies in the database showing the particular gene-condition association. For each gene, a summary page is available, describing the behaviour of the gene across all studies present in the Atlas (figure 2). Currently the database contains information on more than 200,000 genes from nine species and almost 4,500 biological conditions studied in over 30,000 assays from over 1,000 independent studies.

The Atlas was first released as a production database in March 2009. A substantially improved version was released in June and announced by an EBI press release. Since June, there have been three more releases of the Atlas, adding a new DAS track, extensive support for programmatic access, ontology-powered queries and data downloads. We obtained industry funding from Pfizer Inc. for developing an additional functionality and to produce a stand-alone resource for integrating transcriptomic data, which will be completed in December. A publication about the Atlas has been accepted in the *Nucleic Acids Research Database* issue 2010 and will appear as a special ‘featured article’.

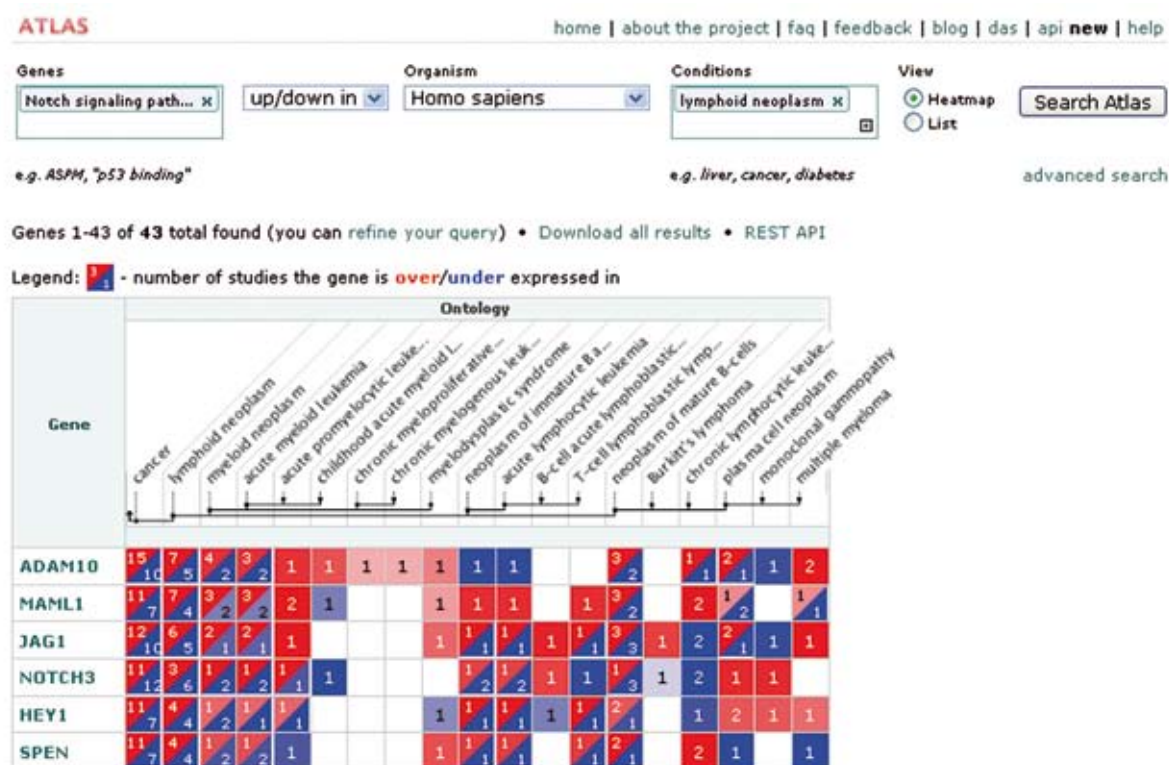


Figure 1. Query results for human genes matching GO term ‘Notch signalling pathway’ expressed in condition ‘lymphoid neoplasm’.

The team has also been working on the development of the EBI R Cloud, which was launched in September at MGED 2009, providing an easy-to-use workbench to access the R/Bioconductor computational framework at EMBL-EBI.

Towards the EBI Sample Database

Maria Krestyaninova, Helen Parkinson, Susanna-Assunta Sansone, Ugis Sarkans, Marco Brandizi, Mike Gostev, Natalja Kurbatova, Eamonn Maguire, Philippe Rocca-Serra, Johan Rung, Nataliya Sklyar

As molecular profiling (including sequencing) moves from creating reference datasets to profiling individuals and assaying specific conditions, it becomes increasingly important to record the information about the assayed sample and specific biological conditions. The information required is all the sample-associated metadata, which, for example, could specify the material sampled, the site (including organ, tissue etc.), phenotypic information (including disease states and clinical information about the individual), and the experimental conditions (drug dosage, treatments etc.). The same sample may be assayed by several different technologies, for instance, the same individual can be genotyped and profiled for gene expression (expression Quantitative Trait Locus – eQTL), and in such cases it is important to record which datasets have originated from profiling the same individual or the same sample. It is advantageous to record sample information in a separate database, which then can link out to the assay data stored in particular molecular databases. This is the task of the EBI Sample Database (ESD), which we have started building.

Substantial design and development work has already gone into existing systems developed in the team: SIMBioMS (Krestyaninova *et al.*, 2009) and BioInvestigation Index (BII) (Sansone *et al.*, 2008), as well as the new ArrayExpress architecture, which like the BII is largely based on modelling Sample Data Relationship graphs (Rayner *et al.*, 2006). This work gives the EBI a sound basis for the ESD, however we will adopt neither in its entirety. This reflects our desire to tightly couple the scope of the ESD to the needs of the real data streams entering our databases. As such, the ESD is a new development drawing on previous work rather than an adaption of the existing systems.

R&D FOR MEDICAL AND TRANSLATIONAL BIOINFORMATICS

Maria Krestyaninova, Helen Parkinson, Susanna-Assunta Sansone, Ugis Sarkans, Tomasz Adamusiak, Marco Brandizi, Mikhail Gostev, Natalja Kurbatova, Eamonn Maguire, James Malone, Philippe Rocca-Serra, Johan Rung, Nataliya Sklyar, Chris Taylor, Holly Zheng-Bradley

Several projects contribute to this activity. The NET Project, led by Susanna-Assunta Sansone and Philippe Rocca-Serra includes a series of collaborative projects with nutrigenomics, environmental genomics and toxicogenomics commu-

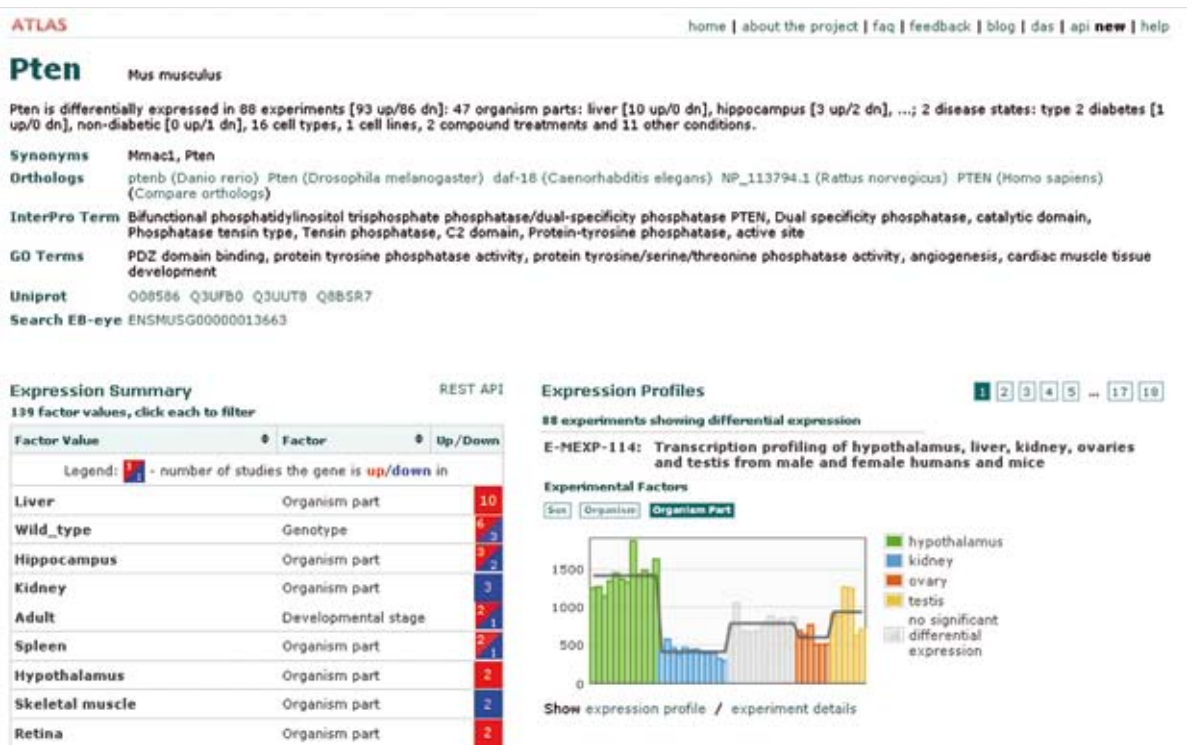


Figure 2. ‘Gene page’ for *Mus musculus Pten*. *Pten* is observed over-expressed in liver, in ten independent studies, and shows the highest significance of differential expression in the experiment E-MEXP-114.

nities. The main achievements have been the finalisation of the ISA (Investigation-Study-Assay) tools (<http://isatab.sf.net>), which are currently being prepared for a release as open source software, the launch of the BioInvestigation Index (www.ebi.ac.uk/bioinvidex) at EMBL-EBI with data dispatched to other EBI databases (thus serving as a prototype for the ESD), and work on the OBO Foundry (and in particular the OBI workshops organised in June 2009). The team also substantially contributed to the work of the ELIXIR project on standards. The team is leading a community effort to develop and promote standards in functional genomics.

Maria Krestyaninova coordinates bioinformatics and information management in two genetic epidemiology projects ENGAGE (EU FP7) and MUTHER (Wellcome Trust). A System for Information Management in BioMedical Studies – SIMBioMS, which was developed over the last five years, was published (www.simbioms.org). SIMBioMS is widely used in ENGAGE for providing data management services for high-impact scientific community studies (61 user accounts) in biobanking and genetic epidemiology (data from 800 GWAS have been uploaded). New functionality was developed for data transfer to public archives and used to submit data to EGA (HapMap3 data). A special instance of SIMBioMS was developed for the Tara Oceans project, which is a completely new application domain of marine biology. A new component – the Sample Availability system (SAIL) – has been released and a publication is in preparation. Currently SAIL contains availability data for 30,000 samples across ten European biobanks. Maria Krestyaninova is co-chairing an informatics working group in P3G (www.p3g.org/), and is contributing to standards promotion and development. Case studies have been provided for ESFRI projects BBMRI, EATRIS and ELIXIR.

As a part of the Gen2Phen project, coordinated by Helen Parkinson in collaboration with Dr Morris Swertz (University of Groningen), we have developed a model for representing phenotypes. It has been implemented using the Molgenis Software. We have populated the database with human and mouse data and are developing mappings between mammalian phenotypes. We have continued developing the Experimental Factor Ontology (EFO), which now has monthly releases and is used in the ArrayExpress and Atlas queries. A publication is in preparation.

RESEARCH

Nils Gehlenborg, Angela Gonzales, Misha Kapushesky, Katherine Lawler, Margus Lukk, Helen Parkinson, Johan Rung, Gabriela Rustici, Holly Zheng-Bradley

Human and mouse gene expression maps

While there is only one genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental or disease states. The size and structure of this ‘expression space’ is still largely unknown; most transcriptomics experiments typically focus on sampling small regions. We have constructed a global gene expression map by systematically integrating data from 5,372 human samples representing 369 different cell and tissue types, disease states, and cell lines processed in 163 different laboratories. We found that the major structure of this space is described by a small number of distinct expression profile classes, which sort according to phenotypic attributes: a haematopoietic axis discriminating the haematopoietic system, solid tissues and incompletely differentiated cell types; a malignancy axis arranging cell lines, neoplastic samples and non-neoplastic primary tissue derived samples; and a neurological axis separating nervous system from the rest of the samples. In particular, most cell lines cluster together rather than with their tissues of origin. Analysing a similar dataset for mouse from data obtained on a different platform we found that mouse expression space has a similar structure. Based on these comprehensive datasets we have implemented a prototype for a novel online resource that allows the user to query for a gene of interest to find in which conditions it is over- or underexpressed, and for every condition of interest, which genes are differentially expressed in it. This will become part of the Gene Expression Atlas. A manuscript describing this work has been provisionally accepted for publishing. This work was done in collaboration with Helsinki University of Technology and Wolfgang Huber’s group at EMBL.

Data retrieval

As ArrayExpress and other repositories of genome-wide experiments are reaching a mature size, it is becoming more meaningful to search for related experiments, given a particular study. We introduce methods that allow for the search to be based upon measurement data, instead of the more customary annotation data. The goal is to retrieve experiments in which the same biological processes are activated. This can be due either to experiments targeting the same biological question, or to as yet unknown relationships. We use a combination of existing and new probabilistic machine learning techniques to extract information about the biological processes differentially activated in each experiment, to retrieve earlier experiments where the same processes are activated, and to visualise and interpret the retrieval results. Case studies on a subset of ArrayExpress show that, with a sufficient amount of data, our method indeed finds experiments relevant to particular biological questions. Results can be interpreted in terms of biological processes using the visualisation techniques. This work was done in collaboration with Helsinki University of Technology and was accepted as a plenary talk at ISMB 2009 (Caldas *et al.*, 2009).

Transcriptional regulation of the fission yeast cell cycle

Cell division is a highly regulated process and periodic gene expression is crucial to this regulation. Gene subsets are induced at specific times during the cell cycle, when their function is required, and are then repressed when they are no longer needed; mistakes in this regulation can lead to alterations in cell proliferation which will in turn result in diseases such as cancer.

In humans, as well as in lower eukaryotes, a well-known regulatory complex is responsible for controlling the onset of DNA replication by inducing genes required for this transition. We show that in fission yeast this regulatory complex (MBF) also induces a gene whose encoded protein (Yox1p) in turn binds to MBF and represses MBF-regulated genes. In the absence of Yox1p, the MBF-regulated genes do not fluctuate during the cell cycle but remain constantly induced. Thus, MBF sets up not only the induction but also the timely repression of its target genes via Yox1p. We also provide a global analysis of all the genes regulated by Yox1p and MBF. Together, our data uncover a new negative control loop, further highlighting the sophistication of gene regulation during the cell cycle, and illustrating regulatory similarities and differences between organisms. This work was done in collaboration with Jürg Bähler's group at UCL and the work is published in *PLoS Genetics* (Aligianni *et al.*, 2009).

Transcriptional and post-transcriptional regulation of gene expression: computational analysis of microarray studies in fungal species

Katherine Lawler, a PhD student in the group, finished and submitted her PhD thesis. The thesis presents two related computational studies of the genome-wide regulation of gene expression based on the analysis of microarray datasets. The first study concerns the dynamics of a global gene expression response. The regulation of mRNA abundance by both transcriptional and post-transcriptional control implies a range of possible strategies for shaping gene expression in response to a stimulus. Katherine's work investigated strategies for shaping gene expression and the strength of evidence for regulated mRNA stability from microarray time series in the fission yeast *Schizosaccharomyces pombe*. A dynamic model of mRNA abundance was applied to simultaneous time series of mRNA abundance (DNA microarray) and transcription rate (RNA polymerase II ChIP-chip) datasets. Candidate genes were identified for which the gene expression response appears to be driven by a change in mRNA stability rather than by transcriptional control. The second study used expression analysis combined with recently predicted transcription associated proteins to identify genes co-expressed with putative DNA binding transcription factors in the recently sequenced fungal crop pathogen *Fusarium graminearum*.

TRAINING

Gabriella Rustici, Tomasz Adamusiak, Ibrahim Emam, Margus Lukk, Misha Kapushesky, Maria Krestyaninova, Helen Parkinson, Susanna-Assunta Sansone, Eleanor Williams

We have organised or participated in over 25 training events in the past year. The EMBO course on Analysis and Informatics of Microarray Data was one of the most successful training workshops at EMBL-EBI in 2009. Funding to repeat this course next year has been secured.

FUTURE PROJECTS AND GOALS

Our main goals for the foreseeable future will be

- to develop and release the fully functional EBI Sample Database and populate it with sample information from the existing core databases at EMBL-EBI;
- to continue developing the Gene Expression Atlas, enriching it with new functionality, new data including the next-generation sequencing experiments and data from protein expression;
- to increase the robustness of the established links and pipelines with ENA and EGA, with regards to the shared sequencing and genotyping data, and full metadata exchange with GEO;
- to continue our involvement in medically relevant collaborative projects to develop tools for data management, representation and analysis and to contribute to data analysis in these projects;
- to continue research in integrative data analysis, in particular using next-generation sequencing data and integrating genotype and gene expression data, and building systems biology models.

Among the concrete plans is organising a Wellcome Trust sponsored conference 'Bridging the gap between bioinformatics and medical informatics' in 2010, to work on genotype imputations using the 1000 Genomes Project data jointly with the EGA and the ENGAGE project; to work on next-generation sequencing-based transcriptomics data analysis jointly with the Cancer Research UK in Cambridge; and to start working on new collaborative projects, including SYBARIS (biomarker discovery for fungal diseases), and CAGEKID (kidney cancer).

Team Members**Technical Team Leader**

Ugis Sarkans

Coordinators

Misha Kapushesky
 Maria Krestyaninova
 Helen Parkinson
 Susanna-Assunta Sansone

Technical Coordinator

Philippe Rocca-Serra

Software Developers

Tony Burdett
 Marco Brandizi
 Mike Gostev
 Pavel Kurnosov
 Eamonn Maguire
 Nataliya Sklyar
 Andrew Tikhonov
 Andrey Zorin
 Anna Farne

Scientists

Johan Rung
 Gabriela Rustici

**Scientific Curators/
Bioinformaticians**

Tomasz Adamusiak
 Ele Holloway
 Natalja Kurbatova
 Chris Taylor
 Anna Farne
 Margus Lukk
 Eleanor Williams
 Holly Zheng-Bradley*
 James Malone

PhD Students

Nils Gehlenborg
 Angela Gonzalves
 Katherine Lawler*

Visitors

Vincenzo Belcastro
 Juok Cho
 Richard Evans
 Talay Djumabaev
 Morris Swertz
 Anna Zhukova

Personal Assistant

Lynn French

* Indicates part of the year only

Publications**2008**

Rustici, G., *et al.* (2008). Data storage and analysis in ArrayExpress and expression profiler. *Curr. Protoc. Bioinformatics*, unit 7.13, Suppl 23,1-27

Schmidt, A., *et al.* (2008). An integrated, directed mass spectrometric

approach for in-depth characterization of complex peptide mixtures. *Mol. Cell Proteomics*, 7, 2138-2150

Vinken, M., *et al.* (2008). The carcino-GENOMICS project: Critical selection of model compounds for the development of omics-based *in vitro* carcinogenicity screening assays. *Mutat. Res.-Rev. Mutat. Res.*, 659, 202-210

2009

Aebersold, R., *et al.* (2009). Report on EU-USA Workshop: How Systems Biology Can Advance Cancer Research (27 October 2008). *Mol. Oncol.*, 3, 9-17

Aligianni, S., *et al.* (2009). The fission yeast homeodomain protein Yox1p binds to MBF and confines MBF-dependent cell-cycle transcription to G1-S via negative feedback. *PLoS Genet.*, 5, 1-12

Brazma, A. (2009). Minimum Information About a Microarray Experiment (MIAME) – successes, failures, challenges. *TheScientificWorldJournal*, 9, 420-423

Brazma, A., *et al.* (2009). Introduction. *J. Bioinform. Comput. Biol.*, 7, 5

Caldas, J., *et al.* (2009). Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25, i145-i153

Field, D., *et al.* (2009). Omics data sharing. *Science*, 326, 234-236

Gehlenborg, N., *et al.* (2009). Prequips – An extensible software platform for integration, visualization and analysis of LC-MS-MS proteomics data. *Bioinformatics*, 25, 682-683

Harttig, U., *et al.* (2009). Owner controlled data exchange in nutrigenomic collaborations: the NuGO information network. *Genes Nutr.*, 1-10

Hwang, D., *et al.* (2009). A systems approach to prion disease. *Mol. Syst. Biol.*, 5, 1-23

Kauffmann, A., *et al.* (2009). Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics*, 25, 2092-2094

Krestyaninova, M., *et al.* (2009). A System for Information Management in BioMedical Studies – SIMBioMS. *Bioinformatics*, 20, 2768-2769

Lefever, S., *et al.* (2009). RDML:

Structured language and reporting guidelines for real-time quantitative PCR data. *Nucleic Acids Res.*, 37, 2065-2069

Orchard, S. & Taylor, C.F. (2009). Debunking minimum information myths: one hat need not fit all. *Nat. Biotechnol.*, 25, 171-172

Parkinson, H., *et al.* (2009). ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, 37, D868-872

Prokopenko, I., *et al.* (2009). Variants in MTNR1B influence fasting glucose levels. *Nat. Genet.*, 41, 77-81

Rayner, T.F., *et al.* (2009). MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. *Bioinformatics*, 25, 279-280

Rung, J., *et al.* (2009). Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat. Genet.*, in press

Schober, D., *et al.* (2009). Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*, 10, 125

Vingron, M., *et al.* (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol.*, 10, 8

Other EMBL publications

Brazma, A., *et al.* (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8, 1202-1215

Rayner, T.F., *et al.* (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7, 489

Rustici, G., *et al.* (2004). Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, 36, 809-817

Sansone, S. A., *et al.*, (2008). The first RSBI (ISA-TAB) workshop: 'CAN a simple format work for complex studies?' *OMICS A Journal of Integrative Biology*, 12, 143-149

Schlitt, T. & Brazma, A. (2006). Modelling in molecular biology: describing transcription regulatory networks. *Philos. Trans. R. Soc. Lond., B*, 361, 483-494



Ugis Sarkans

PhD in Computer Science, University of Latvia, 1998.

Postdoctoral research in University of Wales,

Aberystwyth, 2000.

At EMBL-EBI since 2000.

The Microarray Software Development Team

85

INTRODUCTION

Our team has been developing software for ArrayExpress since 2001. As of October 2009 ArrayExpress holds data from more than 260,000 microarray hybridisations and is one of the major data resources of EMBL-EBI.

The software development team has built the following components of the ArrayExpress infrastructure:

- Repository – the archival MIAME-compliant database for the data that support publications;
- Data Warehouse – a query oriented database of gene expression profiles (in 2009 replaced by the Gene Expression Atlas, work coordinated by Misha Kapushesky from the Microarray Informatics team);
- MIAMExpress – a data annotation and submission system;
- Expression Profiler – a web-based data analysis toolset;
- components used internally by the ArrayExpress production team.

In 2009 we have been working on overhauling the entire ArrayExpress infrastructure. The bulk of the work has been done and we aim at migrating data from the old infrastructure at the beginning of 2010.

ARRAYEXPRESS – NEW INFRASTRUCTURE

Nikolay Kolesnikov, Mohammadreza Shojatalab, Niran Abeygunawardena, Mirosław Dylag, Ibrahim Emam, Ekaterina Pilicheva, Anjan Sharma, Roby Mani

ArrayExpress is a public repository for microarray data that supports community standards – MIAME (Minimum Information About a Microarray Experiment), MAGE-ML (Microarray Gene Expression Markup Language), and MAGE-TAB (MAGE tabular).

In 2008 we started to work on replacing the MAGE-ML centred infrastructure with one based around MAGE-TAB. This effort will significantly simplify all internal data management tasks and will enable us to concentrate more on providing added value for our users.

The most significant development in our team in 2009 is not a concrete finished product, but rather the way that we work. Since the beginning of 2009 we have adopted the agile development methodology, characterised by the following main principles:

- valuing individuals and interactions more highly than predefined processes and supporting tools;
- striving to produce working software in small increments, rather than producing extensive specifications upfront and then working against those;
- working in close collaboration with the ArrayExpress production team;
- frequently re-evaluating our plans and resetting priorities if necessary.

In summary, the team is working in a more collaborative and flexible manner. The result is that we have been able to progress at a better pace than in 2008 when we tried to implement various aspects of the new infrastructure in a piecewise fashion, following the more conventional ‘specification – implementation’ process paradigm.

In 2009 most components for loading, exporting and deleting various parts of MAGE-TAB into/from the database have been implemented, in addition to the creation of a gene reannotation pipeline. The next step requires that all these components are connected together in a consistent framework, which is then extensively tested prior to migration of the data from the old infrastructure.

Another significant process-related development in 2009 has been emphasis on the quality of software that we build, and we have adopted a more comprehensive approach to testing. Previously, our software was tested in the traditional ‘manual’ testing manner, where developers implement a new module or a new feature and relay it to testers who produce a bug report if necessary and pass everything back to developers. This method of working does not align with agile development, where frequent release of new features is encouraged, and there is a collective ‘team spirit’ of project ownership.

Modern software development methodologies recognise four types of testing:

- automated programmer tests (‘unit tests’) that are built into the code and help ensure that software works correctly on a local, per method, level;
- automated functional tests that validate whether or not software works correctly from the user point of view, i.e. whether a certain sequence of interactions with the software produces the expected output;
- exploratory tests that most closely correspond to old fashioned testing and that help with finding usability problems and bugs that could not be predicted by automated tests;
- automated performance and scalability tests.

We have incorporated the first two types of tests into our work practices, and continue performing manual testing. Comprehensive performance and scalability testing is next on our agenda, in preparation for data migration and production deployment of the new architecture. Although these activities have reduced the overall progress from the short term point of view, we already have obtained concrete evidence that we can continue working in a more confident manner, knowing that we cannot inadvertently perturb something that was working before, and that automated testing works behind the scenes and will highlight any mistakes.

ARRAYEXPRESS REPOSITORY USER INTERFACE

Nikolay Kolesnikov

In 2009 the ArrayExpress Repository interface was restructured by introducing a new powerful search mechanism, based on Apache Lucene, the most popular open source, full-text indexing and search technology. This development makes it easier to add new visual search capabilities to the interface. Keyword-based search behaviour is closer to that provided by PubMed which is widely regarded as the *de facto* standard in bioinformatics. The interface has already been deployed to work on top of the old ArrayExpress back-end, and it can work equally well with the new data management infrastructure. Therefore this investment will not be lost after switching to the new ArrayExpress back-end, and the change will be transparent for end users.

The web services for accessing ArrayExpress data programmatically were improved by adding more powerful search capabilities and making the interface more consistent. The user interface has also been improved by making the experiment view layout easier to use. The login mechanism for access to private submissions has been incorporated into the browse interface – this functionality previously existed only for the old advanced interface.

There are some exciting developments that we have worked on in 2009 but that will be fully integrated into existing interfaces only in 2010, such as ontology-aware search expansion (based on synonyms and ‘part_of’ relations in Experimental Factor Ontology), query term prompting and an advanced search mechanism.

ARRAYEXPRESS MAINTENANCE

Nikolay Kolesnikov, Mohammadreza Shojatalab, Niran Abeygunawardena, Mirosław Dyląg, Ibrahim Emam, Ekaterina Pilicheva, Anjan Sharma, Roby Mani

A continued focus of the team has been on improving the robustness and usability of ArrayExpress, and troubleshooting when necessary. Data flow into and from ArrayExpress needs to be maintained before the new software comes online and during the data migration process. Having to balance between two sets of software components during these processes presents an additional challenge to the team.

The new infrastructure of ArrayExpress is based around MAGE-TAB data exchange format, while the old software components were derived from MAGE-OM data model and MAGE-ML XML language. For 80% of all submissions, the abstraction levels of MAGE-TAB and MAGE-ML are similar and the mapping between the two is simple to perform. However, due to the extreme expressivity of MAGE-OM and MAGE-ML, export of the remaining 20% of legacy data has proved to be a challenge. In 2009, in preparation for the data migration, a considerable effort was invested

into ensuring that all the information components transformed from MAGE-OM into MAGE-TAB are correct also for various ‘corner cases’. This work required great care, since there are tools that depend on MAGE-TAB representation of ArrayExpress experiments (such as the Bioconductor package for importing ArrayExpress data).

The other area of significant effort has been the system for reannotating the array design descriptions submitted to ArrayExpress and used in the Gene Expression Atlas (previously ArrayExpress Warehouse). The reannotation process is performed for each Ensembl release. In 2009 the pipeline was modified to make gene orthologue information available, along with other annotations.

MEDICAL INFORMATICS

Sudeshna Guha Neogi

In 2009 we continued working with MolPAGE Data Warehouse (MoDa), a proof of concept database for managing multi-omics data. Significant effort was devoted to reannotating bioentities measured in methylation, metabolomics and protein and tissue array experiments. A report of this work has been submitted for publication.

FUTURE PROJECTS AND GOALS

The main goal for 2010 is finishing the testing of the new generation ArrayExpress infrastructure, data migration and roll out.

The ArrayExpress repository interface will continue to receive incremental updates; in particular, there are still some aspects of the interface that are served by older software (viewing array designs and protocols), and these will be replaced with new components.

There is a recent initiative in EMBL-EBI to clean up and aggregate aspects of biological sample information that are served by different EBI data resources. We will exploit our experience with dealing with various aspects of biological sample information management and reuse and adapt the relevant parts of the ArrayExpress software for these purposes.

<p>Team Members</p> <p>Software Development Project Leaders</p> <p>Nikolay Kolesnikov Mohammadreza Shojatalab</p> <p>Software Engineers</p> <p>Niran Abeygunawardena Miroslaw Dylag Ibrahim Emam Sudeshna Guha Neogi* Ekaterina Pilicheva Anjan Sharma</p>	<p>Database Administrator</p> <p>Roby Mani</p> <p><i>* Indicates part of the year only</i></p> <p>Publications</p> <p>2008</p> <p>Rustici, G., <i>et al.</i> (2008). Data storage and analysis in ArrayExpress and expression profiler. <i>Curr. Protoc. Bioinformatics</i>, 7.13.11-17.13.27</p>	<p>2009</p> <p>Krestyaninova, M., <i>et al.</i> (2009). A system for information management in BioMedical studies – SIMBioMS. <i>Bioinformatics</i>, 25, 2768-2769</p> <p>Parkinson, H., <i>et al.</i> (2009). ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. <i>Nucleic Acids Res.</i>, 37, D868-872</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Gerard Kleywegt (from July 2009)

*PhD in Chemistry, University of Utrecht, 1991.
Postdoctoral research, University of Uppsala, 1991–1996.
Coordinator and then programme director of the Swedish
Structural Biology Network (SBNNet), 1996–2009.
Appointed Professor of Structural Molecular Biology,
University of Uppsala, 2009.
At EMBL-EBI since 2009.*



Kim Henrick

(until June 2009)



The Protein Data Bank in Europe (PDBe) Team

89

INTRODUCTION

The Protein Data Bank in Europe (PDBe, www.ebi.ac.uk/pdbe) is one of the five core molecular databases (genomes, nucleotides, proteins, 3D structures, and expression data) hosted by EMBL-EBI. PDBe holds detailed knowledge of the structure and function of biological macromolecules (see figure 1) and access to this information is vital for many different users, for example in the identification of potential targets for therapeutic intervention as well as of lead structures for pharmaceutical use. PDBe usage averages approximately 1.8 million web hits per month from around 22,000 unique web hosts. In addition, we serve over 1,350 FTP addresses with 375GB of data on average per month. A combination of EMBL and Wellcome Trust funding supports the core staff and the computer hardware that is essential for the PDBe to store, give access to and integrate the deluge of data relating to 3D molecular structures.

Through our membership of the Worldwide Protein Data Bank (wwPDB) organisation we are an equal PDB partner with the United States (RCSB) and Japan (PDBj) and all partners work closely to maintain the single international archive for structural data. We also integrate the experimental data derived by 3D cryo-electron microscopy and electron tomography techniques, and derive the molecular biological assemblies of structures held in the PDB.

WORLDWIDE PROTEIN DATA BANK ACTIVITIES

A major joint project between PDBe and RCSB is chemistry remediation in the PDB. This involves two classes of molecules (inhibitors and antibiotics) which have been inconsistently handled in the past. Some of these molecules are best represented as a single molecule whereas others can be represented by a sequence of biological origin. In the former case, some 420 instance of linked HET group compounds have been converted to new three-letter coded single residues (for example, the various PPACKs and penicillins). This area of work is carried out at RCSB, whereas the remediation of cyclic, modified and conjugated peptides is undertaken by PDBe. Broadly speaking, these molecules can be considered to be antibiotics and they may be of ribosomal or non-ribosomal origin. All instances in this set are to be represented as a SEQRES of linked residues.

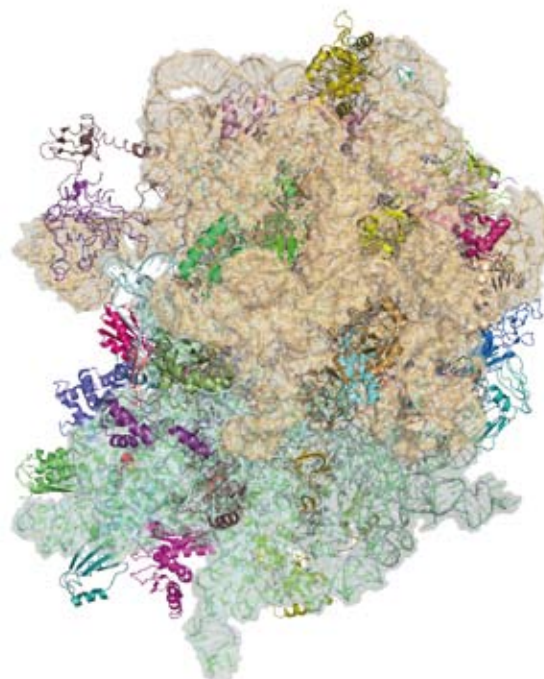


Figure 1. In October 2009, the Nobel Prize for Chemistry was awarded to three prominent structural biologists (Venki Ramakrishnan, Tom Steitz and Ada Yonath) for their studies on the structure and function of ribosomes. All the results of their structural studies (and those of many others) are available from the PDB and EMDB. This figure, created by Jawahar Swaminathan (PDBe), shows the structure of the *Thermus thermophilus* 70S ribosome (PDB entries 2WDI and 2WDG). The PDBe has created a website with information about the prize-winning structures; see www.ebi.ac.uk/pdbe/docs/nobel/.

The non-ribosomal, non-gene peptide-like molecules such as actinomycin require a gene cluster and are archived in the Norine database (<http://bioinfo.lifl.fr/norine/>). Over 1,300 PDB entries have been identified as containing either anti-bacterial, anti-viral, anti-microbial, anti-fungal, antibiotic, anti-cancer, anti-inflammatory, immunosuppressant, herbicide or toxin molecules. Some of these have only recently been recognised as gene products, albeit after extensive post-translational modification, and PDBe is working with the UniProt and RESID databases to make the linked database entries.

PDBe ACTIVITIES

The team underwent a change of leadership in 2009 with Gerard Kleywegt assuming the leadership in July 2009 and Kim Henrick leaving at the end of July. Overall, the transition has been carried out smoothly with Gerard taking over and continuing the European support for the wwPDB activities (www.wwpdb.org) while overseeing the introduction of new services. There has also been a change in staffing with many members of the team reaching the end of their nine-year contract and there will be a continued staff turnover for the next 18 months. Another change is in the team's name from MSD (Macromolecular Structure Database) to PDBe to reflect its major role in the management of the PDB itself through the wwPDB activities.

In 2009, PDBe has continued to attract funding with the award of a BBSRC grant to expand its Electron Microscopy Data Bank activities (EMDB; see figure 2; www.ebi.ac.uk/pdbe/emdb). Moreover, a five-year grant was awarded by the Wellcome Trust with the Grant Committee acknowledging the importance of PDBe as a critical resource serving a very wide scientific community, our proposal being strongly supported by the scientific community.

The annual PDBe Scientific Advisory Committee (SAC) meeting was held in February 2009. The current membership of the SAC includes Prof. Keith S. Wilson (York) as Chair, Prof. Andreas Engel (Basel), Prof. Udo Heinemann (Berlin), Prof. Ernest Laue (Cambridge), Dr Tomas Lundqvist (Gothenburg), Prof. Andrea Mattevi (Pavia), Prof. Randy J. Read (Cambridge), Prof. Helen Saibil (London), Prof. Michael Sattler (Munich), Prof. Titia Sixma (Amsterdam) and Prof. Torsten Schwede (Basel).

PDBe service	Description	URL
BIObar	Search system implemented as a toolbar application for Mozilla browsers	www.ebi.ac.uk/pdbe/docs/biobar.html
PDBeStatus	Search system to query the status of PDB entries	www.ebi.ac.uk/pdbe-as/pdbStatus
PDBeMapQuick	Quick access to cross-referenced information in external databases based on PDB ID	www.ebi.ac.uk/pdbe-as/PDBeMapQuick/
PDBeView	Text-based and advanced PDB search tool	www.ebi.ac.uk/pdbe-srv/view
PDBeLite	Search system based on the relational PDBe database	www.ebi.ac.uk/pdbe-srv/pdbelite
EMsearch	Search system for the Electron Microscopy database	www.ebi.ac.uk/pdbe-srv/emsearch
PDBeChem	Ligand search using the PDB reference dictionary	www.ebi.ac.uk/msd-srv/chempdb
PDBeMotif	Small 3D motif statistics and searches	www.ebi.ac.uk/pdbe-site/PDBeMotif/
PDBeSite	Ligand environment search	www.ebi.ac.uk/pdbe-site/PDBeSite
PDBePisa	Search and analysis of Protein Interfaces, Surfaces and Assemblies	www.ebi.ac.uk/msd-srv/prot_int/pistart.html
PDBeFold	Secondary Structure Matching (SSM) service for comparing protein structures in 3D	www.ebi.ac.uk/msd-srv/ssm/
PDBeTemplate	Search of local residue interactions in the PDB	www.ebi.ac.uk/pdbe-as/PDBeTemplate/
PDBeAnalysis	Validation and analysis of PDBe data	www.ebi.ac.uk/pdbe-as/PDBeValidate
OLDERADO	Clustering information for NMR entries in the PDB	www.ebi.ac.uk/pdbe/olderado/
PDBeMine	Supports <i>ad hoc</i> queries and data analysis based on the relational PDBe database	www.ebi.ac.uk/pdbe-srv/msdmine

Table 1. PDBe services.



Figure 2. It is now possible for electron microscopists to deposit their maps at EMDB and seamlessly deposit any fitted coordinates at the PDB. This image, created by Glen van Ginkel (PDBe), shows EM volumes of influenza virus ribonucleoprotein complex (RNP) into which polymerase domains have been placed. This study provides clues about the interactions between the functional elements of the RNP and the possible location of the viral RNA (EMDB entry EMD-1603 and PDB entry 2WFS).

Depositions: the PDBe has continued to process deposited entries via the EBI PDB deposition tool, AutoDep. Deposition is the process whereby scientists submit experimental structural data, currently from X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy or electron microscopy (EM), to one of the wwPDB sites which then processes, validates and annotates the data before passing it on to the central archive. The current release of AutoDep (4.3) contains significant improvements over previous releases. PDBe is committed to providing value-added data to the depositor as part of the deposition process. In addition to providing the annotated structures, we also include detailed chemical dictionary descriptions for new ligands, Electron Density Server (EDS) validation reports for X-ray entries, quaternary structure descriptions from PISA and sequence and taxonomy cross references. AutoDep supports integration of EMDB volume deposition and PDB coordinate submission. Depositors of EMDB volume data have the option to deposit fitted coordinates at the PDB. To this end, extensions have been implemented to both EMDep and AutoDep which allow information from EMDep to initiate a seamless PDB coordinate deposition via AutoDep. The latest release of AutoDep further supports the uploading of NMR structures and data as a CCPN project.

Services: the group has developed new services that allow users to carry out simple textual queries or more complex 3D structure-based queries. The newly designed ‘PDBeView Atlas pages’ provide an overview of an individual PDB entry in a user-friendly layout which serves as a starting point to further explore the information available in the PDBe database. New services include a mapping service between PDB entries and associated databases at the chain level and a detailed system to check on an entry’s status in the annotation pipeline. Table 1 summarises the main services.

FUTURE PROJECTS AND GOALS

For all biologists seeking to understand the structural basis of life, for researchers looking for the causative agents of disease and diagnostic tools and for the pharmaceutical and biotech industries, the aim of PDBe is to continue to provide integrated data resources that evolve with the needs of structural biologists. To achieve this goal, the main future objectives for PDBe are to:

- expertly handle deposition and annotation of structural data as one of the wwPDB deposition sites. We aim for an average turn-around time of one working day or less, employ expert annotators with experience in structure determination, and eventually intend to handle approximately one third of all depositions worldwide;
- provide an integrated resource of high-quality macromolecular structures and related data. This is implemented by developing and maintaining advanced structural bioinformatics databases and services that are, or even define, the state of the art. They should be kept up to date to keep pace with the growth of the PDB archive and ideally be available on a 24/7 basis for 360+ days per year;
- maintain in-house expertise in all the major structure determination techniques (X-ray, NMR and EM) in order to stay abreast of technical and methodological developments in these fields, and to work with the community on issues of mutual interest (e.g. data representation, harvesting, formats and standards, and validation of structural data).

The wwPDB partners are committed to the development of a common deposition and annotation tool. The steering committee reviewed the proof-of-concept testing phase in July 2009. The key infrastructure architecture components covering the data sharing technology and workflow technology were demonstrated and the team will produce a significant workflow example by January 2010 and a completed component by June 2010. The project is expected to be finished in 2011.

Team Members**Deposition Curators**

Barbara Beuth
Glen van Ginkel
Richard Newman
Gaurav Sahni
Sanchayita Sen
Jawahar Swaminathan

Database Development

Harry Boutselakis
Norman Cobley
Dimitris Dimitropoulos
Jorge Pineda
Robert Slowley
Antonio Suarez Uruena

Database Administrator

Melford John

Search and Retrieval

Adel Golovin
Matt Harrison*
Hiren Joshi*
Eugene Krissinel
Thomas Oldfield

Structural Bioinformatics

Christoph Best
Miriam Hirshberg
Anne Pajon*
Chris Penkett
Sameer Velankar
Wim Vranken
Alan Wilter Sousa da Silva*

Team Secretary

Celia Copp*

* Indicates part of the year only

Publications**2008**

Davis, A.M., *et al.* (2008). Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today*, 13, 831-841

Strombergsson, H., *et al.* (2008). Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *J. Chem. Inf. Model.*, 48, 2278-2288

Wilkins, M.R., *et al.* (2008). Information management for proteomics: A perspective. *Expert Rev. Proteomics*, 5, 663-678

2009

Berman, H., *et al.* (2009). The Worldwide Protein Data Bank. In 'Structural Bioinformatics', Weissig, H. & Bourne, P.E. (eds), 293-303, John Wiley & Sons

Kirmizis, A., *et al.* (2009). Distinct transcriptional outputs associated

with mono- and dimethylated histone H3 arginine 2. *Nat. Struct. Mol. Biol.*, 16, 449-451

Kleywegt, G.J. (2009). On vital aid: The why, what and how of validation. *Acta Crystallogr. Section D*, 65, 134-139

Krissinel, E. (2009). Crystal contacts as nature's docking solutions. *J. Comput. Chem.*, in press

Laughton, C.A., *et al.* (2009). COCO: A simple tool to enrich the representation of conformational variability in NMR structures. *Proteins: Structure, Function and Bioinformatics*, 75, 206-216

Read, R.J. & Kleywegt, G.J. (2009). Case-controlled structure validation. *Acta Crystallogr. Section D*, 65, 140-147

Strombergsson, H. & Kleywegt, G.J. (2009). A chemogenomics view on protein-ligand spaces. *BMC Bioinformatics*, 10, Suppl 6, article S13

Vranken, W.F. & Rieping, W. (2009). Relationship between chemical shift value and accessible surface area for all amino acid atoms. *BMC Struct. Biol.*, 9, 20



Peter Rice

*BSc 1976, University of Liverpool, UK.
Previously at EMBL Heidelberg (1987–1994), The Sanger Centre (1994–2000) and LION Bioscience
(2000–2002).
At EMBL–EBI since 2003.*

Developing and Integrating Tools for Biologists

93

INTRODUCTION

The team's focus is on the integration of bioinformatics tools and data resources. We also have the remit to investigate and advise on the e-Science and Grid technology requirements of EMBL-EBI, through application development, training exercises and participation in international projects and standards development. Our group is responsible for the EMBOSS open source sequence analysis package, the Taverna bioinformatics workflow system (originally developed as part of the myGrid UK e-Science project) and the EMBRACE project that integrates access to bioinformatics tools and data content through standard-compliant web services.

THE GRID

Grid technology is proposed as the next-generation infrastructure necessary to support and enable the collaboration of people and resources through highly capable computation and data management systems. Current Grid projects in high-energy physics focus primarily on the sharing of computational resources, large-scale data movement and replication for simulations, remote instrumentation steering or high-throughput sequence analysis. The most visible Grid project is currently CERN's Large Hadron Collider (LHC) Computing Grid and the EGEE project to distribute and analyse the data resulting from the LHC experiments. Such infrastructures are generally termed 'Computational Grids'. However, much bioinformatics requires support for a scientific process that has relatively more modest computational needs, but has significant semantic complexity. These are generally termed 'Data Grids'. There are hundreds of resources and applications available to today's biologist via either 'command line' applications, databases, flat files, web forms or graphical user interfaces. These may be either local to the user, or provided by remote sites. What is more, these resources are updated frequently. A user needs to find, discriminate among and choose the most appropriate services, and may need to be notified when resources are changed or updated. In an 'e-Science' context, this necessitates adapting and wrapping resources so that they comply with existing and emerging standards, specifications and technologies.

To date, Grid development has focused on the basic issues of storage, computation and the resource management needed to make a global scientific community's information and tools accessible in a high-performance environment. However, from the e-Science point of view, the purpose of the Grid is to deliver a collaborative and supportive environment that enables geographically distributed scientists to achieve research goals more effectively, while allowing their results to be used in developments elsewhere.

EMBOSS

The European Molecular Biology Open Software Suite (EMBOSS) is a collaborative open source sequence analysis package originally started in 1996 by Peter Rice at the Sanger Centre in Hinxton, in collaboration with EMBL-EBI and with Alan Bleasby at the Rosalind Franklin Centre for Genomics Research (formerly HGMP) in Hinxton. It is particularly appropriate that the whole project moved in August 2005 to EMBL-EBI as it had its origins in software developed by Peter Rice at EMBL Heidelberg, and after a period of uncertainty EMBOSS is now funded for a further three years of core development and support. The EMBOSS project is jointly coordinated by Peter Rice and Alan Bleasby and EMBOSS is available from <http://emboss.sf.net/>.

A key factor in the success of EMBOSS, and in particular its selection as the application platform for the EMBRACE and myGrid projects, has been its development and implementation of the AJAX Command Definition standard or

ACD files. These define the interface of each EMBOSS application, and are directly used by the application on startup for all processing of the command line and interaction with the user. Because the ACD file has a full description of all inputs, output and parameters, and provides full control over the input and output data formats, many other projects have used EMBOSS as the core applications suite. In SoapLab and myGrid we go further, by extending the ACD file syntax to define all other command line-driven applications. The ACD definitions are first converted into an XML style that is compatible with Object Management Group (OMG) application standards, and then used to define two web service interfaces – one general string-based interface for all applications, and a more type checked ‘derived’ interface.

In the past year we have been able to treble our effort on EMBOSS development. This allows us to pursue several new approaches. The EMBOSS libraries are expanding to support a far broader set of public and proprietary data resources, including non-sequence data, with the aim of making data available from any data resource cross-referenced by an existing entry in (among others) the EMBL nucleotide archive or the UniProt Knowledgebase. We are also adding support for Ensembl, BioMart and DAS servers. We are working with the other projects in the Open-Bio Foundation (BioPython, BioPerl, BioRuby) to agree common standards, beginning with the interpretation of the various ‘FASTQ’ format standards for next-generation short read sequence and quality data. We are building a controlled vocabulary and hierarchical structure (the EDAM ontology) within the EMBRACE consortium to describe all data inputs, outputs and analysis methods. These are annotated directly within the EMBOSS application definitions and automatically built into web service definitions in SoapLab.

We have already started an ambitious programme of new developments in database indexing, graphics outputs, pattern searches, automated documentation and help text. Support for GFF3 and Sequence Ontology (SOFA) annotation of protein features was added in the new EMBOSS 6.1.0 release in July 2009.

We are keen to encourage contributions from outside developers, and have completed a major refactoring and re-documenting of the source code and the programming interface together with new documentation for system administrators which is to be published as open source books through Cambridge University Press.

Together with partners in the EMBRACE project, we are looking to build an extensive set of ‘adapter’ or ‘shim’ services from EMBOSS to interconvert bioinformatics data structures between the output and input formats of various remote services. This has been shown to be a very useful approach to linking services in both the myGrid and BioMoby environments. As EMBOSS includes complete definitions of inputs and outputs, we are now working on adding the necessary metadata to describe exactly how each output is generated from the input data and other options selected by the user. This will provide the foundation for an ontological description of EMBOSS applications and of services derived from them, and could be applied to other non-EMBOSS services made available through SoapLab and also defined through the ACD language.

The need to provide support for the biological community has required us to build and support a native Windows implementation (mEMBOSS). The current EMBOSS release (6.1.0, July 2009) has a fully supported Windows implementation with the Jembo interface to support the needs of EMBOSS and general bioinformatics course providers. We are now working on automated generation of BioPerl-based wrappers covering all EMBOSS applications in collaboration with SciTegic Inc., with the Galaxy project team, and are looking to work with several other EMBOSS interface developers to provide notification and proposed solutions for future changes as they appear in the development version of EMBOSS.

EMBRACE

The EMBRACE project, an EU-funded Network of Excellence, is now in its fourth year, with the aim of defining and implementing a consistent standard interface to integrate data content and analysis tools across all EMBL-EBI’s core databases and those provided by our partners. The early focus of this five-year project was on the sequence and structure data resources at EBI and the EMBOSS applications. Our group is also active in defining the core technologies used by EMBRACE, including BioMart data federation methods, web services provided by the EBI External Services group, and the Taverna workbench as an end-user client.

SoapLab (<http://soaplab.sf.net/>) does not access individual analysis programmes directly but uses a general purpose wrapping system that hides all the details about finding, starting, controlling and using applications. The advantages of SoapLab and the OMG LSAE specification are that, in a standard way, it allows analyses and their input and output data to be specified using an XML-based metadata description. The SoapLab web service interface allows clients access to the metadata.

Together with Martin Senger, the developer of SoapLab, we developed a new version of SoapLab, SoapLab2 (<http://soaplab.sf.net/soaplab2/>). SoapLab2 is a major redevelopment of the existing SoapLab code and was first released in October 2007. Our group has continued to contribute to the further development of SoapLab by implementing new features and fixing performance-related issues. The current release is SoapLab 2.1.1.

The group is maintaining the EBI SoapLab server together with the External Services team. The services are now using the LSF system for batch job execution. Existing web services were extended by adding new SoapLab services for the latest release of EMBOSS (6.1.0), including for the serving all the third-party EMBASSY applications whose licensing allows open public services (a few depend on third party utilities which have restrictions on their use – we avoid such dependencies in the main EMBOSS package). These services were made available using the document/literal wrapped protocol as recommended by the EMBRACE Technology group, as well as retaining previously supported RPC/encoded versions for backward compatibility with existing Taverna workflows and other users. We have comprehensively tested the new services and ensured that migration of users from the old to the new will require only minor changes to workflows. Strongly typed services have been generated using a variant of SoapLab developed by Peter Ernst at the German Cancer Research Centre. Our group has also contributed to this variant of SoapLab by implementing support for a common sequence datatype.

We are also exploring the suitability of the DAS annotation protocol as an interface to EMBOSS applications. We have implemented a prototype DAS server for a few EMBOSS applications, based on the myDAS implementation of DAS-1.

An EMBRACE internal wiki page is used internally to maintain information on the development of analysis tool services by all partners. Service documentation pages for individual SoapLab services were prepared and linked to the main EMBRACE project page. These service documentation pages include a description of the services, list of inputs/outputs, WSDL reference, and an example usage of each service. We plan to integrate these pages with the new Spinet module in SoapLab2 to allow users to test run the services from their browsers before they actually use the web service interface from their own client applications (e.g. Taverna).

Data services for EMBRACE depend on existing services provided by EMBL-EBI. We do not intend to reinvent the wheel, rather we plan to provide sufficient metadata for these services to enable the EMBRACE application programming interface to publish services that are automatically well defined and interoperable. These include the search and retrieval services from the External Services team and the BioMart services from the Ensembl team. This work will continue through the coming year.

BioMart provides a generic data warehousing solution for fast querying of large biological databases and integration with third-party data and tools. The system consists of a query-optimised database and interactive user-friendly interfaces written in both Java and Perl. The project has successfully evolved from the original Ensembl specific ‘EnsMart’ to a generic system renamed ‘BioMart’. Our group aims to support the project by enhancing the existing system so it can manage more of the data resources maintained at EMBL-EBI and other partners in EMBRACE. Following the recent development and restructuring of the BioMart configuration system, Perl API and web service interfaces, we have successfully attracted increased traffic on BioMart servers around the world. In addition we are now managing web services and the BioMart central server (at www.biomart.org). The BioMart project relocated earlier this year to the Ontario Institute for Cancer Research in Toronto, Canada. In collaboration with OICR we are in the process of major architectural redesign in order to extend data federation, scalability, security and optimisation of the system. In addition to existing features, the software will support analysis and visualisation plugins, as well as secure data submission to marts. The redesign is driven by the cancer data management and analysis platform which will bring forward use cases from wet lab biologists, and subsequently add to the usefulness of the mart data. The BioMart central portal and web services are managed by our group at EMBL-EBI. We have also conducted several training workshops on BioMart web services during the last year to reach a wider bioinformatics community.

THE TAVERNA PROJECT

Our group developed the Taverna workbench as the biological specialist participant in the UK-funded myGrid project and we have continued this collaboration through Tom Oinn’s participation in the Open Middleware Infrastructure Institute (OMII-UK). This project was aimed at developing and maintaining open source high-level service-based middleware to support the construction, management and sharing of data-intensive *in silico* experiments in biology. Taverna currently has over 9,000 downloads, 1,500 installations, and users in Europe and worldwide, including in industry and in areas beyond bioinformatics (for example in astronomy).

Taverna has already been adopted as the interface of choice by the EMBRACE project, by BioMoby in Canada and others. Taverna allows a user to search for available services using a variety of ‘scavenging’ mechanisms, to link these services together into workflows, and to specify inputs and parameters. The workflow is then launched under Taverna, and both the results and the provenance data are made available on completion. Services made available include the EMBRACE SoapLab services, BioMoby, and EMBL-EBI’s EB-eye web services and the BioMart databases.

Taverna has been redesigned over the past two years and is now available as the legacy 1.7.2 release and as the new 2.1 beta version. Taverna 2 has many enhancements to provide improved scalability for large data-intensive workflows, reduce memory usage, and a graphical workflow editor.

Taverna is made available under the Lesser General Public License through SourceForge at <http://taverna.sourceforge.net/> and www.mygrid.org.uk/tools/taverna/taverna-2-0/.

FUTURE PROJECTS AND GOALS

The services provided by the group remain largely SOAP-based web services. These have proved themselves to be highly useful to prototype and develop service and metadata standards. We are looking, especially through the EMBRACE project, to migrate to true Grid services, but like many other groups we are waiting for the long-anticipated merging of web and grid service standards.

The EMBOSS project plans to expand in the coming few years to cover bioinformatics more generally, including genomics, protein structure, gene expression, proteomics, phylogenetics, genetics and biostatistics. This will require the participation of external groups to expand the project beyond its current EBI base, and we are actively seeking potential partners in each area. We will expect to build a service-based e-Science architecture around the applications and data resources through the EMBRACE project, with support and guidance from the community of users in academia and industry.

The EMBRACE project is now in its final phase. The services have been developed and standardised. We are now developing metadata standards including an ontological description of the data types used and the methods provided. These will be annotated within each service’s Web Service Definition Language (WSDL) file, and queried through the EMBRACE registry portal.

Team Members	Publications	
Scientists Alan Bleasby Syed Haider* Jon Ison Shaun McGlinchey* Tom Oinn* Mahmut Uludag	2009 Carver, T., <i>et al.</i> (2009). DNAPlotter: Circular and linear interactive genome visualization. <i>Bioinformatics</i> , 25, 119-120 Gibson, A., <i>et al.</i> (2009). The data playground: An intuitive workflow	specification environment. <i>Fut. Gen. Comp. Systems</i> , 25, 453-459 Haider, S., <i>et al.</i> (2009). BioMart central portal – Unified access to biological data. <i>Nucleic Acids Res.</i> , 37, W23-W27 Smedley, D., <i>et al.</i> (2009). BioMart – Biological queries made easy. <i>BMC Genomics</i> , 10, 22
*Indicates part of year only		

Johanna McEntyre (from May 2009)

*PhD Plant Biotechnology, 1990, Manchester Metropolitan University.
Editor, Trends in Biochemical Sciences, Elsevier, Cambridge.
Staff Scientist, NCBI, National Library of Medicine, NIH, USA.
At EMBL-EBI since 2009.*



Peter Stoehr

(until May 2009)



Literature Resource Development

97

INTRODUCTION

The biomedical literature is the formal record of achievement and scientific understanding of the biomedical research community. As such, searching, reading and browsing the literature are key components of all research strategies. Furthermore, many biological and biomedical databases, including those produced at EMBL-EBI, are heavily linked to relevant portions of the scientific literature to provide functional information and supporting evidence.

Aside from the clear value of a stand-alone search and retrieval system for the literature, it will become increasingly important to provide deeper integration of the scientific literature with the information held in the ever-expanding databases. As the amount of available data grows, finding useful information becomes more challenging. The scientific literature has a role to play in bringing research data closer to the surface of our search systems by acting as an integrative force for related underlying data.

Advances in online publishing and web technologies over the past decade are providing researchers with new approaches to extracting useful biomedical information from the scientific literature. In particular, open access publishing promises greater flexibility for manipulating and redistributing the knowledge found in research papers.

The EBI currently hosts a database of biomedical abstracts called CiteXplore, which provides a good foundation for further growth of the literature resources. CiteXplore content is imported from a number of sources, including: PubMed (from the US National Library of Medicine); AGRICOLA (from the USDA-National Agriculture Library); patents (from the European Patent Office); Chinese Biological Abstracts (CAS-SICLS) and CiteSeer. There are currently over 22 million abstracts available, of which about 19 million are PubMed abstracts. There are approximately 1.9 million patents, with the remaining data sources completing the full dataset.

Access to the data in CiteXplore is provided via both a website (www.ebi.ac.uk/citexplore/) and web service (SOAP). Both use Lucene, an open source search engine, as the search and retrieval system. The query functions that come as part of the Lucene package are customised to some extent, to reflect the specific data structure of the scientific literature. We also broaden the search by using the MeSH dictionary to expand query terms; therefore a search for 'atorvastatin' will also return abstracts in which only Lipitor (and not atorvastatin) are mentioned.

ACTIVITIES DURING 2009

CiteXplore growth

Sharmila Pillai, Peter Stoehr

Building on the content additions from previous years, CiteXplore grew in 2009 by approximately 1.23 million records, from the following sources:

- 900,000 PubMed records;
- 300,000 patents;
- 25,000 AGRICOLA records;
- 2,500 UK Health Information Resource records.

The Health Information Resource records represent a new data source and have been deposited in CiteXplore as a part of the UK PubMed Central (UKPMC) project (described in further detail below). These records are harvested from the NHS Evidence hub and cover best practices and treatment guidelines in medicine. We have developed new protocols to convert records from MARC format to the pubxml format used by the CiteXplore database. The records are updated on a weekly basis.

Database cross references

Several of the EBI databases are linked to CiteXplore and database cross references continue to be created to UniProt, IntAct, PDBe, InterPro and EMBL-Bank, based on the literature citations by both submitters and curators within those databases. We have also taken the first steps towards incorporating text-mining solutions in an integrated fashion into the CiteXplore interface. This makes use of the Whatizit text-mining resource built by the Rebholz-Schuhmann research group at EMBL-EBI. This function highlights named entities such as genes/proteins, GO terms and organisms in the text of CiteXplore records. The highlighted terms are then linked to underlying data sources.

Origin	Articles with citations	Total citations	Citations per article	Citations matched	Articles cited ¹	% citations matched ²
UKPMC XML	294002	11441260	38.92	9660898	2586159	84.44
CrossRef	2394648	67666153	28.26	45157083	7968318	66.74
UKPMC OCR	575420	10617137	18.45	10617137	2699659	100.00
Overall	3264070	89724550	27.49	65435118	8827245	72.93

Table 1. Account of database links available in CiteXplore (October 2009).

¹ The 'Overall' figure for the number of articles cited is not simply the sum of the articles cited from the three sources because some articles are cited more than once. ² The number of citations matched for the OCR UKPMC data is 100% because only the references within articles that resolve to PubMed have been used; therefore all citations can be matched to PubMed. This has the likely effect of falsely raising the overall number of citations matched.

Citation cross references

Information regarding the references cited within an article is required to calculate citation network information and allow bibliometric analysis for the CiteXplore dataset. CiteXplore uses two data sources for this purpose: 1) the UKPMC archive of full-text articles; and 2) the CrossRef metadata web service. We have access to the UKPMC archive of data files via our participation in the UKPMC project and access to the CrossRef metadata web service is by membership subscription. The CrossRef subscription allows us to query and retrieve the metadata for articles from participating publishers, about a quarter of which also contain the references cited within those articles.

About 90 million references from 3.3 million articles are now used to calculate citation information (see table 2). This is a large increase over the course of a year (in 2008, there were 60 million references from 2 million articles in the network). This step increase is largely due to the full utilisation of the data in the CrossRef web service, and the inclusion of the scanned portion of the UKPMC dataset (about two thirds of the content). In 2008, use of the CrossRef web service was just being explored, and only the XML portion of UKPMC (about one third) was being used to calculate citation information.

In total, almost nine million records have been cited at least once. This corresponds to approximately 45% of PubMed. In the absence of the use of the CrossRef dataset, this drops to about five million articles (approximately 25% of PubMed), demonstrating that, even though only one quarter of the CrossRef metadata records contain bibliographic information, this adds significant value to the citation network. As such, CiteXplore now contains the biggest citation network available in the public domain.

In CiteXplore, both the articles that cite a given record, plus the articles that the record itself cites, are displayed.

Database	No. Database Records	No. Articles
EMBL	1,364,314	308,822
UniProt	4,654,178	228,339
InterPro	41,821	16,805
IntAct	10,819	4,064
PDBe	52,109	27,618

Table 2. Summary of sources of citation information used in CiteXplore's citation network (as of 8 September 2009).

Links to full-text articles

Peter Stoehr

We continue to maintain a database of links to full-text articles, synchronised with updates to PubMed and AGRICOLA, created on the basis of publisher-specific rules, or Digital Object Identifiers (DOIs) harvested via CrossRef. New full-text links were added to key data sources, such as the Cochrane Reviews and Health Technology Assessments. All the newly added UK Health Information Resource records contain links to full text.

CiteXplore now includes over 20 million links to external sources, with over 7.8 million DOIs and links to 2.7 million PDFs. This represents an increase of three million external sources, 1.4 million DOIs, and 100,000 PDF links in 2009.

UK PubMed Central (UKPMC)

Johanna McEntyre, Peter Stoehr, Sharmila Pillai, Alan Horne, Dietrich Rebholz-Schuhmann

UKPMC is a freely accessible archive of the full text of peer-reviewed biomedical research articles. It is based on PubMed Central, built at the National Library of Medicine (NLM), USA. UKPMC contains around 1.5 million full-text articles that are hosted at the University of Manchester (MIMAS). The underlying article data and the article display mechanism are packaged in a 'black box', known as PMCi or 'PMC international', shared by the US PMC site and updated with content on a daily basis. The search function is also currently supplied by the US site, and there is a manuscript submission and grant reporting function customised for UK biomedical research funding agencies.

UKPMC is supported by all major UK funding agencies, each of which has mandated that their grantees place all papers published as a result of that funding into UKPMC, preferably as open access articles. The funding agencies are: the Wellcome Trust, MRC, BBSRC, British Heart Foundation, Cancer Research UK, Arthritis Research Campaign and the National Institute for Health Research.

The project entered a new phase in July 2008, with a three-year grant being awarded by the Wellcome Trust (on behalf of the UKPMC Funders Group) to the University of Manchester (MIMAS and the National Centre for Text Mining [NaCTeM]), the British Library, and EMBL-EBI (text mining from the Rebholz-Schuhmann research group and services from the Literature Services group). As the basic UKPMC database and grant reporting is in place, the goal of this round of funding is to add value to the search and display of articles by integrating the content with underlying data resources, and applying text-mining techniques to this task.

This involves: 1) the creation of a web interface for search and retrieval of full-text articles; 2) creating a new full-text search index; 3) the addition of semantic mark-up to that index by use of text-mining techniques; 4) teasing apart the rendering of the articles themselves from the surrounding areas of the page, such as the margins, header and footer, so that UKPMC-specific functionality can be added to these areas; and 5) addition of new content.

The role of the Literature Services group in these activities has been to broker the search and retrieval mechanisms between the data (i.e. full-text research articles) and the web interface, include the added value from text mining (see figure 1), and make new content available via CiteXplore.

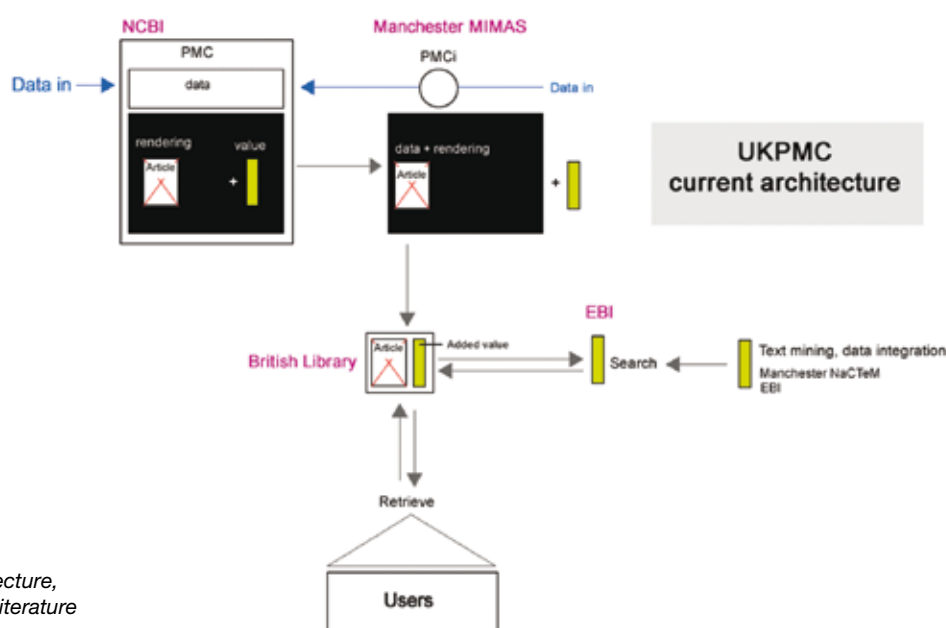


Figure 1. UKPMC architecture, showing the role of the Literature Services group.

During 2009, much of the activity of the Literature Services group has been focused on creating the infrastructure to perform these tasks. In particular, we have provided programmatic access to the 22 million bibliographic metadata records in CiteXplore to UKPMC via web services. Including these data in the UKPMC website is considered an essential part of a scientific literature search, providing greater breadth of coverage than searching the 1.5 million articles in UKPMC alone. This has required quality assessment of our service, plus refinement and customisation of both the indexed fields and the web service response to meet the needs of the UKPMC interface.

The second major portion of our work has been to assist in the creation of a full-text semantic index. This has involved building a full-text database and import of records into this database from the PMCi database at MIMAS. This required a database design compatible with the existing metadata database, so that corresponding records in each database can be related to each other. Approximately two thirds of the full-text records in the new database are also represented as PubMed records in the metadata (CiteXplore) database. On loading the content, a full evaluation and inventory check were completed to ensure the integrity of the 1.5 million articles, and to decide which fields should be indexed for the search function. Finally, a mechanism was created for timely content updates. The maximum number of new articles that can now enter the database in one day is 10,000, taking about four hours to load. This ensures a minimal lag in content availability via the UKPMC interface.

The creation of the full-text database at the NCBI has been vital for the creation of the index with semantic enrichment, under development at EMBL-EBI by the Rebholz-Schuhmann text-mining research group. Towards the end of 2009, the creation of this index is in progress and the first set of named entities – gene/proteins, Gene Ontology (GO) terms, and organism names, are being annotated. This index will shortly be made available to the UKPMC project via a web service.

In addition to the production of the semantic index, we are also in the process of creating text-mined article summaries for each UKPMC full-text article. These will include a list of all gene/proteins, GO terms, and organism names found in each article, plus a count of each term. This is a work in progress, and requires the creation of additional tables in the full-text database for the storage of the terms linked to each article. The result will be an alternate way of quickly ascertaining the content and emphasis of an article, compared to manually reading the abstract. The display of these terms in the UKPMC interface requires that they are delivered with the metadata for articles in the full-text search web service.

Web services

Sharmila Pillai

The CiteXplore database is accessible via SOAP web services. This service is in the process of being improved and extended to provide access to the full-text UKPMC index, as described in the section above. The full-text service will, as far as possible, leverage the existing architecture of the CiteXplore web service. This has required effort to ensure search field compatibility between the two data sources as well as use of the same features of the Lucene index to ensure the same behaviour for both metadata and full-text searching. In the case of full-text searches, the service is being extended to not only include abstract, full-text links, database links, and citations, but also text-mining article summaries. The web service used to access to CiteXplore uses JAX-WS technology and is described here: www.ebi.ac.uk/citations/webservices.

The vast majority of CiteXplore use is through the web service (about 95%), serving, under normal use, over 500,000 page views a month. On some occasions, traffic can be much heavier (for example, 55 million page views in February 2009), demonstrating the robustness of the service for future growth.

FUTURE PROJECTS AND GOALS

Improvements to CiteXplore

Content growth: in collaboration with the British Library, we will identify further additional content for UKPMC, for example, theses, which will be added to CiteXplore and exposed in the UKPMC interface. We also plan to add all non-PubMed indexed UKPMC records to CiteXplore (currently about 475,000 records, although this number may change depending on which of these PubMed chooses to index).

Improving search logic: while Lucene is an excellent basis for the search functions in CiteXplore, there are several customisations and developments that would improve the search capability. These include: the use of dictionaries to expand searches to synonyms of the query term; inclusion of special search logic for author and journal searches, customisation of stemming functions; use of new sort orders; and advanced search functions.

Enhancing database and citation link data: the basic information regarding database links and citation information is displayed in CiteXplore records. However, this information should be indexed and further exploited to extend browsing and cross-database integration.

Full text capabilities

The UKPMC collaboration will continue to move towards a search and retrieval interface that incorporates new text mining features. The Literature Services group has a key role to play in this process as a provider of the search index, as well as helping to productise the results of text mining from the Rebholz-Schuhmann research group. A significant piece of work within this remit will be to form a robust pipeline for content, text mining and indexing updates within the context of UKPMC.

The Literature Services group is also a SLING partner, with a deliverable of indexing the full text of patents supplied by the European Patent Office. This will augment the UKPMC full-text research articles and provide further breadth and depth to the full-text search services.

Leveraging text mining

We plan to maximise the use of the Whatizit text-mining resource and other resources built by the Rebholz-Schuhmann research group, with related components (such as vocabularies or ontologies) to enhance search, retrieval, and browsing, and to improve integration of the literature resources with other databases at EMBL-EBI.

Team Members

Consultant

Peter Stoehr

Senior Software Engineers

Sharmila Pillai

Alan Horne*

**Indicates part of year only*

Section 3

Research in 2009

103

The Bertone Group: differentiation and development	105
The Enright Group: functional genomics and analysis of small RNA function	111
The Goldman Group: evolutionary tools for sequence analysis	117
The Le Novère Group: computational systems neurobiology	123
The Luscombe Group: genome-scale analysis of regulatory systems	129
The Rebholz-Schuhmann Group: semantic standardisation of the scientific literature	135
The Thornton Group: computational biology of proteins	141





Paul Bertone

*PhD 2005, Yale University.
At EMBL-EBI since 2005.
Joint appointments in Genome Biology and
Developmental Biology Units.*

Differentiation and development

105

INTRODUCTION

We investigate the cellular and molecular processes underlying mammalian stem cell differentiation, using a combination of experimental and computational approaches. Embryonic stem (ES) cells are similar to the transient population of self-renewing cells within the inner cell mass of the pre-implantation blastocyst (epiblast), capable of pluripotential differentiation to all specialised cell types comprising the adult organism. These cells undergo continuous self-renewal to produce identical daughter cells, or can develop into specialised progenitors and terminally differentiated cells. Each regenerative or differentiative cell division involves a decision whereby an individual stem cell remains in self-renewal or commits to a particular lineage. Pluripotent ES cells can produce lineage-specific precursors and tissue-specific stem cells, with an accompanying restriction in commitment potential. These exist *in vivo* as self-renewing multipotent progenitors localised in reservoirs within developed organs and tissues. The properties of proliferation, differentiation and lineage specialisation are fundamental to cellular diversification and growth patterning during organismal development, as well as the initiation of cellular repair processes throughout life.

A number of molecular pathways involved in embryonic development have been elucidated, including those influencing stem cell differentiation. As a result, we know of a number of key transcriptional regulators and signalling molecules that play essential roles in manifesting nuclear potency and self-renewal capacity of embryonic and tissue-specific stem cells. Despite these efforts however, only a small number of components have been identified and large-scale characterisation of cellular commitment and terminal differentiation to specific cell types remains incomplete. Our research group applies the latest high-throughput technologies to investigate the functions of key regulatory proteins and their influence on the changing transcriptome. We focus on early lineage commitment of ES cells, neural differentiation and nuclear reprogramming. The generation of large-scale data from functional genomic and proteomic experiments will help to identify and characterise the regulatory influence of key transcription factors, signalling genes and non-coding RNAs involved in early developmental pathways, leading to a more detailed understanding of the molecular mechanisms of vertebrate embryogenesis.

CURRENT PROJECTS

Genomic and proteomic technology development

Mali Salmon-Divon, Remco Loos, Diva Tommei and Heidi Dvinge, in collaboration with Vladimir Benes, EMBL Genomics Core Facility, and Kathryn Lilley, University of Cambridge

Functional genomic studies undertaken by the group have been enabled by new analytical strategies and software infrastructure, allowing us to efficiently manage and process high-throughput genomic data. We routinely use the Solexa/Illumina Genome Analyzer platform, based on solid-phase sequencing by synthesis. In this system, reactions take place within an optically transparent flow cell which can accommodate eight separate lanes where independent samples can be loaded for sequencing. Several million trace reads are generated on a single lane during each sequencing run, allowing us to perform a variety of large-scale experiments at an unprecedented level of detail.

We have implemented efficient software components for the processing and analysis of these data, which included a detailed performance assessment of the leading short-read alignment methods to obtain the maximum read placement onto the reference genome. Members of the group have developed algorithms for optimal peak detection, allowing the automated, genome-wide scanning of high-throughput sequencing data for binding site occupancy (in the case of

ChIP-seq) and transcribed sequences (for RNA-seq). We also use a variety of microarray formats, and these data are processed using a combination of open source programming tools, augmented by software developed by the group.

Post-transcriptional regulation by microRNAs constitutes a particular focus in the group, as microRNA activity plays a significant role in differentiation and development. We profile the expression of known microRNAs using specialised microarrays constructed from sugar-modified oligonucleotides termed Locked Nucleic Acids (LNA; Petersen & Wengel, 2003). The incorporation of cyclohexene nucleosides greatly increase the stability of RNA:DNA duplexes (Hakansson & Wengel, 2001; Wang *et al.*, 2001). LNA probes anneal to short complementary sequences with high affinity, such that the improved detection sensitivity of this array platform allows the expression profiling of mature microRNAs as well as the accurate discrimination of distinct microRNA species exhibiting even single base differences (Castoldi *et al.*, 2006).

Genome-wide analysis of transcription factor-mediated gene regulation

Kairi Tammoja, Mali Salmon-Divon and Heidi Dvinge, in collaboration with Boris Lenhard, Bergen Center for Computational Science

Several of our larger projects involve the genome-wide mapping of transcription factor binding sites, using the ChIP-seq method to resolve the locations of immunoselected DNA fragments via high-throughput sequencing. Such investigations are crucial to our understanding of the functional roles of key transcriptional regulators. Many studies report such genome-wide DNA association profiles as a compendium of loci, and while useful, this yields little information about the usage of these sites in different cellular contexts. To address this shortcoming, our projects incorporate comprehensive transcriptome analysis in parallel with ChIP sequencing. Through this approach we can discern many functional regulatory elements in a global fashion, along with the genes exhibiting the effects of transcription factor-mediated activation and repression.



Figure 1. A. ChIP-seq analysis of transcription factor binding (red track) reveals occupancy of promoter elements 5' of a known protein-coding target gene. Another binding site is also identified within the gene, although its putative function cannot easily be inferred. B. The same locus with differential expression information, derived from exon array time series analysis. Here we detect a shift in fluorescence levels from several exonic probe sets relative to the rest of the transcript, indicating the differential expression of an alternative splice product. The two isoforms expressed can be readily identified first as NP_149081.1, then NP_683701.2. Using this approach, we can now assign functional consequences to the binding of transcription factors to alternate promoters internal to gene loci.

One of the key advances in this area has been the use of exon-based microarrays, where individual transcript components are associated with a unique probe set. Alternate splicing of mRNA transcripts constitutes an important source of gene product diversity across multiple cell types, and the more uniform probe representation afforded by the exon array format allows both the quantitation of differential gene expression, and the identification of particular transcript isoforms present. We can therefore measure the expression of particular splice variants across different cell populations, as well as the changes in exon usage within messages expressed during specific cellular transitions.

Using this combination of technologies we are now able to resolve numerous regulatory binding events affecting not only gene expression, but splice variation across the genome. The advantages of this technique become particularly evident when transcriptional status is measured over time, as differential exon usage can then be observed in response to varying promoter occupancy (figure 1). This analysis is greatly enabled by direct programmatic access to Ensembl (Flicek *et al.*, 2008), facilitating the accurate integration of current genome annotation data.

Transcriptional regulation of neural stem cell differentiation

Diva Tommei and Kairi Tammoja, in collaboration with Steven Pollard and Austin Smith, Wellcome Trust Centre for Stem Cell Research, University of Cambridge, and Peter Dirks, University of Toronto

One of the principle cell lines we study are neural stem cells, which can either be converted from ES cells or derived from fetal forebrain tissue. In feeder- and serum-free culture conditions, ES cell self-renewal is maintained by exposure to leukaemia inhibitory factor (LIF) and bone morphogenic protein (BMP) in the culture media. Differentiation is blocked by LIF through LIF-receptor/GP130 signalling and STAT3 activation, and by BMP via SMAD-mediated Id signalling. Upon withdrawal of LIF and BMP, ES cells begin to differentiate; lineage selection is determined by specific culture conditions and the introduction of various inductive cytokines. In basal media, spontaneous ES cell differentiation is driven by the ERK signalling pathway, activated in response to autocrine production of fibroblast growth factor 4 (FGF-4).

When ES cells are differentiated in this manner, lineage selection is predominantly neuroepithelial and results in the emergence of a large fraction (50-80%) of Sox1-positive neural precursors. A reporter cell line in which the open reading frame of Sox1 is replaced with eGFP is used to monitor neuroepithelial differentiation, and the expression of a variety of other differentiation markers can be detected in the same manner. Subsequent application of fibroblast growth factor 2 (FGF-2), in combination with epithelial growth factor (EGF), supports the expansion of a clonogenic population of neural progenitor cells that over several passages acquire homogeneous morphology and immunological reactivity, and exhibit characteristic stem cell properties.

Specifically, these neural stem (NS) cells divide indefinitely in culture, exhibit a stable karyotype and retain neuronal multipotency (Glaser *et al.*, 2007). Even after greater than 100 passages, NS cells can differentiate into all three major cell types of the nervous system (neurons, astrocytes and oligodendrocytes) and demonstrate electrophysiological activity (Conti *et al.*, 2005). NS cells lose Sox1 expression but uniformly express the neuronal marker Sox2 and the intermediate filament nestin, undergo proliferation and expansion in the presence of FGF-2 and EGF, and continuously self-renew by symmetrical division.

NS cells are morphologically similar to radial glia, the developmental precursors of neurons and glial cells, and display common genetic and surface markers including RC2, Lex1, Pax6, GLAST and brain lipid binding protein (BLBP), among others (figure 2). Immunological identification and isolation of homogeneous ES and NS populations by fluorescence-activated cell sorting (FACS), using markers such as LeX/CD15 (SSEA1), is therefore efficient and robust. Preliminary microarray analysis of FACS-selected ES and NS cell populations has been performed, revealing distinct transcriptional profiles that comprise a multitude of differentially-expressed genes.

The combined ES/NS system constitutes a reproducible and well-defined model of *ex vivo* stem cell differentiation. Previous studies have reported the identification of neural stem cells from neurospheres, used as a vehicle to proliferate the stem cell population in suspension. In contrast to suspension cultures that comprise a heterogeneous population of cells – some in self-renewal and others exiting the cell cycle and committing to differentiative lineage selection – NS cells are cultured as a stably proliferating monolayer of adherent cells, permitting straightforward maintenance, immunological identification and sorting/selection.

Importantly, both ES and NS cells exhibit strong morphological and behavioural similarities to *in vivo* cell types (cells of the inner cell mass and radial glia, respectively). The progression of ES and NS cellular differentiation events is likely to be a useful *in vitro* model of early development. Thus, ES cell conversion to NS cells and differentiated neurons and glia provides an unlimited cellular resource to study cell commitment, fate choice and differentiation within the developing mammalian nervous system.

A related collection of cell lines have also been derived from human glioma multiforme tumour samples. Gliomas are driven by subpopulation of cancer stem cells which display striking similarities to normal NS cells. These glioma neural stem (GNS) cells have been isolated and expanded using the same culture conditions previously used for the

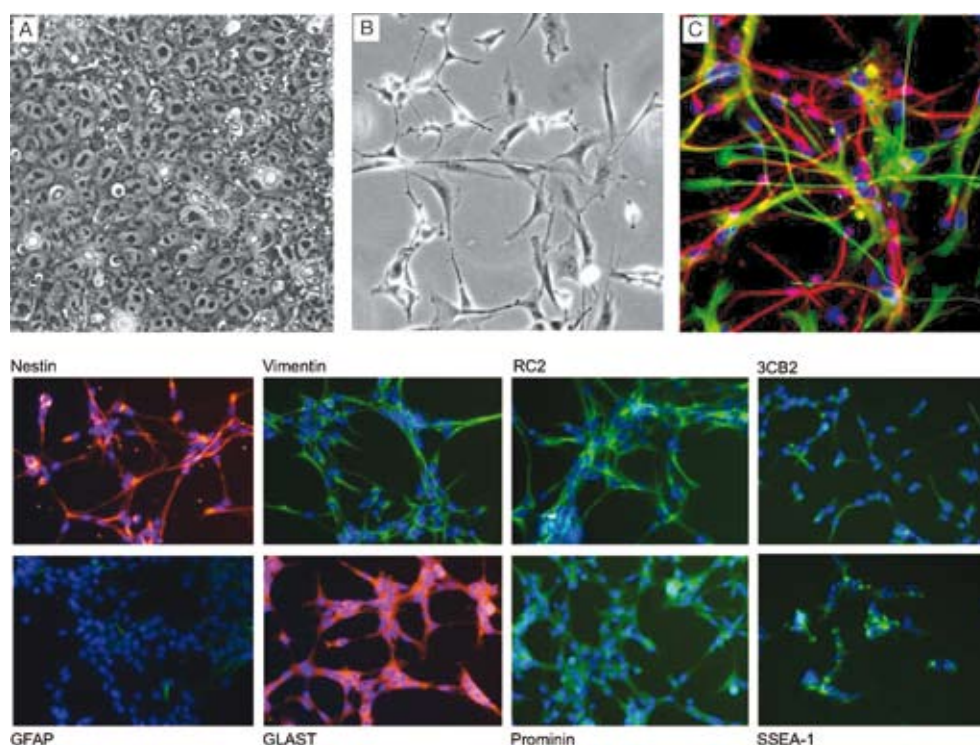


Figure 2. Top: Differentiation into neural stem (NS) cells from neural-rosette structures. A) ES cell primary culture, B, C) immunostaining for specific surface markers. Bottom: NS cells express markers characteristic of radial glia, permitting both accurate identification of differentiation stages and efficient FACS selection of homogeneous cell populations for genomic analysis. (Images: Steve Pollard, University of Cambridge; adapted from Conti et al., 2005).

establishment of NS cells. The normal and diseased counterparts are morphologically and immunohistologically indistinguishable, and yet the differentiation behaviour of the cancer stem cells is clearly aberrant.

We are now applying high-throughput transcriptome sequencing to define the comprehensive transcriptional status of stem cells during neural lineage commitment and differentiation to neurons and oligodendrocytes, an event which is positively correlated with patient survival rates in cases of glioblastoma multiforme. This involves whole-transcriptome shotgun sequencing to provide unbiased quantitation of coding and non-coding transcripts. During this project we will also analyse both GNS and NS cell populations using a combination of real-time assays and LNA microarrays to identify microRNAs whose transcriptional status is altered in the disease versus normal cell states.

Functional characterisation of non-coding RNAs in neural system development

Pär Engström, in collaboration with Ramesh Pillai, EMBL Grenoble

Eukaryotic gene expression is modulated at many layers of regulatory control. It is becoming apparent that differentiation and development involves the action of numerous regulatory non-protein coding RNAs (ncRNAs). We are therefore establishing computational resources for the study of ncRNAs, and conducting experiments to investigate their expression and function during the development of the mammalian central nervous system.

To identify ncRNAs involved in this process, we are using custom microarrays and strand-specific sequencing protocols to measure ncRNA expression in the developing mouse brain. To further prioritise ncRNAs that are likely to be functional, we make use of the large number of sequenced genomes to distinguish ncRNAs that have been conserved during evolution. In the absence of a general model for the molecular function of long ncRNAs, we are considering evolutionary conservation at three different levels: structure, sequence and expression. Genome-wide searches for structurally conserved ncRNAs are facilitated by recent algorithmic innovations (Washietl *et al.*, 2007). For RNAs that are neither conserved in sequence nor in structure, the act of transcription itself can serve a regulatory role (Martens *et al.*, 2004, Hirota *et al.*, 2008). Orthologous RNAs that lack sequence and structure conservation can be identified by making use of the extensive cDNA and EST collections available for human and mouse (Engström *et al.*, 2006).

Characterisation of novel RNAs includes targeted amplification using RACE PCR to determine precise transcription sites, followed by reciprocal overexpression and knockdown studies. In the latter case a panel of RNA-interference screens are performed; this entails the design of siRNA sequences specific to each target, introduction via transfection vectors and assessment of delivery efficiency, GFP-monitored siRNA expression and measurement of transcriptional

repression of target RNAs using quantitative real-time PCR. Additionally, we will test the hypothesis that some non-coding RNAs may be expressed as antisense targets of specific microRNAs, thereby depleting active pools of these by acting as a competitor and attenuating their associated regulatory influence.

Deep sequencing and analysis of small RNAs in embryonic development

Diva Tommei, in collaboration with Dónal O'Carroll, EMBL Monterotondo

Increasing attention has been paid to the involvement of non-coding antisense RNAs in the attenuation of message levels and/or inhibition of translation. In particular, microRNAs (miRNAs) are a class of short (~22mer) regulatory non-coding RNAs that have been shown to mediate mRNA degradation or translational inhibition through complete or partial duplex formation with target mRNA transcripts. miRNAs share functional and structural similarities to short interfering RNAs (siRNAs), and like siRNAs are processed by the dsRNA-specific ribonuclease Dicer and later engaged by PAZ/PIWI domain (PPD) proteins to confer post-transcriptional stability.

MicroRNAs are initially expressed as long non-coding sequences which undergo cleavage by the RNase III protein Drosha, which reduces the primary transcripts to 70 nucleotides (nt) precursor miRNAs having characteristic hairpin secondary structures. These are exported from the nucleus to the cytoplasm by Exportin-5, where they are digested further by Dicer to yield mature miRNAs of 19–23nt in length. In association with RNA-Induced Silencing Complex (RISC)/Argonaute proteins, miRNAs are then directed to their target mRNA transcripts and attenuate their expression levels in one of two ways: perfect complementarity between the miRNA and the mRNA sequence induces degradation of the RNA duplex; imperfect base-pairing inhibits message translation by blocking and disengaging the ribosomal complex.

A recent advance in this area has been the identification of Piwi-interacting RNAs (piRNAs), 26–30nt non-coding RNAs whose expression and function is restricted to the germline. The exact mechanism of piRNA biogenesis is unknown, although they derive from developmentally-regulated genomic clusters. These loci vary in length (generally from 1-100kb), encode numerous piRNAs (between ten and 4,500 per cluster) and are thought to produce single large transcripts that are processed to release mature piRNA sequences. The execution of piRNA activity is mediated by the Piwi sub-family of Argonaute proteins, via a pathway distinct from miRNA and siRNA function. Piwi proteins have been shown to regulate mobile genetic elements, where a large number of piRNAs are complementary to endogenous transposons. Sequence analysis of expressed piRNAs during mouse spermatogenesis, as well genetic studies of the mouse Piwi proteins Mili and Miwi2, reveal a similar function for mouse piRNAs in regulating transposable elements. While the functions of transposon-related piRNAs is appreciated, those of non-transposable element-related piRNAs (ntr-piRNAs) is not yet clear.

In collaboration with the O'Carroll group at EMBL Monterotondo, we are studying the functions of these small RNA regulators in mouse germline development using a combination of experimental and computational approaches. This involves deep sequencing of size-exclusion RNA libraries for the detection and quantitation of known and novel RNA species, followed by in-depth molecular characterisation. Using the Solexa platform we are able to ascertain small RNA expression with a high degree of precision; previously annotated RNAs are identified through alignments to miRbase and an internal piRNA database, while unknown transcripts are subjected to secondary structure predictions to determine if they are likely to adopt favourable energy conformations. This approach has enabled us to generate numerous candidates for subsequent single molecule assays. Through extensive experimental and bioinformatic investigation, we wish to gain a better understanding of the functional roles of small RNA regulators in early embryogenesis and tissue differentiation.

Computational resources for studying non-coding RNA

Pär Engström, in collaboration with John Mattick, University of Queensland

As a foundation for various non-coding RNA studies, we are building a generic computational framework for identifying ncRNAs in sequence data and cataloguing them. This involves developing bioinformatic approaches to integrate RNA-related data from different sources, perform quality controls and accurately distinguish ncRNA from mRNA. In designing and implementing these methods we are building on tools previously developed by group members for handling transcript sequence data (Engström *et al.*, 2006). Our studies are focused on human and mouse, but we aim to make our computational tools sufficiently generic to allow their application to any animal genome. Together with John Mattick's research group, who maintain the RNAdb database of mammalian ncRNAs (Pang *et al.*, 2007), we are producing a comprehensive and regularly updated ncRNA sequence collection to be made available to the RNA research community.

The problem of distinguishing ncRNA from mRNA has not been satisfactorily solved and, as a result, reference transcript collections contain many ncRNAs mistakenly annotated as mRNAs (Clamp *et al.*, 2007). While several highly accurate methods for mRNA/ncRNA discrimination have been described (Frith *et al.*, 2006, Liu *et al.*, 2006, Kong *et al.*, 2007, Lin *et al.*, 2008), there is no publicly available implementation of these methods that can be easily deployed

for whole-transcriptome analysis. We will therefore be implementing this as part of our computational framework.

FUTURE PROJECTS AND GOALS

A long-term goal of this work is to elucidate accurate models of stem cell differentiation and lineage commitment at various biological levels. Despite the importance of transcription factors and the interaction of co-factor proteins on the repression and activation of genes, eukaryotic cells utilise many layers of regulatory control. These range from histone acetylation and methylation events affecting chromatin accessibility, variations in transcript splicing producing alternate isoforms in certain cell types or conditions, the attenuation of message levels and/or inhibition of translation by antisense RNAs, and myriad post-translational modifications affecting protein function and subcellular localisation. Computational approaches will be vital for the analysis and integration of these data in context with existing knowledge.

We eventually wish to characterise the complex interaction of signalling pathways, gene regulation by key transcription factors and non-coding RNAs, and chromatin modifications that function in concert to induce distinct morphological and physiological outcomes. A first step in the process of system-level modelling is the construction of regulatory networks from time-resolved gene expression profiles. Such an approach can be applied to data generated from the projects described above to build regulatory networks from experimental results, augmented by existing information from external resources. Using this approach, we can examine changes in network topology and gene expression patterns in response to permutations of the system. Linked to biological data from many sources, this will become a powerful framework for exploring the biological activities and system-wide impact of transcriptional and translational regulators at various stages of cell differentiation.

Group Members

Staff Member

Mali Salmon-Divon

Postdoctoral Fellows

Pär Engström

Remco Loos

PhD Students

Heidi Dvinge

Myrto Areti Kostadima*

Tamara Steijger*

Diva Tommei

Visitors

Mia Rööm

Yoshihiro Taguchi

Kairi Tammoja

* Indicates part of the year only

Publications

2008

Loos, R., *et al.* (2008). Descriptive complexity of splicing systems.

International Journal of Foundations of Computer Science, 19, 813-826

Loos, R. & Ogihara, M. (2008). Time and Space Complexity for Splicing Systems. *Theory of Computing Systems*, 1-16

2009

Fredman, D., *et al.* (2009). Web-based tools and approaches to study long-range gene regulation in Metazoa. *Brief Funct. Genomic. Proteomic*, 8, 231-242

Johansson, S., *et al.* (2009). Probing natural killer cell education by Ly49 receptor expression analysis and computational modelling in single MHC class I mice. *PLoS ONE*, 4, e6046

Tan, D.J.L., *et al.* (2009). Mapping

organelle proteins and protein complexes in *Drosophila melanogaster*. *J. Proteome Res.*, 8, 2667-2678

Other EMBL publications

Engström, P.G., *et al.* (2006).

Complex loci in human and mouse genomes. *PLoS Genet.*, 2, e47

Flicek, P., *et al.* (2008). Ensembl 2008. *Nucleic Acids Res.*, 36, D707-714

Other publications

Castoldi, M., *et al.* (2006). A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). *RNA*, 12, 913-20

Clamp, M., *et al.* (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA*, 104, 19428-19433

Conti, L., *et al.* (2005). Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol.*, 3, 1596-1606

Frith, M.C., *et al.* (2006). Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol.*, 3, 40-48

Glaser, T., *et al.* (2007). Tripotential differentiation of adherently expandable neural stem (NS) cells. *PLoS ONE*, 2, e298

Hakansson, A.E. & Wengel, J. (2001). The adenine derivative of alpha-L-LNA (alpha-L-ribo configured locked nucleic acid): synthesis and high-affinity hybridization towards DNA, RNA, LNA and alpha-L-LNA complementarity sequences. *Bioorg. Med. Chem. Lett.*, 11, 935-938

Hirota, K., *et al.* (2008). Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature*, 456, 130-134

Lin, M.F., *et al.* (2008). Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. *PLoS Comput Biol.*, 4, e1000067

Liu, J., *et al.* (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet.*, 2, e29

Kong, L., *et al.* (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, 35, W345-349

Martens, J.A., *et al.* (2004). Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature*, 429, 510-511

Pang, K. C., *et al.* (2007). RNAdb 2.0 – an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, 35, D178-182

Petersen, M. & Wengel, J. (2003). LNA: a versatile tool for therapeutics and genomics. *Trends Biotechnol.*, 21, 74-81

Wang, J., *et al.* (2001). Cyclohexene nucleic acids (CeNA) form stable duplexes with RNA and induce RNase H activity. *Nucleosides Nucleotides Nucleic Acids*, 20, 785-788

Washietl, S., *et al.* (2007). Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.*, 17, 852-864



Anton Enright

PhD 2003, University of Cambridge.
Postdoctoral research, Memorial Sloan-Kettering Cancer Center, New York.
Junior Investigator at the Wellcome Trust Sanger Institute.
At EMBL-EBI since 2008.

Functional genomics and analysis of small RNA function

111

INTRODUCTION

Complete genome sequencing projects are generating enormous amounts of data. Although progress has been rapid, a significant proportion of genes in any given genome are either not annotated or possess a poorly characterised function. The goal of our group is to predict and describe the functions of genes, proteins, and in particular, regulatory RNAs and their interactions in living organisms. Regulatory RNAs have recently entered the limelight as the roles of a number of novel classes of non-coding RNAs have been uncovered.

Our work is computational and involves the development of algorithms, protocols and datasets for functional genomics. Our research currently focuses on determining the functions of regulatory RNAs. We are also interested in analysis of biological networks, protein-protein interactions, clustering algorithms and visualisation techniques. We collaborate extensively with experimental laboratories on both the commissioning of experiments and analysis of experimental data. Some laboratory members take advantage of these close collaborations to gain hands-on experience in the wet lab and perform relevant experiments to support their computational projects.

MICRORNA TARGET PREDICTION AND ANALYSIS

Nenad Bartonicek, Stijn van Dongen

The discovery of widespread translational regulation by microRNAs (miRNAs) highlights the enormous diversity and complexity of gene regulation in living systems and the need for computational techniques to help understand these systems. Regulation by miRNAs involves the binding of a mature miRNA (18-24nt) to the mRNA of a target gene. This binding event is assisted by a number of protein complexes and involves complementarity between the miRNA and its target (usually in the 3'UTR). This binding has a number of effects on the target mRNA including translational repression and mRNA destabilisation. A key specificity determinant of the binding event appears to be the 5' end of the miRNA, the so-called 'seed region'.

We developed the miRanda algorithm for miRNA target detection in collaboration with the Computational Biology Center at Memorial Sloan-Kettering Cancer Center in New York. We predict large-scale miRNA-target networks for mammalian, fish and insect genomes using the miRanda algorithm and undertake cross-species sequence analysis as part of the miRBase database. Our target predictions for miRNAs are now available through our MicroCosm web resource and database (www.ebi.ac.uk/enright-srv/microcosm/). Our lab will continue to develop and improve methods for computational detection of miRNA target sites to investigate other possible aspects of miRNA target specificity, including sequence and structural motifs.

In particular we are focused on the development of a new version of miRanda that uses a large set of features for target prediction including: 3'UTR position, structural accessibility of the UTR and proximal motifs. We are building control sets of miRNA targets validated by our collaborators and are assessing the usefulness of approaches such as Support Vector Machine (SVM) algorithms to accurately predict miRNA targets from larger feature sets.

SYLAMER – DETECTION OF MIRNA SIGNATURES FROM GENOME-WIDE STUDIES

Cei Abreu-Goodger, Nenad Bartonicek, Stijn van Dongen

Purely computational methods for miRNA target prediction perform well but suffer from over-prediction. To combat this we have developed methods that can integrate other sources of experimental information for detection of miRNA

signatures and candidate regulatory targets. Examples of this include mRNA expression data, proteomics data or HITS-CLIP. In particular mRNA expression data is useful as the technology is advanced, easy to use and is both practical and inexpensive. Although some miRNA regulatory effects do not appear to be observable at the mRNA level it seems that for the most part, miRNA–target gene binding is observable as a significant shift in mRNA expression of the target gene. Given an experiment where a miRNA is perturbed (e.g. knockout, overexpression or knockdown) we can assess the effect of that miRNA on global mRNA expression levels, query whether this is a significant and direct effect and also identify likely target genes whose expression levels have responded appropriately.

We have developed Sylamer, a new system for finding significantly over or under-represented sequences according to a sorted gene list. Analysis of over-represented features in lists of genes is a powerful tool for associating function with biological effects. Instead of using a single cut-off and thus a single gene list, gene set enrichment analysis uses all the genes, ranked according to how they change during the experiment. This approach obviates the need for cut-offs, instead searching for coordinated shifts in complete pathways or gene sets of biological interest, even if many individual genes might not be at the top of the ranked gene list. Sylamer rapidly assesses over- and under-representation of nucleotide ‘words’ of specific length in ranked gene lists. Using multiple cut-offs, it determines whether each word is more abundant at one end of the list than expected when compared to the rest, calculating significance using hypergeometric statistics. The method takes into account compositional biases in 3'UTRs and multiple testing.

In miRNA knockout experiments, transcripts that are actively downregulated by an miRNA will be upregulated in the knockout and shifted toward the top of the gene list as determined by differential expression. It is to be expected that leading subsets of the gene list are enriched in transcripts that are *in vivo* targets of this miRNA. Sylamer can be used for fast verification and quantification of this hypothesis by gauging the significance of the enrichment p-value of seed matches relative to background p-values of all other words. An intuitive way to visualise the results is to generate a landscape plot showing the associated log-transformed p-values for each word. Over- and under-representation are plotted on the positive and negative y-axis respectively (figure 1). We have validated the Sylamer algorithm on a number of datasets to detect miRNA effects, their extent and to predict likely target genes.

The Sylamer algorithm was designed to be fast and efficient and is able to process a genome-wide dataset for hundreds of miRNA signatures in seconds. The algorithm is freely available as a stand-alone package (www.ebi.ac.uk/enright/sylamer) and we are also constructing a fully-featured online analysis resource to make this approach easily accessible to the community.

SMALL RNA GENOMICS

Harpreet Saini

A large proportion of miRNAs are intergenic and encoded on their own non-coding transcript, in contrast to those which lie within the introns of protein-coding genes. For the majority of intergenic miRNAs little or nothing is known about the host transcript. It is desirable to identify the boundaries of this transcript and its regulatory features, in particular for knockout studies and prediction of miRNA transcriptional activators. We have developed a genomic pipe-

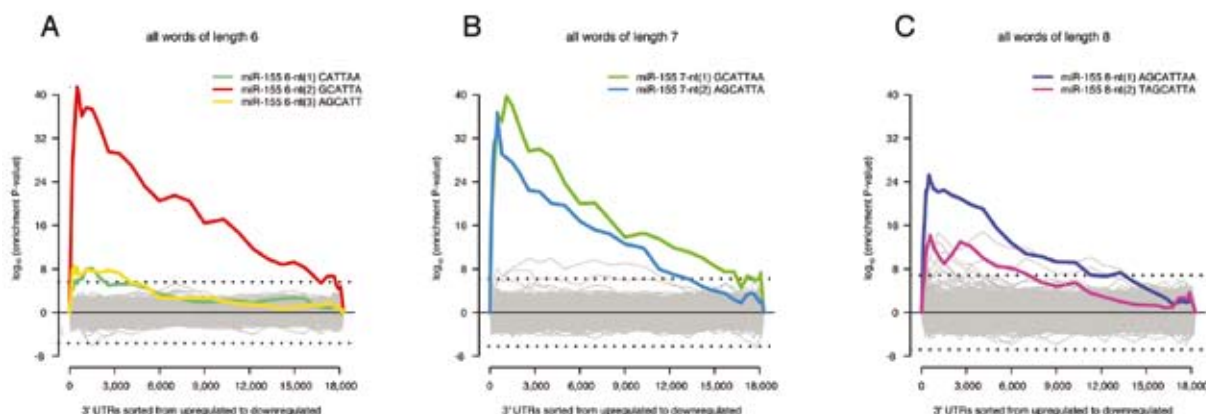


Figure 1. Sylamer results for the miR-155 microRNA in mouse TH1 cells. A clear signal is observed in gene expression data for the seed region of the miR-155 miRNA in wild-type versus knockout samples. The three panels show 6, 7 and 8nt motifs respectively (panels A–C) with miR-155 seed signals coloured. Most other miRNA signatures show no significant enrichment across the gene list (grey lines) as expected.

line for the analysis of miRNA transcripts and prediction of their boundaries and regulatory features. This resource is available as part of the miRBase database (miRBase::Genomics; www.mirbase.org/genomics.shtml). The pipeline integrates tightly with the Ensembl database and assembles information surrounding intergenic miRNAs such as EST/cDNA matches, TSS, PolyA, DiTag, 5' Cage Tag and CpG islands. This information is used to assess the likely boundaries of miRNA containing non-coding transcripts, and upstream regulatory regions are scanned for potential transcriptional activators. Additionally, we integrate variation information from public sources and highlight the location of significant SNPs in each transcript. These transcript annotations are available to the public through DAS sources that can be visualised using the Ensembl genome browser (www.ebi.ac.uk/enright-srv/microcosm/genomics/).

We also work on the prediction and analysis of other regulatory RNAs including piwi-associated RNAs (piRNAs) and small non-coding RNAs (sncRNAs) in bacteria. Part of this work involves prediction of the transcriptional units of common RNAs and their upstream regulatory factors. We are also interested in the evolution of regulatory RNAs and developing phylogenetic techniques appropriate for short non-coding RNA. Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks that may help us better understand the context of RNA in cellular networks.

EVOLUTIONARY ANALYSIS OF MIRNAS

José Afonso Guerra-Assunção

Understanding the evolution of miRNAs is difficult. Small non-coding molecules do not lend themselves well to standard sequence-based phylogenetic approaches. However, even with these challenges it is possible to learn a great deal about the evolution of small RNA regulation in vertebrates. We are approaching the problem in a number of ways. Our first approach is to identify likely orthologues and paralogues of known miRNAs across large numbers of species. The miRBase database of miRNA sequences tends to focus on the species from which an miRNA was first isolated and does not always attempt to find its counterparts in diverse organisms. To this end we have developed a novel mapping strategy for identifying likely miRNA loci in multiple organisms given a query miRNA. We have mapped all miRBase miRNAs across the animal genomes available in Ensembl. We provide an online tool (MapMi) for performing this mapping or querying the results of the miRBase mapping (www.ebi.ac.uk/enright-srv/MapMi/).

For each animal miRNA we can now build phylogenetic profiles of their presence/absence/copy number across large numbers of organisms. In particular we have the ability to build alignments for initial phylogenetic analysis. We intend to identify those miRNAs which have been amplified or deleted in specific lineages and to examine syntenic relationships between miRNAs and their neighbouring genes (either non-coding or coding) to identify potential interactions. We will also cross-compare miRNA alignments with likely target alignments to identify possible cases of correlated evolution where changes in miRNA sequence are compensated by changes to the regulatory target. Another area of interest is the identification of polymorphisms around miRNAs or their targets that may have an effect on disease susceptibility or other traits.

STUDYING REGULATORY RNAS IN MODEL SYSTEMS

Much of our work depends on having access to large-scale biological data obtained from model organisms. Hence, we collaborate extensively with laboratories interested in the role of miRNAs and piRNAs in various systems. In particular we work with our collaborators to engineer miRNA perturbation experiments followed by genome-wide assessment of the effect on mRNA or protein levels. We have developed pipelines for the analysis of microarray and new-technology sequencing data using the R/Bioconductor framework.

These collaborations are briefly listed below together with more detailed descriptions of work published over the previous year.

- Antonio Giraldez (Yale) – The role of miRNAs in both embryonic development and muscle formation.
- Antony Rodriguez (Baylor College) – Analysis of large collections of miRNA knockouts and their effects across multiple tissues in mice.
- Duncan Odom (CRI UK, Cambridge) – Developing an integrated approach for analysis of miRNA function involving regulatory networks including transcriptional activation and repression.
- Dónal O'Carroll (EMBL Monterotondo) – Role of miRNAs in murine haematopoiesis.

STUDYING THE EFFECTS OF MIRNAS IN MOUSE NEURONAL DEVELOPMENT

Sergei Manakov with Seth Grant (Wellcome Trust Sanger Institute)

We study the roles of miRNAs in the development and functioning of neurons in mice. We have performed extensive miRNA and mRNA profiling of primary mouse neurons harvested from embryos and grown in culture. This work has identified three major classes of miRNAs: 1) those which remain highly expressed at steady-state levels; 2) those which

are expressed early; and 3) those which are expressed later as the neurons begin to form connections and become active. We are evaluating the function of these miRNAs through perturbation experiments involving overexpression and knockdown experiments. We aim to dissect the role of miRNAs individually or in concert through analysis of the phenotypic consequences of these perturbations. Using large-scale mRNA expression analysis and Sylamer we will investigate likely regulatory targets of miRNAs and their function in neuronal development.

THE ROLE OF MIR-96 IN DEAFNESS

Cei Abreu-Goodger, Stijn van Dongen with Morag Lewis and Karen Steel, Wellcome Trust Sanger Institute

Recently we have worked with Karen Steel to assess the effects of a single nucleotide mutation in the miR-96 miRNA. This mutant was identified in an ENU screen for mouse deafness. The mutation lies in the key specificity determining region of miR-96. We utilised mRNA profiling data comparing inner ear from wild-type and mutant mice to detect significant shifts in mRNA level. Using the Sylamer algorithm we identified a significant and strong miRNA signature in these genes which was specific to miR-96. The most upregulated genes were enriched in motifs corresponding to the wild-type miR-96 and correspond to genes which had previously been under miRNA control in the wild-type, but which have been de-repressed in the mutant. Additionally, we identified a strong signature for a sequence corresponding to the mutant miR-96 miRNA in the genes that are downregulated in the mutant. This corresponds to novel regulatory targets being acquired by the mutant molecule. The observed phenotype is hence likely to be a combination of both a loss-of-function and a gain-of-function. However, a similar mutation identified recently in a human population at a different position in the miRNA would appear to indicate that the primary effect is the loss of wild-type repression by miR-96.

THE ROLE OF THE MIR-302 CLUSTER IN GERM CELL TUMOURS

Harpreet Saini with Matthew Murray and Nicholas Coleman, Department of Pathology, University of Cambridge

We are working with clinicians to better understand the role of miRNAs in paediatric malignant germ cell tumours (GCTs). Malignant GCTs have a common origin from the primordial germ cell, regardless of patient age. In this study, we have investigated the patterns and consequences of miRNA expression across paediatric malignant GCT samples and their subtypes. To date, there is no published miRNA profiling data for GCTs of paediatric patients. This represents a first study where Sylamer analysis clearly demonstrated that miR-302 and miR-371 clusters, which share an identical 2–7nt seed region, have a fundamental role in the pathogenesis of malignant GCTs by down-regulating functionally significant target genes. As both miRNAs clusters are embryonic stem cell-specific pluripotency markers, this raises the possibility that the expression of miR-302 and miR-371 clusters may present the persistence of an embryonic pattern of miRNA expression, which is absent in normal tissues. We have also identified sets of differentially expressed miRNAs distinguishing between GCT subtypes, such as yolk sac tumours (YSTs) from seminomas and intracranial from extracranial seminomas. From our analysis it appears that in paediatric malignant GCTs, YSTs and seminomas, the downregulated genes mediate cellular processes important in oncogenesis and malignant progression.

CLASSIFICATION AND CLUSTERING OF BIOLOGICAL DATA

Stijn van Dongen

Cluster analysis encompasses a broad category of unsupervised classification methods which are a cornerstone of many large-scale bioinformatics learning and mining techniques. Over the past decade the methodology called network clustering has been shown to be particularly effective. Networks form a natural framework for encoding and analysing objects such as genes or proteins and the links between these objects. Any simple numerical measurement can be encoded as the strength of a link. Examples are the correlation between two gene expression profiles, sequence similarity between two proteins expressed as an E-value, the direct-contact association of protein–protein interactions, and many more. In a network the local connectivity structure is highly informative and utilising this topology leads to increased predictive power.

In our lab we developed the Markov Cluster algorithm (MCL; <http://micans.org/mcl>), an elegant, fast, and highly scalable approach for network clustering (figure 2). It is widely used for protein family classification based on sequence similarity, such as the Ensembl Compara gene families. We collaborate with Ensembl Compara to further improve, scale and fine-tune MCL. The method is also seeing steadily increasing use in the analysis of gene expression data, and we have developed robust protocols in support of this. MCL is part of a highly integrated network analysis workbench. It has a rich set of tools for generating and analysing multi-level clusterings, and additionally provides scalable tools for the computation of network attributes such as betweenness centrality, diameter, eccentricity, shortest paths and clustering coefficients.

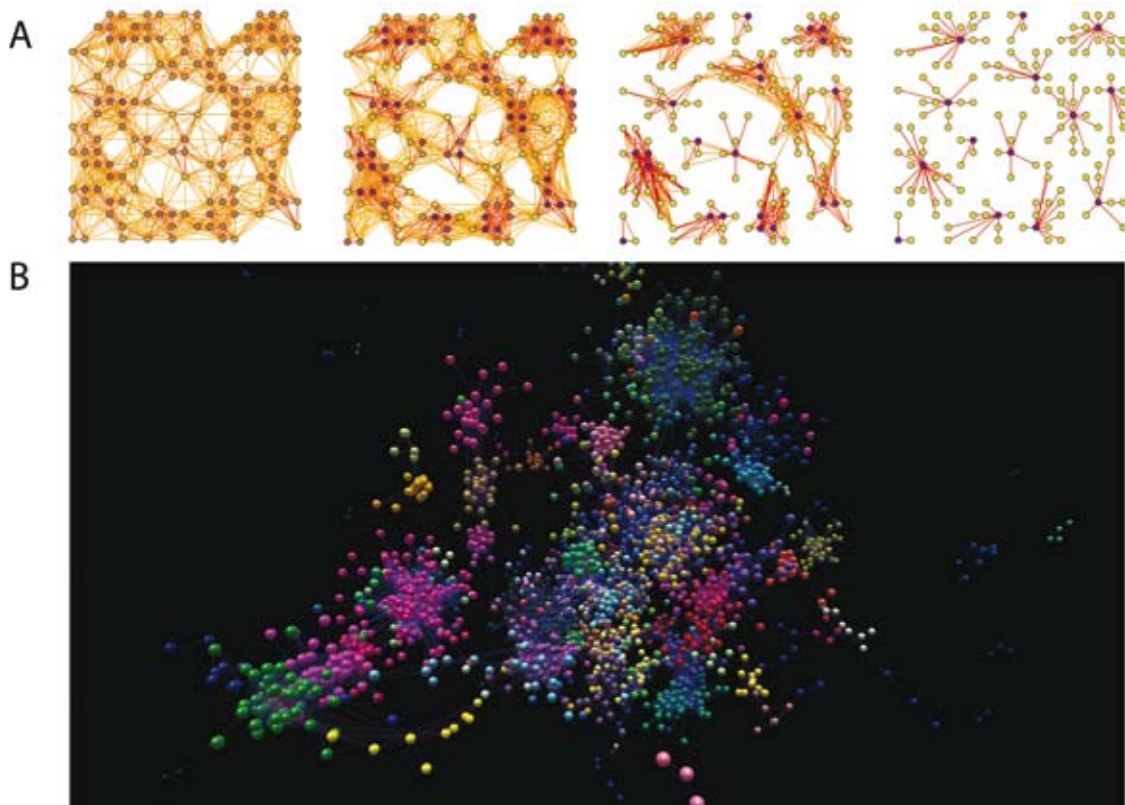


Figure 2. A. The MCL graph clustering algorithm alternates dissipation and reinforcement steps. This process is shaped by the connectivity structure in the initial network and leads eventually to a segmented set of smaller networks, interpreted as a clustering of the input. B. A network of the mouse transcriptome linked according to degree of co-expression correlation, clustered using MCL within the BioLayout 3D visualisation system.

VISUALISATION OF LARGE BIOLOGICAL NETWORKS

In collaboration with Tom Freeman at the Roslin Institute, Edinburgh

We use visualisation tools to combine our ideas and algorithms for graph-based clustering and analysis of biological data. One of our methods, BioLayout, is an integrated network visualisation tool that uses the OpenGL 3D system for fast display of complex graphs. Using OpenGL allows us to take advantage of the rapid improvement of graphics card technology. The method allows immersive 3-dimensional browsing of very large biological networks. For smaller networks there are a variety of tools available (e.g. Cytoscape), however large networks (>5000 nodes) tend to be difficult to display and interact with using standard tools. The current version, BioLayout Express 3D (www.bioblayout.org/), is available and integrates this 3D visualisation framework with MCL-based clustering and data mining of annotations. The method can display graphs with tens of thousands of nodes and hundreds of thousands of edges (figure 2B). We have tested this approach using large-scale gene expression data.

FUTURE PROJECTS AND GOALS

Our long-term goal is to combine regulatory RNA target prediction, secondary effects and upstream regulation into complex regulatory networks. We hope that by building these integrated networks we will be able to place miRNAs into a functional context that will help us to better understand the function and importance of these regulatory molecules. We are extremely interested in the evolution of regulatory RNAs and in developing phylogenetic techniques appropriate for short non-coding RNA. We will continue to build strong links with experimental laboratories working on miRNAs in different systems. In particular such work allows us to build better datasets with which to train and validate our computational approaches. The use of visualisation techniques to assist with the interpretation and display of complex multi-dimensional data will continue to be an important parallel aspect of our work.

Group Members**Senior Software Engineer**

Stijn van Dongen

Postdoctoral Fellows

Cei Abreu-Goodger

Harpreet Saini

PhD Students

Nenad Bartonicek*

José Afonso Guerra-Assunção*

Sergei Manakov

Visitor

Mat Davis

* Indicates part of the year only

Publications**2008**

Saini, H.K., *et al.* (2008). Annotation of mammalian primary microRNAs. *BMC Genomics*, 9, 564

van Dongen, S., *et al.* (2008). Detecting microRNA binding and

siRNA off-target effects from expression data. *Nat. Methods*, 5, 1023-1025

2009

Lewis, M.A., *et al.* (2009). An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat. Genet.*, 41, 614-618

Manakov, S.A., *et al.* (2009). Reciprocal regulation of microRNA and mRNA profiles in neuronal development and synapse formation. *BMC Genomics*, 10, 419-420

Mishima, Y., *et al.* (2009). Zebrafish miR-1 and miR-133 shape muscle gene expression and regulate sarcomeric actin organization. *Genes Dev.*, 23, 619-632

Other publications

Griffiths-Jones, S., *et al.* (2009). miR-Base: microRNA sequences, targets and gene nomenclature. *Nucleic*

Acids Res., 34, 140-144

Saini, H.K., Griffiths-Jones, S. & Enright, A.J. (2008). Genomic analysis of human microRNA transcripts. *Proc. Natl Acad. Sci. USA*, 45, 17719-24

Rodriguez, A., *et al.* (2007). Requirement of bic/microRNA-155 for normal immune function. *Science*, 316, 608-611

Giraldez, A.J., *et al.* (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312, 75-79

Freeman, T.C., *et al.* (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.*, 10, 2032-2042

Nick Goldman

*PhD 1992, University of Cambridge.
Postdoctoral work at National Institute for Medical Research, London, and University of Cambridge.
Wellcome Trust Senior Fellow 1995-2006.
At EMBL-EBI since 2002.*



Evolutionary tools for sequence analysis

117

INTRODUCTION

Research in the Goldman group concentrates on methods of data analysis that use evolutionary information in sequence data and phylogenies to infer the history of living organisms, to describe and understand processes of evolution, and to make predictions about the function of genomic sequence. The group maintains a good balance between phylogenetic methodology development and the use of such techniques, focusing on comparative genomics and the bulk analysis of biological sequence data. Continued fruitful collaborations with major sequencing consortia provide the essential state-of-the-art data and challenges to inspire and confront these new methods of sequence analysis. Intra-group collaborations between members involved in theoretical development and those who carry out comparative analysis of genomic data remain a stimulating source of inspiration in all of our research areas.

The group has traditionally been strong in examining the theoretical foundations of phylogenetic reconstruction and analysis. In 2009, the group has confirmed its growing strength in analysing biological data, helping bring new insight in the evolution of organisms ranging from bacteria to human, while still developing data analysis theory. Our aim is to continue to increase our understanding of the process of evolution and to provide new tools to elucidate the changing function of biological molecules.

PHYLOGENETIC METHODOLOGY

Multiple sequence alignment

Ari Löytynoja, Nick Goldman

We have developed webPRANK, an easy-to-use web interface to the PRANK phylogeny-aware sequence alignment algorithm (Löytynoja & Goldman, 2005). In addition to standard DNA and amino acid alignments, webPRANK can align protein-coding DNA data as codon or amino acid sequences and then back-translate the resulting alignment to DNA. It can also align DNA sequences using evolutionary models of sequence structure (such as fast versus slowly evolving non-coding regions, or non-coding versus coding regions) and infer the sequence structure along the alignment (Löytynoja & Goldman, 2008a). webPRANK includes a powerful alignment browser with features similar to those found in our PRANKSTER stand-alone program. The browser allows for visual inspection of the results with site-wise estimates of alignment reliability and post-processing of results by removal of the most uncertain alignment sites (figure 1). The webPRANK server can be found at www.ebi.ac.uk/goldman-srv/webprank.

The phylogeny-aware sequence alignment algorithm implemented in the PRANK software package was shown to improve the alignment of sequences with insertions (Löytynoja & Goldman, 2008b). The algorithm is, however, greedy and may perform poorly if the phylogeny of sequences is incorrect or changes across the sites. As it relies on a correct phylogeny of the sequences, the original algorithm's performance may paradoxically suffer from increasingly dense sampling of sequences (due to recombination and incomplete lineage sorting which alter the underlying phylogeny across the sites). The modelling of sequences as graphs, where nodes represent characters and edges connect adjacent characters, allows a more flexible description of the uncertainty of the site presence/absence at ancestral sequences. Furthermore, the graph edges can be described with a probabilistic model that accounts for their phylogenetic history, and thus allow a less greedy inference of insertion/deletion events. We are thus currently re-implementing our phylogeny-aware sequence alignment as a graph-based alignment, which should largely resolve the problem.



Figure 1. The TSPAN6 dataset is aligned using the webPRANK alignment server and the result is displayed in a web browser window. In addition to automatic colour coding, display of the evolutionary tree and horizontal scrolling, the alignment browser allows for post-processing of the results. Here, alignment columns with low reliability score (lighter shades in the track at the bottom) are selected and shaded light grey. The filtered set of columns can be exported in several different alignment formats for further analyses.

Assessing multiple patterns of heterogeneity in phylogenetic models of evolution

Samuel Blanquart, Nick Goldman

Mixture models in phylogenetics make the assumption that a given (Markovian) process of substitution persists throughout a sequence site's history. It has, however, been shown that this assumption does not hold systematically. For example, site rates in the Rates Across Sites model (RAS; Yang, 1994) are not always constant over time. They often vary over the site's history, e.g. switching from fast to slow. This is denoted as Single Site Variation (SSV) of rates, or heterotachy (Galtier, 2001). Other works have demonstrated that SSV phenomena also apply to other features of molecular evolution, for example biochemical constraints on amino acid sites (e.g. hydrophilic/hydrophobic SSV, Holmes & Rubin, 2002; Blackburne *et al.*, 2008), or on coding DNA sequences (Whelan, 2008), introducing more complex SSV patterns. This also concerns the selective pressures (dN/dS ratio) applied at the codon level (positive/negative selection SSV; Guindon *et al.*, 2004).

The previously mentioned studies investigating occurrences of SSV phenomena during molecular evolution all use the Markov-Modulated Markov (MMM) formalism. An MMM substitution process involves a set of 'classical' stochastic substitution processes, plus an additional stochastic process allowing switches from each of the classical substitution processes to the others (Galtier, 2001). Our work in this area focuses on the design of a general framework for SSV models involving MMM formalisms. This will generalise the more sophisticated approach of Whelan (2008) by allowing virtually all parameters of a classical phylogenetic model (rates, stationary probabilities, exchange rates etc.) to exhibit SSV behaviour. While the new model will first be applied to coding DNA sequences with the aim of reproducing the results and of improving the fit of the Whelan model, future applications will investigate SSV phenomena for amino acid and codon sequences, and provide new model combinations which have never been tested before.

Comparative prediction of protein-coding genes

Stefan Washietl, Nick Goldman

The detection of protein-coding genes in genomic DNA is a classical problem in computational biology. Using machine learning techniques, sophisticated models of genes have been built that can be used to annotate whole genomes. However, new types of high-throughput data, such as genome-wide transcription maps and massive comparative sequencing, have led to new challenges beyond classical gene finding. Many transcripts have been found that do not overlap known or predicted genes and statistical methods are necessary to assess the coding potential of this 'black matter' transcription. Similarly, comparative sequencing has revealed a plethora of evolutionarily conserved regions without annotation. A reliable analysis of their coding potential is an essential step preceding any further analysis.

We have developed RNACode, a new program to detect coding regions in multiple sequence alignments. RNACode analyses typical evolutionary patterns such as synonymous/conservative amino acid substitutions, conservation of

reading frames and absence of stop codons. Based on a simple statistical model, a dynamic programming algorithm is used to predict locally optimal coding regions together with an intuitive p-value.

Our method is a true *ab initio* approach, as it relies on evolutionary signals only and does not require any training or machine learning steps. It can thus be applied 'out of the box' to data from all living organisms. The method yields accurate results in diverse test sets ranging from archaea to human. As a first application, we have used RNAcode to revisit the protein gene annotation in *Escherichia coli*. RNAcode not only recovered known proteins almost perfectly, it also predicted a set of 33 novel small peptides. Preliminary results from mass spectroscopy experiments confirm the existence of at least 16 of these peptides *in vivo*.

RNAcode is available as open source C-implementation that can be used for manual analysis of selected regions or in annotation pipelines of larger scale. It can be found at <http://github.com/wash/rnacode>.

GENOME EVOLUTION

Evolution of transcription regulation

Jacky Hess, Ari Löytynoja, Emeric Sevin, Martin Taylor, Nick Goldman

It has been widely speculated, but as yet not convincingly demonstrated, that changes in the regulation of gene expression underlie much of the divergence between species (King & Wilson, 1975) and the phenotypic variation evident within populations. Following our earlier work investigating a transcriptional regulatory network that controls cellular differentiation (Suzuki *et al.*, 2009), we have been studying how the regulation of gene expression has diverged between mouse and human. This work uses novel, high-quality gene expression and promoter usage data from both mouse and human primary cells, assayed over a time course following a stimulatory signal. The experimental data was generated as part of an international collaboration. Most genes show significant constraint in their expression profile, but up to 23% have significantly diverged in their regulation. We see a significant correlation between the selective pressure on protein-coding sequence and expression divergence, suggesting that adaptation in coding sequence often goes hand-in-hand with adaptation in the expression of a gene. However, several lines of evidence have led us to conclude that the majority of expression divergence is a consequence of trans regulation – changes in the upstream regulatory network that could simultaneously affect multiple genes, rather than mutations in cis-regulatory sequences such as the core promoter. Despite the apparent dominance of trans effects we have found a handful of non-coding regulatory sequences that show compelling evidence for diversifying selection with consequent impacts on cis regulation of the neighbouring gene. In the course of this analysis we have found yet more evidence for highly localised variation in mutation rates, in agreement with other recent work from the group (Taylor *et al.*, 2008; Washietl *et al.*, 2008; Semple & Taylor, 2009). As all of these results are based on data from a very well studied cell type, we are able to directly relate many of the observed differences in gene regulation to cellular and even organism-level differences between human and mouse, supporting the ideas first advanced over 30 years ago by King and Wilson.

Transcription factors (TF) are obvious effectors of trans regulation, and variation in their repertoire and sequences is thus bound to have consequences on the downstream expression of genes. In an ongoing study of TF repertoires in yeasts, we are aiming to address the role of the protein-coding components of transcriptional regulatory circuits in their evolution.

We have collected data from 17 species of hemiascomycetous yeasts comprising a total of 48 families of DNA-binding proteins. The number of TFs per genome ranges from 170 in *Ashbya gossypii* to 247 in *Debaryomyces hansenii*. Generally, TFs were found to comprise between 3.2% and 4.4% of the annotated protein-coding genes. We found the main contributors to yeast TF repertoires to be classical C2H2, and binuclear cluster zinc finger proteins, often in combination with a fungal-specific TF domain. These make up between 46% and 62% of the collected TF repertoires.

TFs often contain a highly conserved DNA binding region surrounded by fast evolving sequence which makes both alignment and subsequent phylogenetic analysis challenging. To overcome some of those difficulties, we have developed an anchored alignment approach based on PRANK (Löytynoja & Goldman, 2005) that guides the alignment based on annotated structural domains and thereby improves the alignment surrounding those. Furthermore, we are currently investigating the effects of different strategies for the inference of duplication and losses along a gene tree when the phylogenetic signal in the data is limited in order to confidently identify groups of orthologous genes for in-depth evolutionary analysis.

Although trans effects may dominate as a source of change in regulation, the evolution of cis-regulatory elements is nonetheless known to have a significant impact on phenotypic variation (Wray, 2007). While the evolutionary process of specific, well-characterised systems has been studied before, we are currently working on giving a broader view of the evolutionary dynamics of transcription factor binding sites (TFBS) in *Drosophila*, such as their gain and loss in promoters through time ('turnover').

From the 16,209 non-redundant promoters in the genome of *Drosophila melanogaster*, we produced multiple alignments with their orthologues in seven other drosophilids thanks to PRANK (Löytynoja & Goldman, 2005), using

ad hoc tree topology selection to tackle reported issues of incomplete lineage sorting. Ancestral sequences were also reconstructed using maximum likelihood methods and included in the alignments. Fifteen known TFBS motifs were mapped in the promoters (in collaboration with Jüri Reimand, BIIT, University of Tartu, Estonia), leading to a filtered set of 591,865 putative binding positions.

We assessed general constraints over the set of promoters by looking at the average substitution rates at each position relative to the transcription start site. This revealed a strong correlation with nucleotide composition, whereby regions with higher concentration of adenine (A) and tyrosine (T) bases seem to change faster. For some TFBS, we were also able to infer spatial clustering constraints on well-conserved sites, as well as inter-TFBS distance preferences. With this in mind, we are investigating the general mechanisms underlying turnover events, looking at the conservation patterns of the original and substituted sites in compensatory cases, and also estimating the local context of selection of TFBS depending on their position in the promoters.

Selective pressure analysis

Greg Jordan, Martin Taylor, Nick Goldman

Working with data from the Mammalian Genome Project led by the Broad Institute at MIT, we undertook an analysis of selective pressures in mammals using a greatly increased amount of data and resolution when compared to previous studies (figure 2). Using comparative genomics data from the Ensembl Compara database (Vilella *et al.*, 2009) and two methods from the Goldman group, namely Tim Massingham's Sitewise Likelihood Ratio (Massingham & Goldman, 2005) and Ari Löytynoja's PRANK (Löytynoja & Goldman, 2005), we; 1) generated the first ever distribution of site-wise selective pressures in mammals; 2) identified new classes of proteins subject to positive selection in mammalian clades; and 3) began to investigate the dynamics of positive and purifying selection in three mammalian sub-clades.

Of particular interest in this study has been the ability to precisely identify the location of evidence for positive selection within mammalian gene families. Previous genome-wide studies focused on identifying entire genes showing evidence for positive selection; our site-wise analysis allows such evidence to be localised to individual residues. Correlating positively-selected sites with various protein-level annotations yields new insights into the biological processes and structural motifs which are most often subject to positive selection in mammals. Gene ontology (GO; Gene Ontology Consortium, 2008) terms such as 'olfactory receptor activity' and 'electron transport' were enriched in genes under strong evolutionary constraint but with a small number of sites showing evidence for positive selection, while Pfam (Finn *et al.*, 2008) domains such as 'protein kinase domain' and 'ion transport protein' showed strong purifying constraint but a disproportionately large number of positively-selected sites. Future work along these lines will involve analysis to better understand the accuracy and sensitivity of site-wise analyses and correlating site-wise selection pressures with protein structures and population genetic datasets.

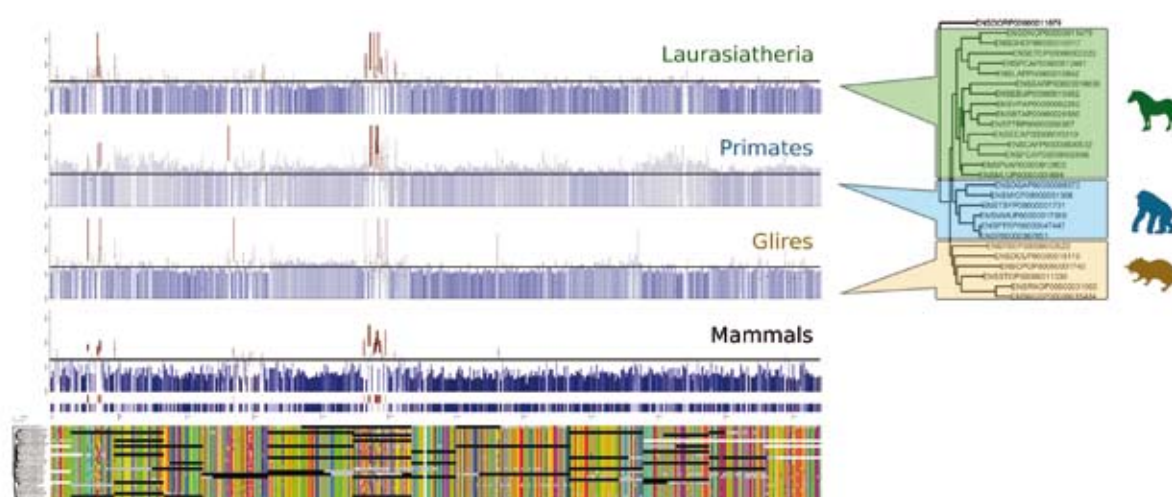


Figure 2. A long protein with very few indels, cytochrome B (ENSG00000165168, CYBB) shows strong purifying selection with two pockets of positively selected sites (bottom; Mammals). The overall dN/dS pattern is similar in the three sub-trees (from top: Laurasiatheria, Primates, Glires), although the exact sites showing evidence of positive selection vary between each sub-tree. The protein sequence alignment is coloured according to the Taylor (1997) colouring scheme except for those bases missing from low-coverage genomes (black). Vertical bars above the alignment represent the 95% confidence intervals of the site-wise dN/dS estimates, plotted on a log scale. Colour represents the strength of evidence for purifying (blue) or positive (red) selection; sites with little evidence for non-neutral selection are coloured grey. Horizontal lines are drawn at the neutral dN/dS value of 1.

As well as acting at the level of the whole organism in the evolution of species, selection is an important process in the development and progression of cancer. Neighbouring cells compete for resources, they must evade the immune system and circumvent normal controls on replication. An understanding of the cell-level selection pressures could help explain how cancers develop and go on to respond to treatments. In collaboration with the group of Janet Thornton at EMBL-EBI, we investigated the differences in selection pressure between cancer and the germline.

We found clear evidence of diversifying selection for disruptive, presumably loss-of-function, mutations in cancer cells. These were typically at sites within proteins that show higher than average conservation between species. These sites are enriched for annotated catalytic and binding sites, and are biased to the surface of proteins. In this same study we were also able to report what is, to our knowledge, the first convincing evidence of transcription-coupled repair active in metazoan somatic cells.

HIGH-THROUGHPUT SEQUENCING

Tim Massingham, Nick Goldman

Massively parallel short-read sequencing machines continue to have a large impact on biology, providing a huge increase in sequencing capacity and opening up new applications. One small institute can now have a sequencing capacity that outstrips what was available in total to the major sequencing centres at the height of the Human Genome Project, allowing sequencing to a depth required for replacing microarrays in expression studies and for cataloguing genomic structure variation.

Calling sequence from the raw output of the machines is not trivial since the errors increase dramatically as the length of the reads increases. A major contribution to these errors is phasing between sequencing cycles (the tendency of signal from the machine to be a mixture of current, past and future bases in the read) and we have previously developed statistical techniques to correct the phasing and thus improve the quality of sequence produced by the machines. These statistical techniques have been implemented in the AYB software, further development of which is now funded by a grant from the Wellcome Trust. AYB has proved particularly effective for long or difficult reads, producing six times as many error-free reads than the software distributed with the machines on a bad run of *Bordetella pertussis* and three times as many on a run of one hundred-cycle bacteriophage data. AYB can be found at www.ebi.ac.uk/goldman/AYB.

Development of AYB has led to the development of other tools for analysing short-read sequencing data, for example software for the probabilistic detection of adapter sequence, which allows the confidence in each position to be correctly down-weighted or deleted entirely from the read.

FUTURE PROJECTS AND GOALS

The study of genome evolution continues to inspire us with novel problems in phylogenetic methodology. The complex nature of the non-independence of sequence data due to their evolutionary relatedness continues to generate statistically challenging problems (Goldman & Yang, 2008) and we will continue to contribute to this theoretical field. We remain dedicated to retaining our interest in the practical applications of these methods in order to promote best practice in computational evolutionary and genomic biology, to keep in touch with the evolving needs of laboratory scientists and to continue to benefit from a supply of motivational biological questions where computational methods can help.

During 2009 we have become increasingly involved in the study of transcriptional regulation. This will continue and will include consideration of both proteins and non-coding DNA and extension of current TFBS work to vertebrates using ChIP-seq data available through external collaboration. We will continue to work with Martin Taylor, who leaves the EBI at the end of 2009 to establish an independent research group at the MRC Human Genetics Unit in Edinburgh. Building on work initiated at the EBI, he will focus on the detection of selection and mutation patterns using both evolutionary and population genetic resources.

There is no doubt that our involvement with the analysis of data arising from next-generation sequencing projects will increase. We also have high hopes for our work on graph alignments. This representation of sequences can be used to account for uncertainty in the input data (e.g. due to high insertion/deletion rates or elevated error rates from some next-generation sequencing technologies), or to capture variation within populations (e.g. for probabilistic modelling of a 'reference sequence' against which to align a population sample).

Group Members

Scientists

Ari Löytynoja
Tim Massingham
Martin Taylor

Postdoctoral Fellows

Alexander Alekseyenko*
Samuel Blanquart*
Emeric Sevin
Stefan Washietl*

Scientific Programmer

Nicolas Rodriguez

PhD Students

Jacky Hess
Greg Jordan
Fabio Pardi*

Visitors

Botond Sipos

* Indicates part of the year only

Publications

2008

Goldman, N. & Yang, Z. (2008). Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. *Philos. Trans. R. Soc. B-Biol. Sci.*, 363, 3889-3892

Löytynoja, A. & Goldman, N. (2008a). A model of evolution and structure for multiple sequence alignment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 3913-3919

Madera, M. (2008). Profile Comparer: A program for scoring and aligning profile hidden Markov models. *Bioinformatics*, 24, 2630-2631

Pain, A., et al. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, 455, 799-803

Taylor, M.S., et al. (2008). Rapidly evolving human promoter regions. *Nat. Genet.*, 40, 1262-1263

Washietl, S., et al. (2008). Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.*, 24, 583-587

2009

Bishop, C.J., et al. (2009). Assigning strains to bacterial species via the internet. *BMC Biol.*, 7, 3

Huang, G.J., et al. (2009). High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res.*, 19, 1133-1140

Löytynoja, A. & Goldman, N. (2009). Evolution. Uniting alignments and trees. *Science*, 324, 1528-1529

Minh, B.Q., et al. (2009). Budgeted phylogenetic diversity on circular split systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 22-29

San Mauro, D., et al. (2009). Experimental design in caecilian systematics: Phylogenetic information of mitochondrial genomes and nuclear rag1. *Syst. Biol.*, 58, 425-438

Semple, C.A.M. & Taylor, M.S. (2009). The structure of change. *Science*, 323, 347-348

Suzuki, H., et al. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, 41, 553-562

Other EMBL publications

The Gene Ontology Consortium (2008). The Gene Ontology Project in 2008. *Nucl. Acids Res.*, 36, D440-D444

Löytynoja, A. & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA*, 102, 10557-10562

Löytynoja, A. & Goldman, N. (2008b). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320, 1632-1635

Massingham T. & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169, 1753-1762

Vilella, A.J., et al. (2009). EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19, 327-335

Other publications

Blackburne, B.P., et al. (2008). Changing selective pressure during antigenic changes in human influenza H3. *PLoS Pathogens*, 4, e1000058

Finn, R.D., et al. (2008). The Pfam protein families database. *Nucl. Acids Res.*, 36, D281-D288

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18, 866-873

Guindon, S., et al. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl Acad. Sci. USA*, 101, 12957-12962

Holmes, I. & Rubin, G.M. (2002). An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.*, 317, 753-764

King, M.C. & Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188, 107-116

Taylor, W.R. (1997). Residual colours: a proposal for amino-chromography. *Protein Eng.*, 10, 743-746

Whelan, S. (2008). Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol. Biol. Evol.*, 25, 1683-1694

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, 8, 206-216

Yang, Z. (1994). Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39, 306-314

Nicolas Le Novère

PhD 1998, Pasteur Institute, Paris.
Postdoctoral research at the University of Cambridge.
Research fellow, CNRS, Paris.
At EMBL-EBI since 2003.



Computational systems neurobiology

123

INTRODUCTION

The Le Novère group's research interests revolve around signal transduction in neurons, ranging from the molecular structure of proteins involved in neurotransmission to signalling pathways and electrophysiology. In particular, we focus on the molecular and cellular basis of neuroadaptation in neurons of the basal ganglia. By building detailed and realistic computational models, we try to understand how neurotransmitter-receptor movement, clustering and activity, influence synaptic signalling. Downstream from the transduction machinery, we build quantitative models of the integration of signalling pathways known to mediate the effects of neurotransmitters, neuromodulators and drugs of abuse. We are particularly interested in understanding the processes of cooperativity, pathway switch and bistability.

The group provides community services that facilitate research in computational systems biology. In particular, we are leading the efforts in encoding and annotating kinetic models in chemistry and cellular biology, including the creation of standard representations, the production of databases and software development. The Systems Biology Markup Language (SBML) is designed to facilitate the exchange of biological models between different types of software. The Systems Biology Graphical Notation (SBGN) is an effort to develop a common visual notation for biochemists and modellers. Moving from the form to the content, we are also developing standards for model curation (MIRIAM, MIASE), a format for describing simulation experiments (SED-ML), and controlled vocabularies (the Systems Biology Ontology, the TErminology for the Description of DYnamics etc.) to improve the models. Finally, a model is only useful if it can be easily accessed and reused. BioModels Database is now the reference resource where scientists can store, search and retrieve published mathematical models of biological interest.

COMPUTATIONAL SYSTEMS BIOLOGY OF DENDRITIC SPINE SIGNALLING

The glutamatergic synapse is one of the main cellular components of the mammalian brain, responsible for most of the cognitive processing and also for learning and memory. It is located on a specific portion of the neuron, the dendritic spine. The spine can be seen as an independent electrical and biochemical compartment, and thus as a unit of signal treatment and integration. The glutamatergic synapse is a very complex structure. The neurotransmitter receptors are embedded in complex multi-molecular assemblies, encompassing proteins of the pre- and postsynaptic sides. Glutamate released by the presynaptic terminals activates glutamate receptors of the AMPA type, which trigger the electrical response. This electrical response in turn allows the opening of glutamate receptors of the NMDA type. Those receptors let calcium flow into the spine, which results in the activation of many signalling cascades, leading for instance to synaptic plasticity or spine remodelling. The research projects of the team are centred on various components of the signal treatment in the spine of a particular neuron, the medium-spiny neuron of the striatum, involved in the control of voluntary motion and processes of reward.

Allosteric models of proteins of the postsynaptic density

Melanie Stefan, Stuart Edelstein, Ranjita Dutta-Roy

Learning processes are thought to rely on modification of synaptic activity such as long-term potentiation (LTP) and depression (LTD). The key event regulating these processes is calcium influx through the NMDA receptor (NMDAR). In the cell, this calcium influx affects many signalling cascades, in particular through the activation of calmodulin. Calmodulin conformation and activation is affected by calcium binding. We have built a full microscopic, kinetic model by extending the framework of concerted allosteric transitions (Stefan *et al.*, 2008, Stefan *et al.*, 2009). This

model provides an explanation of the fact that low concentrations of calcium-activated calcineurin trigger LTD and high concentrations of calcium-activated calcium/calmodulin-dependent protein kinase II (CaMKII) trigger LTP, while in both cases, the effect is mediated by the activation of calmodulin. CaMKII is central to the molecular basis of memory. It is a dodecameric protein that phosphorylates a wide range of targets, including itself and the glutamate receptors. Each monomer can exist in many different states, and the enumeration of all the possible combinations is infeasible. In order to relate the structure of the enzyme to its function as a molecular memory device, we created molecular models of the complex between calcium-calmodulin and CaMKII, and models of the phosphorylated forms of the kinase. Molecular dynamic simulations were used to generate alternative structures. To understand its allosteric properties, we developed highly detailed stochastic models of the function of CaMKII and its associated proteins, such as calmodulin and NMDAR.

Modelling of AMPA receptor function

Dominic Tolle

The position and movements of neurotransmitter receptors in and around synapses influences neuronal signal processing. Moreover, it has long been known that LTP is, in part at least, due to the appearance of new neurotransmitter receptors at the postsynaptic site. We used particle-based stochastic simulations to show that thermal diffusion alone can account for the incorporation of receptors at the synaptic specialisation within the timeframe of LTP expression. Our model predicts how the system behaves under various conditions affecting the free diffusion of receptors in the membrane, such as a change in biophysical parameter values, varying spatial parameters or quantity of interacting components. Receptors accumulate rapidly at the postsynaptic density under a number of biologically observed conditions. This accumulation is controlled by the number of scaffolding proteins and their affinity for the receptors but not the characteristics of the membrane or the initial location.

Signalling pathways involved in the plasticity of striatal neurons

Lu Li, Noriko Hiroi

The projecting neurons of the striatum provide a crucial route for information transfer in the basal ganglia, involved in motor, psycho-motor and behavioural functions. Dopamine modulates the inputs from cortical glutamatergic terminals, providing a measure of the internal (hedonic) state. A protein phosphatase inhibitor, DARPP-32, has been identified as a major target for both dopamine and glutamate signalling. We have extended our model of the regulation of DARPP-32 phosphorylation and dephosphorylation (Fernandez *et al.*, 2006) to accurately account for the variety

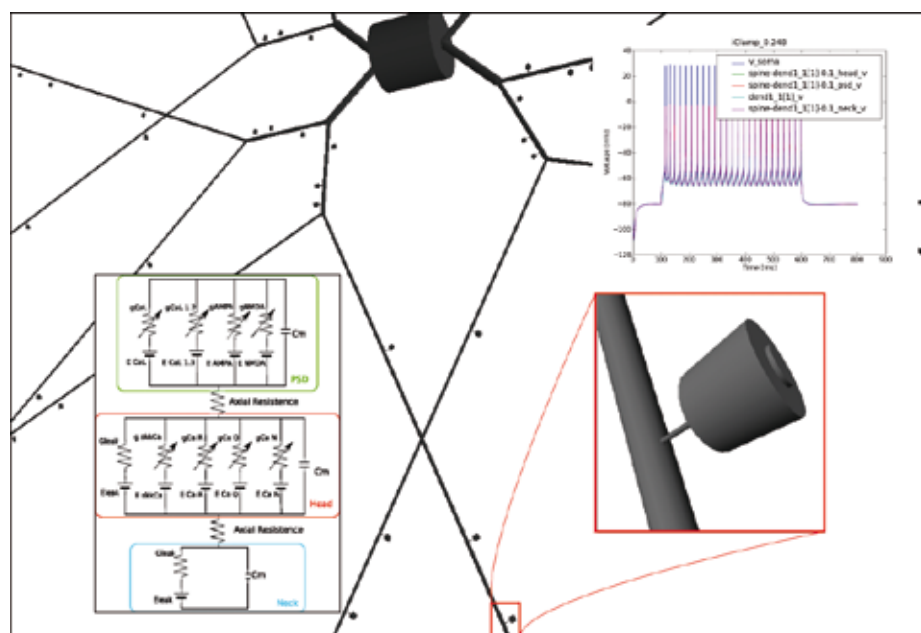


Figure 1. Model of a medium-spiny neuron of the striatum based on the cable approximation. The background image shows the cell body, the dendrites and the dendritic spines. The enlargement (bottom right) shows the structure of a spine. The electrical circuit on the left represents the electrical model of the spine. The plot on the upper right shows a train of action potentials caused by a current clamp.

of calcium modulations and to incorporate the downstream signalling through the MAPK cascade. Furthermore, a multi-compartment model has been developed to represent the range of reactions taking place in the postsynaptic density and the spine cytoplasm. Experiments are under way to verify the predictions and refine the models.

Integration of biochemical and electrical models of the striatal medium-spiny neuron

Michele Mattioni

The function of a neuron can only be understood by taking into account the influence of signalling pathways on its electrical behaviour. We are now developing a model of the entire medium-spiny neuron that incorporates kinetic descriptions of signalling pathways in each dendritic spine and the electrical behaviour of synapses, spines and neuron (figure 1).

COMPUTATIONAL SYSTEMS BIOLOGY

The practice of systems biology relies on interfaces, and in particular interfaces between the entities we study, whether molecules, pathways or cells, or interfaces between tools. If these interfaces are to be generic enough to allow all users to leverage on existing toolkits, the existence of community-developed, well-supported standards is a fundamental requirement, in addition to open resources for tools and parts. Over the last decade or so, several efforts have been launched in this direction, addressing encoding formats, ontologies and databases. Some of these are now well-established in the field and play a significant role in increasing the size and quality of quantitative models. More importantly, they have served as a catalyst to improve the collaborative nature of the computational systems biology community.

Standards of reporting (MIRIAM and MIASE)

Camille Laibe, Nick Juty, Dagmar Köhn

Most published quantitative models in biology are lost to the community because they are insufficiently characterised, which prevents them from being reused. With today's increased interest in detailed biochemical models, it was necessary to define a minimum quality standard for the encoding of these models. The Minimal Information Requested in the Annotation of Models (MIRIAM) is a set of rules for curating quantitative models of biological systems. Their application enables users to search collections of curated models with precision, quickly identify the biological phenomena that a given curated model or model constituent represents, and facilitates model reuse, model composition into large subcellular models, and format conversion. An important part of the standard concerns the controlled annotation of model components, based on Uniform Resource Identifiers (URIs). MIRIAM Resources is an online infrastructure created to enable interoperability of this annotation (www.ebi.ac.uk/miriam/). The core of this resource is a catalogue of data types, whether controlled vocabularies or primary data resources, which provides the means to generate and resolve MIRIAM URIs. The use of MIRIAM annotations by the community is still growing, and software tools have been developed that use URIs as a glue to merge models and to integrate other datasets. MIRIAM's guidelines deal mostly with the structure of the models but in order to use the models to run simulations and obtain numerical results, one needs additional information. The Minimum Information About a Simulation Experiment (MIASE) is a fledgling effort to agree upon a set of mandatory information to include with relevant publications. Both MIRIAM and MIASE are part of MIBBI, a more general effort to coordinate the development of reporting guidelines.

Ontologies in systems biology

Nick Juty, Dagmar Köhn, Camille Laibe

Whilst many controlled vocabularies exist that can be directly used to relate quantitative models to biological knowledge, there was previously no classification of the concepts themselves used in quantitative modelling. One of the goals of the Systems Biology Ontology (SBO; www.ebi.ac.uk/sbo/) is to facilitate the immediate identification of the relationship between a model component and the model structure (Chelliah *et al.*, 2009). SBO is currently made up of six different vocabularies: 1) an ontology of entities which may participate in an interaction, a process or relationship of biological significance (for example: 'enzyme' and 'ribonucleic acid'); 2) a taxonomy of the roles of reaction participants (e.g. 'catalyst', 'competitive inhibitor'); 3) a controlled vocabulary for parameter roles in quantitative models (for instance: 'forward unimolecular rate constant' and 'Michaelis constant'); 4) a list of modelling frameworks that specify how to interpret a mathematical expression (such as: 'continuous framework' or 'discrete framework'); 5) a classification of mathematical expressions used in biochemical modelling (e.g. 'mass action rate law', 'Henri-Michaelis-Menten rate law'); and 6) a catalogue of interactions (for example: 'non-covalent binding' and 'transport reaction'). The annotation of quantitative model components with SBO terms adds a layer of semantics necessary to convert models between different formalisms, to link mathematical representations of biochemical models with graphical notations such as the Systems Biology Graphical Notation (see overleaf), or semantically enriched computing formats to represent biochemical knowledge such as BioPAX. To complete SBO, which is designed to enrich model descriptions, we are developing an ontology of simulation methods (KISAO; www.ebi.ac.uk/compneur-srv/kisao/) aimed to be used with SED-ML (see below), and an ontology to characterise numerical descriptions of dynamic behaviours (TEDDY; www.ebi.ac.uk/compneur-srv/teddy/).

Formal languages to encode models and simulations

Sarah Keating, Dagmar Koehn, Nicolas Le Novère, Nicolas Rodriguez

The Systems Biology Markup Language (SBML) is an XML language designed to facilitate the exchange of biological models between different simulators. SBML is now an established standard in the field of systems biology, and is supported by several EMBL-EBI resources such as Reactome, IntAct and BioModels Database. While bringing minor corrections and clarifications to the current specification of the language (Level 2, Version 4; Hucka *et al.*, 2008), we are now working to develop the new generation of SBML. The field of computational systems biology is now so wide and diverse that a single language, supported by all tools, cannot cover every approach. SBML Level 3 will therefore be modular, with a mandatory core package and optional modules. The group is particularly working on packages to represent multi-component, multi-state species, qualitative models, space and geometry, and hierarchical modelling. We use our generic SBML editor (www.ebi.ac.uk/compneur-srv/SBMLEditor.html) as a benchmark to test possible packages and for various related projects of the group. We also provide software to convert to and from SBML. While SBML encodes the mathematical structure of the models, it does not specify how to obtain numerical results from this description. Together with simulator developers, we are creating a complementary format, the Simulation Experiment Description Markup Language (SED-ML; <http://biomodels.net/sed-ml/>; Köhn & Le Novère, 2008). A SED-ML file defines which models to simulate, how to modify them, which simulation approach to apply, how to post-process the numerical results and how to report them.

Systems Biology Graphical Notation

Nicolas Le Novère, Lu Li

Standard graphical representations have played a crucial role in science and engineering throughout the last century. Without electrical diagrams, it is very likely that our industrial society would not have evolved at the same pace. Similarly, specialised notations such as the Feynman notation or the process flow diagram were instrumental for the adoption of concepts in their fields. With the advent of systems biology, and more recently of synthetic biology, the need for precise and unambiguous graphical descriptions of biochemical processes has become more pressing. While some ideas have been advanced over the last decade, with a few detailed proposals, no actual community standard has emerged. We developed the Systems Biology Graphical Notation (SBGN; www.sbgm.org/, Le Novère *et al.*, 2009), a graphical representation crafted over several years by a community of biochemists, modellers and computer scientists. Three orthogonal and complementary languages have been created: the Process Descriptions, the Entity Relationships and the Activity Flows. These three idioms enable scientists to represent any network of biochemical interactions in a standardised way, which can then be interpreted unambiguously. The set of symbols used is limited and the grammar kept as simple as possible, to also allow its use in textbooks and education. Shared SBGN languages will foster efficient and accurate representation, storage, exchange and reuse of information on biological knowledge, e.g. signalling pathways, metabolic and gene regulatory networks, between the communities of biologists, theoreticians and computational biologists.

BioModels Database

Chen Li, Lukas Endler, Nicolas Rodriguez, Vijayalakshmi Chelliah

For computational modelling to become more widely used in biological research, modellers must be able to exchange and share their results. BioModels Database (www.ebi.ac.uk/biomodels/) is a data resource that allows modellers to store, search and retrieve published mathematical models of biological interest. Models are annotated and linked to other relevant data resources. BioModels Database accelerates computational modelling efforts by allowing researchers to leverage each others' work more directly (Endler *et al.*, 2009). It also supports improved and more accurate communication of research results by allowing journal publishers to encourage the submission of models in the same electronic format, stored in a common, publicly accessible location. Finally, the database provides examples of working models for educational purposes, allowing inexperienced modellers to find ready-to-use models for exploration. BioModels Database has been developed in collaboration with the California Institute of Technology and is now the largest database of curated models worldwide (containing more than 429 models and 39,000 reactions). This status is recognised by BioMedCentral, Nature Publishing Group and the Public Library of Science, all of which request deposition of models upon submission of manuscripts to several hundreds of journals. We regularly release new versions of the database, with new features for both users and curators.

FUTURE PROJECTS AND GOALS

In forthcoming years, the activity of the group will continue along two orthogonal directions. Our research work on modelling neuronal signalling at the level of the dendritic spine will expand to include other signalling pathways (MAPK, TrkB, PI3K) and tackle problems such as the role of scaffolding proteins or the synchronisation of calcium waves and phosphorylation gradients. Building on the growth of the BioModels Database, we will also carry out research on model composition, with the aim of improving component identification and reaction matching to build large-scale models of cellular compartments such as dendritic spines. Our involvement in developing standards and resources for systems biology will continue, with the goal of completing the puzzle of representations and ontologies







Standard specification of quantitative models	Model description	Simulation description	Simulation results description
Minimal requirements			?
Data format		SED-ML	SBRML
Ontologies			

Figure 2. Matrix of the available standards, formats and ontologies for the description of models, simulations and results.

so as to efficiently integrate the different levels of description of biochemical and cellular processes, qualitative, quantitative and experimental (figure 2).

Group Members

Visitor Scientist

Stuart Edelstein

Postdoctoral Fellows

Noriko Hiroi*

Melanie Stefan*

Software Engineers

Sarah Keating*

Camille Laibe

Chen Li

Nicolas Rodriguez (shared among several research groups)

Scientific Database Curators

Vijayalakshmi Chelliah

Lukas Endler

Nick Juty

PhD Students

Lu Li

Michele Mattioni

Dominic Tölle

Trainees

Ranjita Dutta-Roy*

Marine Dumousseau*

*Indicates part of the year only

Publications

2008

Guerlet, G., *et al.* (2008). Comparative models of P2X2 receptor support inter-subunit ATP-binding sites. *Biochem. Biophys. Res. Commun.*, 375, 405-409

Hucka M., *et al.* (2008) Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions. *Nat. Prec.*, doi:10.1038/npre.2008.2715.1

Herrgård M.J., *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology *Nat. Biotechnol.*, 26, 1155-1160

Köhn, D. & Le Novère, N. (2008) SED-ML – An XML Format for the Implementation of the MIASE Guidelines CMSB 2008. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Heiner, M. & Uhrmacher, A.M. (eds), 5307, 176-190, Springer-Verlag

Le Novère, N. (2008). Multiscale modelling of neuronal signalling. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Heiner, M. & Uhrmacher, A.M. (eds), 5307, 176-190, Springer-Verlag

2009

Chelliah, V., *et al.* (2009) Data Integration and Semantic Enrichment of Systems Biology Models and Simulations. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes

in Bioinformatics)', Paton, N.W., Missier, P. & Hedeler, C. (eds), 5647, 5-15, Springer-Verlag

Dräger, A., *et al.* (2009) SBML2LATEX: Conversion of SBML files into human-readable reports. *Bioinformatics*, 25, 1455-1456

Endler, L., *et al.* (2009) Designing and encoding models for Synthetic Biology. *J. Roy. Soc. Int.*, 6, S405-S417

Le Novère, N., *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27, 735-741

Stefan, M.I., *et al.* (2009) Computing phenomenologic Adair-Klotz constants from microscopic MWC parameters. *BMC Sys. Biol.*, 3, 68

Wolkenhauer, O., *et al.* (2009) SysBioMed report: Advancing systems biology for medical applications. *IET Sys. Biol.*, 3, 131-136

Other EMBL publications

Fernandez, E., *et al.* (2006). DARPP-32 is a robust integrator of dopamine and glutamate signals. *PLoS Comput. Biol.*, 2, e176

Stefan M.I., *et al.* (2008). An allosteric model of calmodulin explains differential activation of PP2B and CaMKII. *Proc. Natl Acad. Sci. USA*, 105, 10768-10773

Nicholas Luscombe

*PhD, 2000, University College London.
Postdoctoral work at Department of Molecular Biophysics & Biochemistry, Yale University.
At EMBL-EBI since 2005.
Joint appointment with Gene Expression Unit, EMBL Heidelberg.*



Genome-scale analysis of regulatory systems

129

INTRODUCTION

Cellular life must recognise and respond appropriately to diverse internal and external stimuli. By ensuring the correct expression of specific genes at the appropriate times, the transcriptional regulatory system plays a central role in controlling many biological processes; these range from cell cycle progression and maintenance of intracellular metabolic and physiological balance, to cellular differentiation and developmental time courses. Numerous diseases result from a breakdown in the regulatory system and a third of human developmental disorders have been attributed to dysfunctional transcription factors. Furthermore, alterations in the activity and regulatory specificity of transcription factors are now established as major sources for species diversity and evolutionary adaptation. Indeed, increased sophistication in the regulatory system appears to have been a principal requirement for the emergence of metazoan life.

Much of our basic knowledge of transcription regulation has derived from molecular biological and genetic investigations. In the past decade, the availability of genome sequences and development of new laboratory techniques has generated (and continues to generate) information describing the function and organisation of regulatory systems on an unprecedented scale. Genome-scale studies now allow us to examine the regulatory system from a whole-organism perspective; on the other hand however, observations made with these data are often unexpected and appear to complicate our view of gene expression control.

This continued flood of biological data means that many interesting questions require the application of computational methods to answer them. The strength of bioinformatics is its ability to uncover general principles providing global descriptions of entire systems. Armed with these biological data we are now poised to achieve this.

By integrating diverse data sources – from genome sequence to the results of functional genomics experiments – we study the regulatory system at a genomic scale. Since the start of the group in 2005, we have focused our interests on understanding bacterial and eukaryotic gene regulation. Below we describe some of our findings in these areas.

Our current projects include:

- examining how the metabolic system is controlled at multiple levels through the feedback activity of small molecules;
- analysing the repertoire, usage and cross-species conservation of transcription factors in the human genome;
- wet/dry collaborations to uncover the regulation governing complex organismal behaviour;
- wet/dry collaborations to understand the epigenetic control of dosage compensation in animals.

In 2010 we will continue to advance analysis techniques and our understanding of regulatory systems in microbes and higher eukaryotes. A major focus continues to be our close interactions with research groups performing genome-scale experiments.

REGULATORY NETWORKS IN ENTEROBACTERIA

Aswin Seshasayee, Inigo Martincorena, Nicholas Luscombe

The *E. coli* regulatory system

The *E. coli* K12 genome encodes about 280 transcription factors, most of which bind DNA using variants of the helix-

turn-helix motif (Luscombe *et al.*, 2000). As a long-standing model organism, *E. coli* K12 has much data regarding its regulatory circuitry. The RegulonDB database avails a manually curated list of regulatory interactions identified in molecular and genetic experiments (Salgado *et al.*, 2006), and more recently, several groups published ChIP-chip studies to identify additional targets for the transcription factors Crp and MelR (Grainger *et al.*, 2003; Grainger *et al.*, 2005). An assembly of these sources provides a network comprising over 2,000 regulatory interactions between 156 factors and 1,114 target genes. Although this is a sizeable dataset that enables us to examine the regulatory system on a genomic scale, we still lack information for 130–150 transcription factors (~46–50%) and nearly 3,000 non-transcription factor genes (~70%). Functions that particularly lack information include the regulation of cell division, lipid metabolism and cellular defence mechanisms.

Although transcription factors are most commonly classified according to their DNA binding domains, it is also possible to view them from alternative perspectives that provide additional insights into their regulatory functions. One example is the identity of the partner domain outside the DNA binding regions: 150 factors (over 50%) contain a small molecule binding domain; 25 contain a phosphorylation domain for two-component systems; and 44 comprise a DNA-binding domain only. These domains are strongly indicative of regulatory function; for example, two-component regulators generally target signalling genes, whereas sugar-binding factors predominantly regulate carbohydrate and carbon metabolism.

Regulation of small molecule metabolism

The partner domains also show how transcription factors are themselves regulated; those containing only DNA binding domains tend to be controlled at the level of transcription, whereas the others are controlled post-translationally. This raises the intriguing possibility of a series of feedback loops that regulate the metabolic system, whereby small molecules control the activity of transcription factors, which in turn control the expression of the enzymes that process these metabolites.

The importance of small molecule metabolism is highlighted by the fact that it processes all the core molecules that are required for an organism's survival. Equally important is the regulation of the metabolic system so that the correct metabolites are processed at the appropriate times, at minimal additional cost to the cell. Two well-established mechanisms for regulation are: 1) control of enzyme concentration, largely at a transcriptional level; and 2) control of enzyme activity by post-translational means. The two mechanisms differ widely in the timescales involved. Enzyme activities can change in a matter of milliseconds whereas their concentrations vary over several minutes (over a 10^4 range). The two mechanisms complement each other; rapid control of enzyme activity would prevent the unnecessary loss of small molecules that would occur during the time it takes for transcriptional regulation to take effect. Control of enzyme concentrations, on the other hand, conserves the energy that would be spent in wasteful protein synthesis.

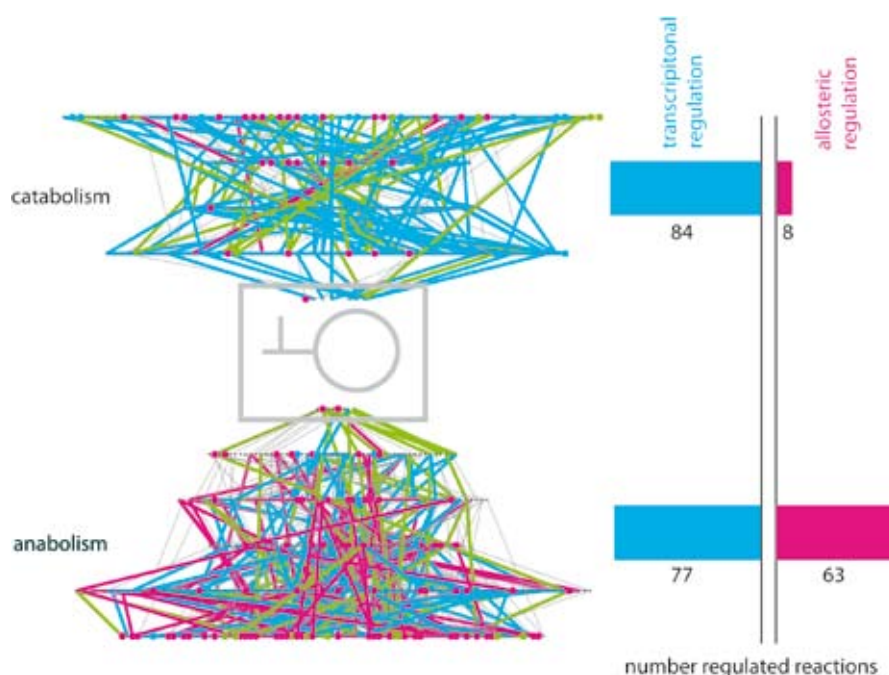


Figure 1. A network representation displays the *E. coli* metabolic system. Nodes represent small molecules and edges depict enzymatic reactions. The reactions are coloured according to whether they are controlled transcriptionally (blue), allosterically (cyan) or by both methods (green). Allosteric feedback predominantly regulates anabolic pathways, whereas transcriptional feedback controls both anabolic and catabolic pathways.

For both types of control, much of the regulation is mediated via the feedback mechanisms from small molecules, either indirectly via an intermediate transcription factor or directly through allosteric binding with the enzymes. For the *E. coli* metabolic system (Keseler *et al.*, 2005), which comprises 158 pathways and 627 small molecules, we estimate that over 30% of metabolites provide regulatory feedback. On comparing the usage of transcriptional and allosteric feedback, we show that the former is distributed evenly throughout the entire system, whereas the latter is almost entirely exclusive to anabolic pathways (figure 1). The partitioning of two modes of regulation allows cells to effectively balance the cost of small molecule depletion and protein synthesis with the benefits of cell and population growth.

Regulation of complex and infectious bacterial behaviour

Our studies of bacterial regulation form part of a broader collaboration with Gillian Fraser's group at the University of Cambridge to study complex bacterial behaviour. Bacteria are single-celled organisms typically viewed as living and acting independently of each other. However, depending on nutrient availability, surface conditions and cell density, they can transform to multicellular behaviour. Such populations have several advantages: 1) they optimise growth and survival by differentiating into distinct cell types with specialised functions; and 2) they construct a defensive matrix from which to deploy further invasive fronts. Swarming is an important manifestation of this behaviour as it enables coordinated bacterial populations to migrate rapidly over surfaces that are otherwise inaccessible to isolated cells (Rather, 2005). In a medical context, the behaviour allows pathogens to reach sites of infection in the host. For example, *E. coli* and *Proteus mirabilis* are leading causes of hospital-acquired infections (Liedl, 2001); swarming bacteria gain access to the urethra, bladder and kidneys by ascending the abiotic catheter surface and host epithelium.

Cells initiate swarming by sensing contact with a surface and with each other. This triggers a metamorphosis in which cells lengthen 20-fold and build long molecular propellers called flagella that extend outwards from the cell surface. These elongated cells (which number in the billions) then align to form large bacterial rafts and migrate away, propelled by synchronised flagellar rotation.

A complex cascade of molecular signals converts the input stimulus into a response by activating and repressing specific sets of genes required for swarming. Multiple signals (such as surface contact and population density) initiate global and reversible changes in gene expression, including increased flagella and virulence factor production, and suppression of cell division. Through 20 years of molecular biological investigations, a collection of swarming genes has been identified such as those involved in flagella construction. However, it is clear that such complex behaviour requires several hundred genes that remain unidentified. Moreover, we have only a basic understanding of how the incoming signals are transmitted to coordinate the activity of these genes.

In collaboration with Dr Fraser, we are combining computational and experimental approaches to elucidate the regulatory mechanisms underlying swarming behaviour in *E. coli* and *Proteus mirabilis*, which have closely related genomes. We are currently interrogating gene expression changes and transcription factor binding during the periodic swarm cycle using high-resolution tiling arrays. In doing so, we will identify the components of the regulatory network that underlie this complex cellular behaviour. Further, by comparing two closely related organisms, we will identify the core processes underlying this behaviour. In the longer term, this may help identify potential targets to inhibit infections caused by swarming bacteria.

MAMMALIAN TRANSCRIPTION REGULATION

Florence Cavalli, Juanma Vaquerizas, Nicholas Luscombe in collaboration with Professor Jussi Taipale, University of Helsinki

Functional census of human transcription factors

A major goal in genomic research is to study the content and usage of the human genome. To this end, we are currently examining the mammalian regulatory system. Despite the importance and popularity of research in this area, it is notable that most studies so far have focused on identifying binding sites, and there has been little attention on the transcription factors themselves.

Previously, we produced a high-confidence dataset of 1,369 gene loci encoding DNA binding transcription factors by manually identifying DNA binding domains and families from the InterPro database, extracting all human transcripts matching one of these domains, and removing false positive hits. In addition, we have used similar methods to identify DNA binding transcription factors in the rat and mouse genomes, as well as to determine the set of basal transcription factors, histones, and chromatin modifying enzymes.

Using publicly available gene expression data (Su *et al.*, 2004), we assessed the usage of these transcription factors in 33 major tissue types (862 factors were represented on the microarrays; figure 2). We found that 354 regulators are not expressed at a detectable level in these tissue types (although some of them are expressed at high levels in tumour and stem cell lines). Of the 508 factors that are expressed, 330 are detected specifically in one to five tissue types and 178 factors in all or most tissues. Interestingly, there are only a few factors with intermediate expression. We also identified

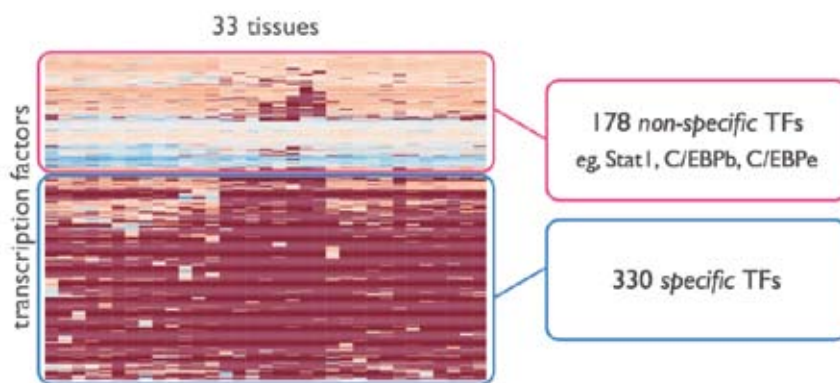


Figure 2. A heatmap displaying the pattern of expression of human transcription factors. The regulators are either specifically expressed in a small number of tissues or ubiquitously expressed across all tissues.

orthologues of these transcription factors in 19 eukaryotic genomes. It is apparent that sets of transcription factors arose in the human lineage at certain points during evolution.

High-throughput determination of DNA binding specificity and affinity

Most sequence-specific DNA binding transcription factors target binding sites by recognising a short nucleotide sequence motif. Unfortunately there is very limited information regarding the binding specificities for human regulators. Professor Jussi Taipale has pioneered a method for high-throughput screening of transcription factor specificities using a competitive binding assay of DNA oligonucleotides (Hallikas *et al.*, 2006). Potential enhancer elements in the genome can then be identified by integrating ChIP-chip data and utilising motif-searching algorithms (e.g. Palin, Taipale & Ukkonen, 2006). In collaboration with Professor Taipale, we are now measuring the DNA binding specificities of all the probable transcription factors in our high-quality dataset.

Expression and regulatory changes during human cellular differentiation

Mammalian development requires the specification of over 400 cell types from a single pluripotent cell (Vickaryous & Hall, 2006). Such stem cells can not only divide to generate pluripotent daughter cells, but they also differentiate to produce all of the cells of the mesoderm, endoderm and ectoderm as well as germ cells. In most cases, stem cells gradually restrict their lineage potential during the course of development and generate tissue-specific multipotent stem cells. There is a lot of interest in identifying genes that provide the ‘stemmy’ character of cells, and studies have so far reported a handful of regulators such as OCT4, SOX2 and Nanog that maintain the pluripotent cell type (Boiani & Scholer, 2005). So far, we have collected over 1,500 Affymetrix published experiments measuring gene expression levels in human and mouse stem cell lines at different stages of development. In collaboration with Paul Bertone, we will be using these datasets to characterise the gene expression patterns that distinguish between pluripotent and multipotent stem cells and examine the expression changes occurring during the differentiation process.

EPIGENETIC CONTROL OF DOSAGE COMPENSATION

Juanma Vaquerizas, Florence Cavalli, Nicholas Luscombe in collaboration with Dr Asifa Akhtar, EMBL Heidelberg (now Max Planck Institute for Immunology)

In higher eukaryotes, one of the most important manifestations of gene expression control is the compensation for the different numbers of sex chromosomes in the two sexes (Straub & Becker, 2007). Diploid cells have two homologous copies of every autosomal chromosome. However the situation is more complex for the sex chromosomes. In mammals and fruit flies (*Drosophila melanogaster*), females are characterised by two X chromosomes, whereas male cells contain only a single X and a Y chromosome. The worm (*Caenorhabditis elegans*) has lost the Y chromosome altogether: males have an XO genotype and hermaphrodites have XX.

Dosage compensation offsets this gross imbalance in gene content by adjusting the expression of the X chromosome. It has been suggested that species employ different strategies for dosage compensation. Male fruit flies re-instate the balance of diploid gene expression by doubling the transcription of genes on the single X chromosome (Lucchesi, 1973). It has been a long-held belief that in humans, female cells inactivate one of the X chromosomes, and that hermaphrodite worms partially repress both X chromosomes. Surprisingly however, most recent evidence suggests that all organisms upregulate the male X chromosome, suggesting a universal mechanism for dosage compensation (Gupta *et al.*, 2006; Nguyen & Disteche, 2006).

Fruit flies operate this system through the Dosage Compensation Complex. In males, the expression of the MSL2 protein stabilises the complex, allowing it to bind the X chromosome. Female cells – lacking MSL2 and therefore the complex – transcribe at the ‘normal’ rate. Despite this knowledge however, the molecular mechanisms underlying dosage compensation have remained unclear.

In collaboration with Asifa Akhtar's group at EMBL Heidelberg, we recently performed a joint wet/dry study to analyse this regulatory mechanism on a genomic scale (Kind *et al.*, 2008). We demonstrated conclusively that the Dosage Compensation Complex functions by targeting the histone modification enzyme MOF to specific sites on the male X chromosome. MOF then acetylates the histone H4 protein, releasing the chromatin-mediated transcriptional repression of the chromosomal region.

Using high-resolution tiling arrays combined with chromatin immunoprecipitation, we identified all the binding sites of the Dosage Compensation Complex, and the effect of MOF activity on chromatin structure. We complemented these data with gene expression profiles of RNAi knockdowns of these proteins. Interestingly, the Dosage Compensation Complex does not appear to operate uniformly across the entire chromosome but on individual genes. This raises the intriguing possibility that dosage compensation could also be a major mediator for phenotypic variation between individuals.

FUTURE PROJECTS AND GOALS

We will continue to develop new techniques to advance our understanding of regulatory systems, and expand our approaches towards alternative regulatory processes. We will continue to interact closely with research groups performing functional genomics experiments. Several of the projects described above are now in the final stages of completion. Clearly the primary aim of the group is to publish our work, and we expect our papers to be presented in the best peer-review journals over the next twelve months.

Group Members

Staff Scientists

Annabel Todd
Juanma Vaquerizas

Postdoctoral Fellow

Kathi Zarnack

PhD Students

Florence Cavalli
Inigo Martincorena
Aswin Sai Narain Seshasayee
Judith Zaugg

Visitors

Nishant Verma*

*Indicates part of the year only

Publications

2009

Adler, P., *et al.* (2009). Ranking genes by their co-expression to subsets of pathway members. *Ann. N. Y. Acad. Sci.*, 1158, 1-13

Illesley, G.R., *et al.* (2009). Know your limits: Assumptions, constraints and interpretation in systems biology. *Biochim. Biophys. Acta*, 1794, 1280-1287

Seshasayee, A.S.N., *et al.* (2009). Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*. *Genome Res.*, 19, 79-91

Vaquerizas, J.M., *et al.* (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, 10, 252-263

Other EMBL publications

Kind, J., *et al.* (2008). Genome-wide Analysis Reveals MOF as a Key Regulator of Dosage Compensation

and Gene Expression in *Drosophila*. *Cell*, 133, 813-828

Luscombe, N.M., *et al.* (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.*, 1, reviews001

Other publications

Boiani, M. & Scholer, H.R. (2005). Regulatory networks in embryo-derived pluripotent stem cells. *Nat. Rev. Mol. Cell Biol.*, 6, 872-884

Grainger, D.C., *et al.* (2003). Binding of the *Escherichia coli* MelR protein to the melAB promoter: orientation of MelR subunits and investigation of MelR-DNA contacts. *Mol. Microbiol.*, 48, 335-348

Grainger, D.C., *et al.* (2005). Studies of the distribution of *Escherichia coli* camp-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc. Natl Acad. Sci. USA*, 102, 17693-8

Gupta, V., *et al.* (2006). Global analysis of X-chromosome dosage compensation. *J. Biol.*, 5, 3

Hallikas, O., *et al.* (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124, 47-59

Jansen, R., *et al.* (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302, 449-453

Keseler, I.M., *et al.* (2005). EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, 33, D334-337

Liedl, B. (2001). Catheter-associated urinary tract infections. *Curr. Opin. Urol.*, 11, 75-79

Lucchesi, J.C. (1973). Dosage compensation in *Drosophila*. *Annu. Rev. Genet.*, 7, 225-237

Palin, K., Taipale, J. & Ukkonen, E. (2006). Locating potential enhancer elements by comparative genomics using the EEL software. *Nat. Protoc.*, 1, 368-374

Nguyen, D.K. & Disteche, C.M. (2006). Dosage compensation of the active X chromosome in mammals. *Nat. Genet.*, 38, 47-53

Rather, P.N. (2005). Swarmer cell differentiation in *Proteus mirabilis*. *Environ. Microbiol.*, 7, 1065-1073

Salgado, H., *et al.* (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, 34, D394-397

Straub, T. & Becker, P.B. (2007). Dosage compensation: the beginning and end of generalization. *Nat. Rev. Genet.*, 8, 47-57

Su, A.I., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, 101, 6062-6067

Vickaryous, M.K. & Hall, B.K. (2006). Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev.*, 81: 425-55

Dietrich Rebholz-Schuhmann

*Master in Medicine, 1988, University of Düsseldorf.
PhD in immunology, 1989, University of Düsseldorf.
Master in Computer Science, 1993, Passau.
Senior scientist at gsf, Munich and LION bioscience
AG, Heidelberg.
At EMBL-EBI since 2003.*



Semantic standardisation of the scientific literature

135

INTRODUCTION

Text mining comprises the fast retrieval of relevant documents from the whole body of the literature (e.g. Medline database) and the extraction of facts from the text thereafter. Text-mining solutions are now becoming mature enough to be automatically integrated into workflows for research work and into services for the general public, for example delivery of annotated full-text documents as part of UK Pubmed Central (UKPMC).

Research in the Rebholz-Schuhmann group is focused on fact extraction from the literature. It is our goal to automatically connect literature content to other biomedical data resources (e.g. bioinformatics databases) and to evaluate the results. Ongoing research targets the recognition of biomedical terms (genes, proteins, Gene Ontology labels) and the identification of relationships between them.

The work in the research group is split into different parts: 1) research work in named entity recognition and its quality control (e.g. UKPMC project); 2) knowledge discovery tasks, e.g. for the identification of gene–disease associations; and 3) further development of the IT infrastructure for information extraction. All parts are tightly coupled.

RESEARCH IN NAMED ENTITY RECOGNITION

Standardisation of the scientific literature: UKPMC and CALBC

Vivian Lee, Jung-Jae Kim, Piotr Pezik, Anika Oellrich, Menaka Naraysamy

The research work of the Rebholz-Schuhmann group is concerned with the integration of the scientific literature with the bioinformatics data resources. One important part of this research work is the identification of named entities, e.g. genes, proteins, diseases, species, from the scientific literature, and subsequently linking the entities to an entry in a reference database, for example UniProtKB for proteins. Both steps are challenging and require the use of natural language processing techniques as well as statistical methods. Several solutions are underway to normalise the representation of concepts in the scientific literature: 1) provision of a standardised lexical resource (BioLexicon); 2) definition of a schema that enables the annotation of entities in the scientific text; 3) availability of an IT infrastructure that annotates the documents with named entities and links the entities to the reference data resources; and 4) means to measure the performance and improve the quality of the annotations.

In order to provide full coverage of domain knowledge in molecular biology, the Rebholz-Schuhmann group has undertaken research to generate a complete terminological resource (BioLexicon) for gene and protein names (GPNs), chemical entities and ontological terms (e.g. Gene Ontology) as part of the European research project 'BOOTStrep' (www.bootstrep.org). A number of bioinformatics resources have been incorporated into this BioLexicon, for example, the BioThesaurus (Liu *et al.*, 2006), to cope with nonsense names and identify ambiguous terms. The quality of the BioLexicon has been assessed in its capability to improve the performance for named entity recognition for genes and proteins. Furthermore, the BioLexicon has been enriched with information from other resources, such as the scientific literature, and includes novel terms and confidence values for their relevance to the contained concepts.

In recent years, we have proposed a schema for the enrichment of the scientific literature with concept mentions (Rebholz-Schuhmann, Kirsch & Nenadic, 2006). This solution has now been implemented into the literature analysis services of the Rebholz-Schuhmann group (WhatizitLeXML) and is used for the comparison and evaluation of annotations delivered from different annotation services. The BioLexicon serves as a standard reference database for

biomedical terms and is similar to the UMLS lexical resource for the medical domain which supports research on the annotation of scientific literature. All the different resources have been integrated into a text-mining solution that indexes the full body of scientific literature as part of the UKPMC project. The annotations are delivered through CiteXplore (www.ebi.ac.uk/citexplore/) to the British Library for public use via the UKPMC interface.

The Rebholz-Schuhmann group is preparing a competition for the annotation and standardisation of the scientific literature called the 'Collaborative Annotation of a Large Biomedical Corpus' (CALBC) (see next page).

Identification of gene/protein named entities, species and diseases in scientific literature

Jee-Hyub Kim, Ian Lewin, Romain Tertiaux, Abhishek Dexit, Anika Oellrich

The identification of named entities for genes and proteins is ongoing work and is embedded into the research work for the UKPMC project (figure 1). The research team is collaborating with the National Centre for Text Mining (NaCTeM, Professor Sophia Ananiadou) in this project. New solutions have been developed over the past year, which combine dictionary-based gene mention identification with a machine learning solution.

The connection of gene/protein entities to database entries requires the identification of species-specific terms from the context of genes and proteins. To improve the normalisation, the Rebholz-Schuhmann group has advanced species identification from the scientific literature using a dictionary-based method. In this approach, statistical information on the distribution of species names in the literature has been used to reduce the false positive rate. Furthermore, the lexical resource has been adapted to the demands of literature analysis by reducing the false identification of species.

The new solution for species recognition also provides advantages for general information retrieval of full-text documents since it identifies not only the species but also the genus and any other name from the upper parts of the taxonomic hierarchy. The final solution is available through the Whatizit infrastructure and is integrated into the UKPMC prototype.

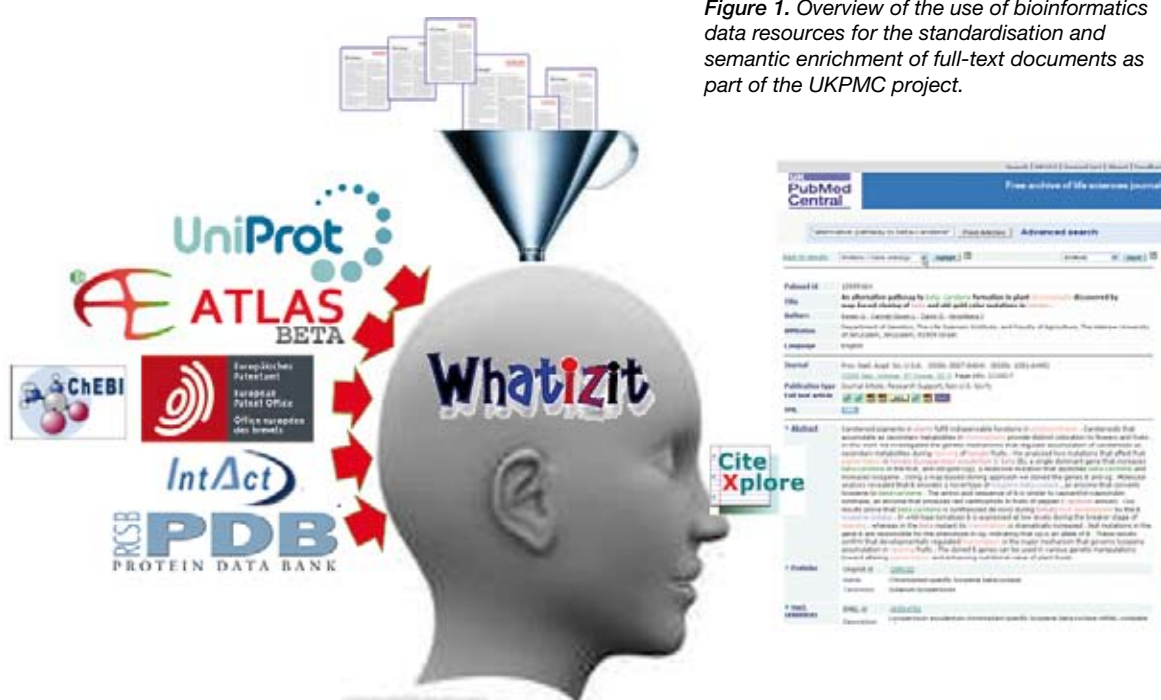
Further research work is concerned with the identification of disease terms and phenotypic information.

Identification of chemical named entities in patent texts

Delphine Bas, Piotr Pezik, Adam Bernard

In collaboration with the ChEBI team and the European Patent Office (EPO), the group is identifying Named Chemical Entities (NCEs) in biochemical patent documents. Members from the EPO and the ChEBI team have provided a manually annotated gold standard corpus that serves as training and test data. The ultimate goal is the automatic extraction of NCEs in patent data, which can then be considered for addition to the ChEBI resource.

Figure 1. Overview of the use of bioinformatics data resources for the standardisation and semantic enrichment of full-text documents as part of the UKPMC project.



As part of the ongoing work, the variety of the chemical entities contained in the gold standard corpus was analysed to determine its ‘representativeness’. Other available baseline solutions were then evaluated against the gold standard corpus, e.g. a dictionary-based NCE system (based on DrugBank and ChEBI), OSCAR3 (Corbett, Batchelor & Teufel, 2007), and a machine-learning (ML) classifier for NCE recognition. The corpus-trained classifier outperforms the other out-of-the-box solutions in terms of identifying the exact boundaries of NCEs in patent documents (precision at 0.6 and recall 0.45 on the gold standard).

Further research work is concerned with the identification of the biological activity of a chemical entity from the contextual information. This information will be used to categorise the chemical entity into different groups of based on biological activity and to cross-validate the results against reference data resources (e.g. ChEMBL).

Collaborative annotation of a large-scale corpus (CALBC)

Antonio Jimeno Yepes, Menaka Naraysamy, Chen Li

The CALBC (www.calbc.eu) initiative aims to provide a large-scale biomedical text corpus containing semantic annotations for tagged named entities of different kinds. The generation of this corpus requires that the annotations from different automatic annotation systems are harmonised.

In the first phase, the annotation systems from five participants (EMBL-EBI, EMC Rotterdam, National Library of Medicine, JULIE Lab Jena, and Linguamatics) were gathered. All annotations were delivered in a common annotation format (including concept ids in the boundary assignments) which enabled comparison and alignment of the results.

During the harmonisation phase, the results from the five participants were integrated into a single harmonised (‘silver standard’) corpus by applying a rating scheme, and the participants’ individual submissions were then evaluated against this corpus. We found that species and disease annotations were better standardised amongst the partners than the annotations of genes and proteins.

The raw corpus is now available for additional named entity annotations. Part of the annotated corpus will be made available later for a public challenge. We expect that we can improve corpus building activities both in terms of the numbers of named entity classes covered, as well as the size of the corpus in terms of annotated documents.

RESEARCH IN KNOWLEDGE DISCOVERY AND NOVEL TEXT-MINING SOLUTIONS

Multi-label classification of text with MeSH terms for Medline abstracts

Dolf Trieschnigg, Piotr Pezik, Vivian Lee

Controlled vocabularies such as the Medical Subject Headings (MeSH) thesaurus and the Gene Ontology (GO) provide an efficient way of accessing and organising biomedical information by reducing the ambiguity inherent in free-text data. Different methods of automating the assignment of MeSH concepts have been proposed to replace manual annotation, but a reliable solution and a thorough analysis of different methods has so far been missing.

The group undertook an analysis to compare the performance of five MeSH classification systems: two classifiers rely on a thesaurus, two unsupervised classifiers are trained on the MeSH annotations of Medline documents and one is a K-Nearest Neighbor (KNN) classifier. All methods were assessed with regards to their capability 1) to reproduce the original manual MeSH annotations and 2) to generate meaningful annotations which complement the manual annotations (verified by a curator).

In our analysis, KNN showed the best performance in all tests. Furthermore, we found that information retrieval can be improved (to a statistically significant degree) when the user’s query is annotated with MeSH concepts (based on KNN) instead of using the original query terms alone. Taken together, these steps mean that automatic annotation of biomedical texts with MeSH terms is ready for widespread application.

Confirmation of gene–disease associations with GO annotations from the literature

Christoph Grabmüller, Darius Sulskus, Antonio Jimeno Yepes

Candidate genes for diseases are identified in association and screening studies, thereby forming gene–disease association pairs (GDAPs). The verification of underlying mechanisms requires sophisticated experiments, whereas the scientific literature delivers this evidence in abundance. Therefore, we propose the use of Gene Ontology concepts to support the identification of GDAPs.

Concept profiles for genes and diseases have been generated based on Gene Ontology concepts from Medline. The ranking of all gene profiles against all disease profiles led to the identification of known and putative novel GDAPs. For 1,154 GDAPs the best-ranked gene for a disease was not referenced in OMIM and 63% are candidates for possibly novel associations based on a curated sample of 30 GDAPs. 57% of the top three ranked GO terms supporting the GDAP are reliable and thus give evidence for the correct interpretation of the association.

GO concept profiles showed improved performance over MeSH concept profiles, but the best results for the extraction task was delivered from the analysis that combined both. For genes that were ranked first for the associated disease, precision reached 35.2% with a recall of 16.9% when compared to OMIM. We found that molecular function concepts from GO are crucial for the performance. Finally, we reproduced existing categorisations of diseases by clustering the disease GO concept profiles (70% agreement to MeSH categories and 54% to GAD disease categories). The results show that GDAPs can be generated with good performance, and that Gene Ontology concepts extracted from literature are well suited to describing the underlying mechanisms of both genes and diseases.

Ontological support for information extraction and retrieval

Antonio Jimeno Yepes

Ontological resources such as controlled vocabularies, taxonomies and ontologies from the OBO Foundry are used to represent biomedical domain knowledge. The development of such resources is a time-consuming task but once complete, they contribute to standardisation of information representation, interoperability of IT solutions, literature analysis and knowledge discovery.

Text mining comprises IT solutions for information retrieval (IR) and information extraction (IE). IR technology exploits ontological resources to select documents that best fit the processed query, for example, through indexing of the literature content with concept ids or through disambiguation of terms in the query. IE solutions make use of the ontological labels to identify concepts in the text. The text passages that denote conceptual entries are then used either to annotate named entities or to relate the named entities to each other. For knowledge discovery solutions the identified concepts in the scientific literature are used to relate entities to each other, e.g. to identify gene–disease relationships based on shared molecular functions.

Decomposition of Gene Ontological terms

Jung-Jae Kim

Gene Ontology describes biological knowledge about gene products using a controlled vocabulary. This vocabulary is important for standardisation of biological knowledge representation and thus has great potential for standardising the representation of facts from the literature. However, the ontology terms are so artificial that they do not literally appear in the literature. To associate the vocabulary with the literature closely, we decompose the ontology terms into words or basic terms that are far better matched to the literature than the ontology terms.

To keep the formality of Gene Ontology, we utilise another ontology, called Gene Regulation Ontology (GRO), as the basis for the decomposition. GRO provides concepts and properties that are logically integrated to represent the semantics of GO terms. Using GRO we develop patterns of the basic concepts combined with semantic information. We apply an event extraction system for automatically matching the patterns to GO terms for the semantic analysis.

The semantic representation of GO terms with GRO can be used for structural comparison between GO terms for the purpose of consistency checking, for deducing implicit knowledge from GO term relations, and for mining GO concepts from the literature.

FURTHER DEVELOPMENT OF THE IT INFRASTRUCTURE FOR INFORMATION EXTRACTION

PaperMaker

Silvestras Kavaliauskas, Piotr Pezik

The automatic analysis of scientific literature can support authors in writing their manuscripts. PaperMaker is a novel IT solution that receives a scientific manuscript via a web interface, automatically analyses the publication, evaluates consistency parameters and interactively delivers feedback to the author. It analyses the proper use of acronyms and their definitions, and the use of specialised terminology. In addition, PaperMaker provides additional services such as GO and MeSH categorisation of text passages, the retrieval of relevant publications from public scientific literature repositories, and the identification of missing or unused references. As a final step, the author receives a summary of the results, the manuscript in its corrected form and a structured digital abstract containing the GO and MeSH annotations (www.ebi.ac.uk/Rebholz-srv/PaperMaker).

FUTURE PROJECTS AND GOALS

The following goals are priorities for the future. Firstly we will continue our ongoing research in term recognition and mapping to biomedical data resources to establish state-of-the-art text-mining applications. This work is constantly monitored by automatic means to measure and evaluate the results to identify the most promising solutions (UKPMC project).

Secondly, we will invest further effort into the extraction of content from the scientific literature. Such solutions will be geared towards the annotation of diseases and the generation of fact databases. As part of this research we will

investigate workflow systems where text mining supports bioinformatics information retrieval solutions. One solution is the integration of public biomedical data resources into the data from the biomedical scientific literature.

Finally, we will increase the availability of information extraction solutions based on SOAP web services for the benefit of the bioinformatics community. This requires standards in the annotation of scientific literature and will automatically lead to semantic enrichment of the literature. Disambiguation of semantic types requires special solutions.

Group Members

Staff Scientists

Vivian Lee
Piotr Pezik
Antonio Jimeno Yepes
Christoph Grabmüller
Jee-Hyub Kim
Ian Lewin
Chen Li

Postdoctoral Fellow

Jung-Jae Kim

Software Engineers

Silvestras Kavaliauskas*
Menaka Naraysamy

PhD Students

Adam Bernard*
Anika Oellrich
Kevin Nagel

Visitors

Dolf Trieschnigg
Pinar Yildirim

Visiting Students

Arun Gupta
Darius Sulskus
Delphine Bas
Rohit Rexa
Abhishek Dixit

* Indicates part of the year only

Acknowledgements

Whatizit has been supported by the EU FP6 Network of Excellence 'Semantic Interoperability and Data Mining in Biomedicine' (NoE 507505). Medline abstracts are provided from the National Library of Medicine (NLM, Bethesda, MD, USA) and PubMed (www.pubmed.org) is the premier web portal to access the data. Sylvain Gaudan is supported by an E-STAR fellowship (EC's FP6 Marie Curie Host fellowship for Early Stage Research Training, MESTCT-2004-504640).

BOOtStrep (FP6-028099) is funded as a STREP project in the EC's FP6 IST programme.

Publications

2008

Altman, R.B., *et al.* (2008). Text mining for biology – The way forward: Opinions from leading scientists. *Genome Biol.*, 9, S7.1-S7.15

Beisswanger, E., *et al.* (2008). Gene Regulation Ontology (GRO): Design principles and use cases. *Stud. Health Technol. Inform.*, 9-14

Cerri, D., *et al.* (2008). Towards Knowledge in the Cloud. In 'OTM 2008 Workshops including SEMELS', Meersman, R., *et al.*, (eds)

Ramialison, M., *et al.* (2008). Rapid identification of PAX2/5/8 direct downstream targets in the otic vesicle by combinatorial use of bioinformatics tools. *Genome Biol.*, 9, r145

Sasaki, Y., *et al.* (2008). BioLexicon: A Lexical Resource for the Biology Domain. In 'Third International Symposium on Semantic Mining in Biomedicine (SMBM)'

2009

Baker, C.J.O. & Rebholz-Schuhmann, D. (2009). Between proteins and phenotypes: Annotation and interpretation of mutations. *BMC Bioinformatics*, 10, 11

Couto, F., *et al.* (2009). Verification of Uncurated Protein Annotations. In 'Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration', 311-325, IGI Global Publishing

Jimeno-Yepes, A., *et al.* (2009). Terminological cleansing for improved information retrieval based

on ontological terms. In 'Proceedings of the WSDM 2009 ACM Workshop on Exploiting Semantic Annotations in Information Retrieval, ESAIR 2009', 6-14

Nagel, K., *et al.* (2009). Annotation of protein residues based on a literature analysis: Cross-validation against UniProtKb. *BMC Bioinformatics*, 10, Suppl 8, S4

Pezik, P., *et al.* (2009). Using Biomedical Terminological Resources for Information Retrieval. In 'Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration', IGI Global Publishing

Trieschnigg, D., *et al.* (2009). MeSH Up: Effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25, 1412-1418

Other EMBL publications

Rebholz-Schuhmann, D., Kirsch, H. & Nenadic, G. (2006). leXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text. *BioLINK*, ISMB 2006, Fortaleza, Brazil

Other publications

Butte, A.J., *et al.* (2006) Creation and implications of a phenome-genome network. *Nat Biotechnol.*, 24, 55-62

Corbett, P., Batchelor, C. & Teufel, S. (2007). Annotation of chemical named entities. *BioNLP 2007: Biological, translational, and clinical language processing*, Prague, Czech Republic, 57-64

Liu, H., *et al.* (2006). BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22, 103-105

Janet Thornton

PhD 1973, King's College & National Inst. For Medical Research, London.
Postdoctoral research at the University of Oxford, NIMR & Birkbeck College, London.
Lecturer, Birkbeck College 1983-1989.
Professor of Biomolecular Structure, University College London since 1990.
Bernal Professor at Birkbeck College, 1996-2002.
Director of the Centre for Structural Biology at Birkbeck College and University College London, 1998-2001.
Director of EMBL-EBI since 2001.



Computational biology of proteins

141

INTRODUCTION

The goal of our research is to understand more about how biology works at the molecular level, how enzymes perform catalysis, how these molecules recognise one another and their cognate ligands, and how proteins and organisms have evolved to create life. We develop and use novel computational methods to analyse the available data, gathering data either from the literature or by mining the data resources, to answer specific questions. Much of our research is collaborative, involving either experimentalists or other computational biologists. During 2009 our major contributions have been in the following areas:

- enzyme structure and function;
- using structural data to predict protein function;
- functional genomics analysis of ageing.

ENZYME STRUCTURE AND FUNCTION

Julia Fischer, Nicholas Furnham, Gemma Holliday, Asad Rahman

In our attempts to understand how enzymes work, we have further studied the chemistry of catalysis (Holliday *et al.*, 2009) and coenzymes. Building on an analysis of overall reaction mechanisms (as stored in MACiE, our database of enzyme mechanisms), we have focused on understanding the roles of individual amino acids in catalysis, identifying stabilisation, general acid/base catalysis (proton acceptor and donor) and nucleophilic addition as the major functions. Different amino acids perform different roles, as expected from their different chemistries, and consequently their use varies between the different EC classes of enzymes.

In collaboration with the group of Professor Ivano Bertini at the University of Florence, we have made a structural analysis of non-heme iron sites in proteins (Andreini *et al.*, 2009) and considered metals ions in biological catalysis in general (Andreini *et al.*, 2008). This has led to the development of a sister database for MACiE which includes these data. In addition, we have studied several individual enzymes in more detail, through *in silico* docking and energetic approaches (e.g. Blum *et al.*, 2009).

Annotations of enzyme function provide critical starting points for generating and testing biological hypotheses, but the quality of functional annotations is hindered by uncertain assignments for uncharacterised sequences and by the relative sparseness of validated experimental data. A review of enzyme annotations in collaboration with UniProt (Furnham *et al.*, 2009) has highlighted the challenges and we are currently using the data in the Catalytic Site Atlas and in MACiE to identify omissions and errors in the public databases and seek ways to update them. A review of protein promiscuity (Nobeli *et al.*, 2009) suggests that promiscuity, not only in interactions but also in the actual function of proteins, is not as rare as was previously thought. This has implications not only for our fundamental understanding of molecular recognition and how protein function has evolved over time but also in the realm of biotechnology. Understanding protein promiscuity is becoming increasingly important not only to optimise protein engineering applications in areas as diverse as synthetic biology and metagenomics but also to lower attrition rates in drug discovery programmes, identify drug interaction surfaces less susceptible to escape mutations and potentiate the power of polypharmacology.

Lastly we have devoted considerable effort to developing new tools to handle small molecules and their reactions. This was necessary so that we can query, cluster and analyse the different reaction mechanisms, including the role of cofactors. We have completed the development of SMSD (Small Molecule Subgraph Detector), a new Maximum Common Subgraph (MCS) tool to compare small molecules, which overcomes some of the issues with current heuristic approaches to small molecule similarity searches. The MCS search implemented in SMSD incorporates chemical knowledge (atom type match with bond sensitive and insensitive information) while searching for molecular similarity. We also propose a novel method by which solutions obtained by each MCS run can be ranked using chemical filters such as stereochemistry and bond energy, etc. The tool was benchmarked by a 50,000 pairwise comparison between KEGG ligands and PDB HET group atoms and showed improvements over the current publicly available tools in speed and efficiency (Rahman *et al.*, 2009). This tool can be applied to various areas of bioinformatics and chemoinformatics for finding exhaustive MCS matches. SMSD is already used widely within the group, and also by other groups around the world. Further research has led to the development of a tool to compare enzyme reactions, which can be applied to the whole reaction, or to the individual steps of a catalytic mechanism as stored in MACiE. This is currently being tested and should allow a more quantitative and powerful analysis of reactions and their mechanisms.

USING STRUCTURAL DATA TO PREDICT PROTEIN FUNCTION

Matthew Bashton, Tjaart de Beer, Nicholas Furnham, Abdullah Kahraman, Roman Laskowski, Rafael Najmanovich, Marialuisa Pellegrini, David Talavera

Understanding the relationship between protein structure and biological function has long been a major goal of structural biology. With the advent of many structural genomics projects, there is a practical need for tools to analyse and characterise the possible functional attributes of a new structure.

In collaboration with Professor David Jones at University College London (UCL), we are developing new tools to characterise, classify and understand the structures and functions of transmembrane proteins. As part of this work, we have studied transmembrane channel proteins, which play pivotal roles in maintaining the homeostasis and responsiveness of cells and the cross-membrane electrochemical gradient by mediating the transport of ions and molecules through biological membranes. We have developed a fully automated method that detects and fully characterises channels in

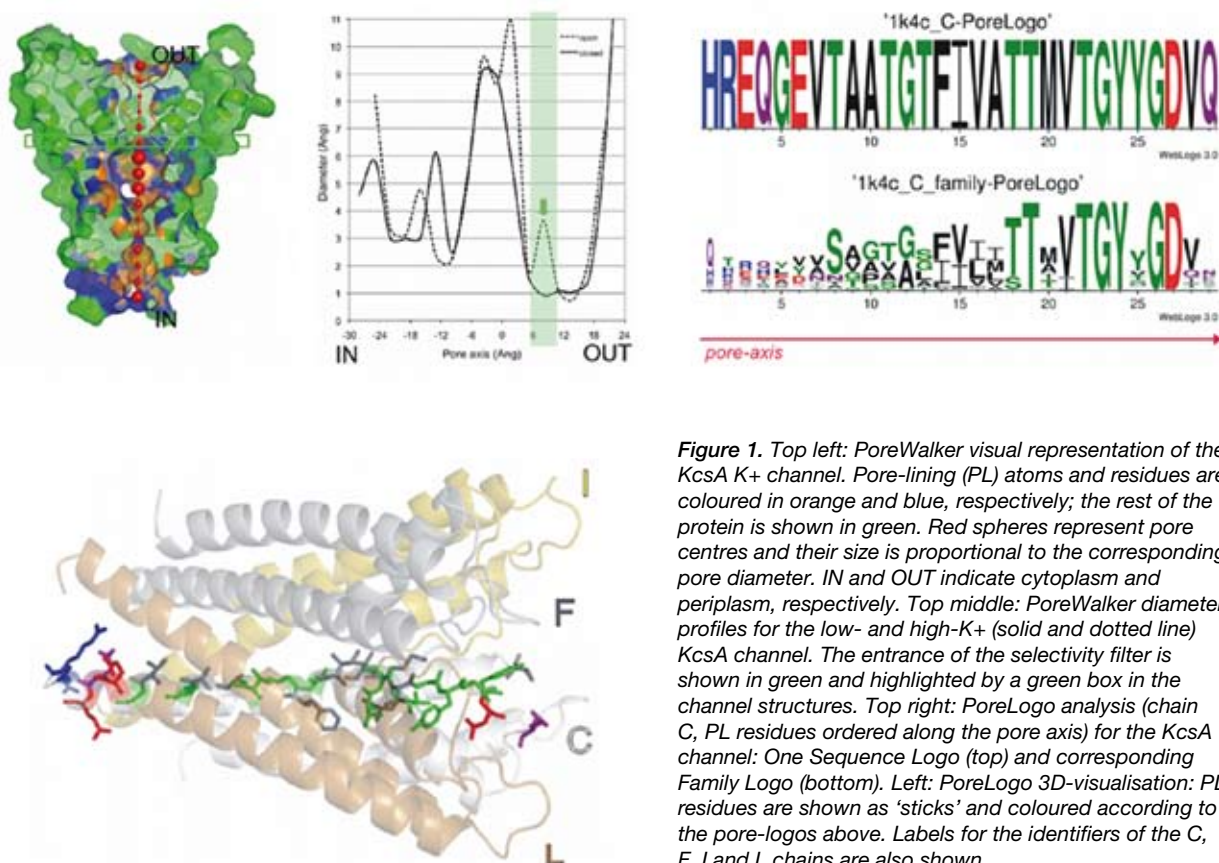


Figure 1. Top left: PoreWalker visual representation of the KcsA K⁺ channel. Pore-lining (PL) atoms and residues are coloured in orange and blue, respectively; the rest of the protein is shown in green. Red spheres represent pore centres and their size is proportional to the corresponding pore diameter. IN and OUT indicate cytoplasm and periplasm, respectively. Top middle: PoreWalker diameter profiles for the low- and high-K⁺ (solid and dotted line) KcsA channel. The entrance of the selectivity filter is shown in green and highlighted by a green box in the channel structures. Top right: PoreLogo analysis (chain C, PL residues ordered along the pore axis) for the KcsA channel: One Sequence Logo (top) and corresponding Family Logo (bottom). Left: PoreLogo 3D-visualisation: PL residues are shown as 'sticks' and coloured according to the pore-logos above. Labels for the identifiers of the C, F, I and L chains are also shown.

transmembrane proteins from their three-dimensional structures. This method has also been extended to protein sequences belonging to families of proteins with known structures. A stepwise procedure is followed in which the pore centre and pore axis are first identified and optimised using geometric criteria, and then the biggest and longest cavity through the channel is detected. Finally, pore features, including diameter profiles, pore-lining residues, size, shape and regularity of the pore are calculated, providing a quantitative and visual characterisation of the channel (see figure 1). The method is currently being applied to all the transmembrane channel proteins in the Protein Data Bank (PDB) to identify shape/size/residue features representative of specific channel families (Pellegrini-Calace *et al.*, 2009). In collaboration with Dr Romina Oliva from the University of Naples, we have used this approach to study the important aquaporin family of proteins, which control the flow of many molecules in and out of cells, to understand how their sequences and structures determine the varied ligand selectivities of the different members of the family.

In collaboration with Professor Christine Orengo at UCL, we are studying protein families for functional and comparative genomics. At EMBL-EBI we are developing approaches based on alignments generated at UCL to automatically explore how large families evolve to bind multiple different ligands, combining the phylogenetic analyses with analysis of the chemical structure of enzymes and their ligands (see figure 2). We have developed integrative tools to handle the data and automate our study and have just started the analysis of six large families of enzymes.

As part of the Midwest Center for Structural Genomics (MCSG), we have continued to analyse the new structures as they are generated and provide functional predictions. An important part of this is to consider binding sites and their variation between proteins, using physicochemical characteristics such as electrostatic potentials and hydrophobicity. Recent studies have revealed the enormous variation seen in binding sites in unrelated proteins for the same molecule e.g. ATP and NAD (Kahraman *et al.*, 2009, in press). In the same study it was demonstrated that most of the analysed protein–ligand complexes lacked perfect physicochemical complementarity. Continuing our analysis of recognising cognate ligands, we have shown that in crystal structures, isolated phosphates from the crystallisation buffer commonly bind in the site occupied by a phosphate moiety of the larger cognate ligand.

As part of the BioSapiens Consortium, we have developed a new ontology for protein sequence annotation (Reeves *et al.*, 2008) and have been analysing cancer mutations (Talavera *et al.*, 2009, in press). This work is continuing with a comparative analysis of variations observed in the 1000 Genomes Project.

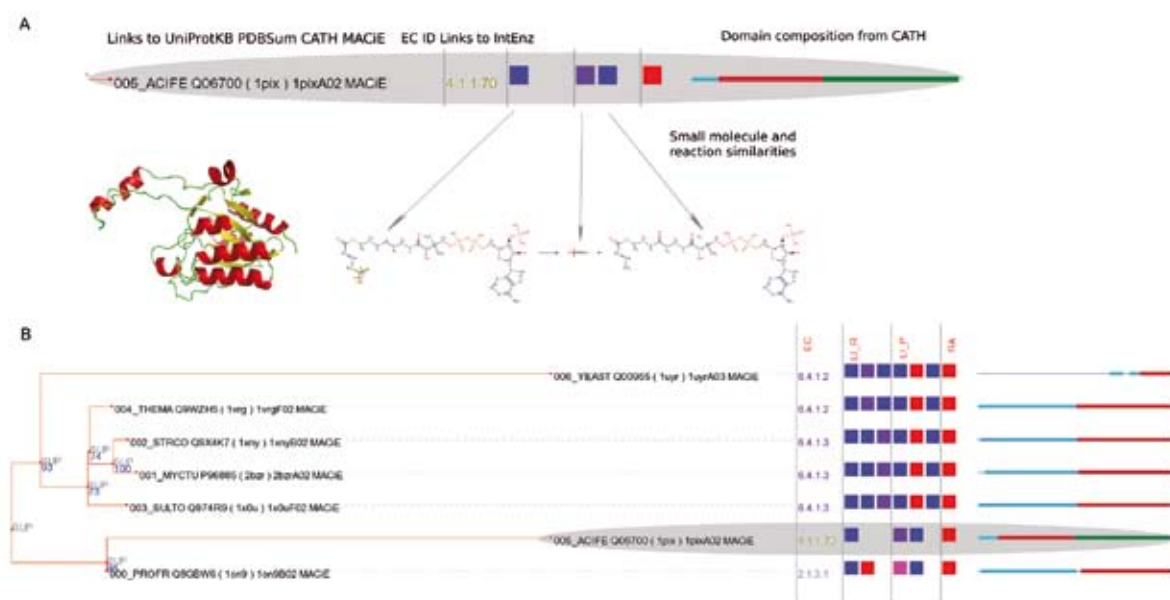


Figure 2. FunTree – a new tool for combining phylogenetic and substrate/reaction comparison data to explore the evolution of new functions within an enzyme family. A: Data for one member of the crotonase superfamily. For each enzyme the domain composition, taken from CATH, is illustrated, with the 'parent' domain highlighted in red and each enzyme is linked to structural data (PDBSum, CATH), sequence data (UniProtKB) and functional data (MACiE, IntEnz) associated with that domain. The similarity between the small molecules and reactions (bond changes are highlighted in green) for all members of this enzyme family are calculated and coloured accordingly in the square boxes. B: The phylogenetic tree of the crotonase superfamily built on the structure-based sequence alignment of the domains assigned to the CATH superfamily. A statistical confidence score for each node is given as well as links to the superimposed structures at each node.

FUNCTIONAL GENOMICS ANALYSIS OF AGEING

Dan Andrews, Eugene Schuster, Daniela Wieser

In collaboration with the Functional Genomics of Ageing Consortium at UCL, we have analysed various functional genomics datasets to understand more about the molecular basis of ageing, involving a variety of different experiments on flies, worms and mice. ChIP-chip and microarray data were used to identify genes which may be regulatory targets of dFoxo in *Drosophila melanogaster*. This work produced a first list of downstream targets of dFoxo regulation and has led to successful follow-up experiments to identify further genes which influence longevity when mutated. To understand how calorie restriction affects longevity, we have shown in mice that deletion of ribosomal S6 protein kinase 1 (S6K1), a component of the nutrient-responsive mTOR (mammalian target of rapamycin) signalling pathway, led to increased life span and resistance to age-related pathologies, such as bone, immune, and motor dysfunction and loss of insulin sensitivity. Deletion of *S6K1* induced gene expression patterns similar to those seen in calorie restriction or with pharmacological activation of adenosine monophosphate (AMP)-activated protein kinase (AMPK), a conserved regulator of the metabolic response to calorie restriction. Our results demonstrate that S6K1 influences healthy mammalian life span and suggest that therapeutic manipulation of S6K1 and AMPK might mimic calorie restriction and could provide broad protection against diseases of ageing (Selman *et al.*, 2009).

These analyses have led us to consider the role of tissue-specific effects in ageing and to develop some new approaches to define tissue specificity of genes in human, fly and mouse data. Using a measure based on Shannon's Entropy, we investigated the capacity of whole animal expression arrays (commonly used in fly experiments) to detect tissue specific differential expression. Filtering tissue-specific genes can help to identify important functional categories which are affected by ageing. This approach showed that the majority of genes with age-dependent transcriptional changes have tissue-specific expression profiles. This complicates comparisons between species and drastically reduces the chances of identifying common genes. However this analysis allowed us to identify a small number of genes common across the species with age-dependent regulation in more than one tissue. These genes perform roles in apoptosis and inflammation.

DEVELOPMENT OF TOOLS AND WEB RESOURCES

Gemma Holliday, Roman Laskowski, Marialuisa Pellegrini, Asad Rahman, David Talavera

Several new tools and resources have been developed as part of our research within the group. These include:

- Metal-MACiE, a new publicly available web-based database, which organises information on the properties and roles of metals in catalysis (www.ebi.ac.uk/thornton-srv/databases/Metal_MACiE/home.html). This has been developed in collaboration with Professor Bertini's laboratory in Florence (Andreini *et al.*, 2009);
- several new attributes added to PDBsum, which provides summary information about each experimentally-determined structure in the PDB, including Pfam domain diagrams, protein-protein interaction diagrams and citation data (www.ebi.ac.uk/pdbsum; Laskowski, 2009);
- two novel tools (PoreWalker and PoreLogo) available as a web-based resource for the identification and characterisation of channels in transmembrane proteins from their three-dimensional structure (www.ebi.ac.uk/thornton-srv/software/PoreWalker/; Pellegrini *et al.*, 2009);
- a new algorithm, SMSD which was developed for small molecular two-dimensional comparison, incorporating chemical knowledge and filters such as stereochemistry and bond energy. This tool is freely available at www.ebi.ac.uk/thornton-srv/software/SMSD; Rahman *et al.*, 2009;
- a new web service for the annotation of functional residues through structural homologues. WSsas uses similarity searches and pairwise alignments to transfer functional information about binding, catalytic and protein-protein interaction residues from solved structures to query sequences. It is available at www.ebi.ac.uk/thornton-srv/databases/WSsas/; Talavera *et al.*, 2009.

FUTURE PROJECTS AND GOALS

We will continue our work on understanding more about enzymes and their mechanisms using structural and chemical information. This will include a study of how the enzymes, their families and their pathways have evolved and how genetic variations in individuals impacts on structure, function and disease. We will apply the new computational tools we have developed to improve the handling of mechanisms and their reactions in order to gain a better understanding of reaction space and its impact on pathways. This will also allow improved chemistry queries across our databases. We will continue to use evolutionary approaches to improve our prediction of protein function from sequence and structure. In the ageing project we are interested in tissue specificity and combining human public transcriptome datasets with results from flies, worms and mice to explore effects related to human variation and age.

Group Members

Staff Scientists

Dan Andrews
Matthew Bashton
Tjaart de Beer*
Angelo Favia*
Nicholas Furnham
Gemma Holliday
Roman Laskowski
Rafael Najmanovich*
Marialuisa Pellegrini-Calace
Syed Asad Rahman
Eugene Schuster*
David Talavera*
Daniela Wieser

PhD Students

Julia Fischer
Fabian Gerick*
Abdullah Kahraman

Personal Assistant to the Director

Helen Barker-Dobson

PA to the Directors Office

Stacy Schab

Visitors

Claudia Andreini
Lorenzo Baldacci
Noa Berkovich
Eric Bornberg-Bauer
Barbara Brodsky
Gabriele Cavallaro
Franz Fenninger
Kyou Hoon
Romina Oliva
Sandya Tiwari
Mauno Vihinen
Paul Wackers
Kazuto Yamazaki

* Indicates part of the year only

Publications

2008

Andreini, C., *et al.* (2008). Metal ions in biological catalysis: From enzyme databases to general principles. *J. Biol. Inorg. Chem.*, 13, 1205-1218

Reeves, G.A., *et al.* (2008). The protein feature ontology: A tool for the unification of protein feature annotations. *Bioinformatics*, 24, 2767-2772

2009

Andreini, C., *et al.* (2009). Metal-MACiE: A database of metals involved in biological catalysis. *Bioinformatics*, 25, 2088-2089

Andreini, C., *et al.* (2009). Structural Analysis of Metal Sites in Proteins: Non-heme Iron Sites as a Case Study. *J. Mol. Biol.*, 388, 356-380

Berka, K., *et al.* (2009). Representative amino acid side chain interactions in proteins: a comparison of highly accurate correlated ab initio quantum chemical and empiri-

cal potential procedures. *J. Chem. Theory Comput.*, 5, 982-992

Blum, A., *et al.* (2009). 11B-Hydroxysteroid dehydrogenase type 1 inhibitors with oleanan and ursan scaffolds. *Mol. Cell. Endocrinol.*, 301, 132-136

Cuff, A., *et al.* (2009). The CATH Hierarchy Revisited-Structural Divergence in Domain Superfamilies and the Continuity of Fold Space. *Structure*, 17, 1051-1062

Cuff, A.I., *et al.* (2009). The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, 37, D310-D314

El-Hawari, Y., *et al.* (2009). Analysis of the substrate-binding site of human carbonyl reductases CBR1 and CBR3 by site-directed mutagenesis. *Chem. Biol. Interact.*, 178, 234-241

Furnham, N., *et al.* (2009). Missing in action: Enzyme functional annotations in biological databases. *Nat. Chem. Biol.*, 5, 521-525

Holliday, G.L., *et al.* (2009). Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis. *J. Mol. Biol.*, 390, 560-577

Kahraman A., *et al.* (2009). On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins*, [Epub ahead of print] PMID: 19927322

Laskowski, R. (2009). Integrated Servers for Structure-informed Function Prediction. In 'Protein Structure to Function with Bioinformatics', Rigden, D.J. (ed), 251-272, Springer

Laskowski, R. (2009). Protein Structure Databases. In 'Data Mining Techniques for the Life Sciences, Methods in Molecular Biology', Carugo, O. & Eisenhaber, F. (eds), 609, Humana Press

Laskowski, R. (2009). Structural Quality Assurance. In 'Structural Bioinformatics', Gu, J. & Bourne, P.E. (eds), 341-375, John Wiley

Laskowski, R.A. (2009). PDBsum new things. *Nucleic Acids Res.*, 37, D355-359

Laskowski, R.A., *et al.* (2009). The structural basis of allosteric regulation in proteins. *FEBS Lett.*, 583, 1692-1698

Laskowski, R.A., *et al.* (2009). The fine details of evolution. *Biochem. Soc. Trans.*, 37, 723-726

Loewenstein, Y., *et al.* (2009). Protein function annotation by homology-based inference. *Genome Biol.*, 10, 207

Nobeli, I., *et al.* (2009). Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.*, 27, 157-167

Pellegrini-Calace, M., *et al.* (2009). PoreWalker: A novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput. Biol.*, 5, e1000440

Persson, B., *et al.* (2009). The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.*, 178, 94-98

Rahman, S.A., *et al.* (2009). Small Molecule Subgraph Detector (SMSD) toolkit. *Journal of Cheminformatics*, in press

Reeves, G.A., *et al.* (2009). Genome and proteome annotation: Organization, interpretation and integration. *J. R. Soc. Interface*, 6, 129-147

Selman, C., *et al.* (2009). Ribosomal protein S6 kinase 1 signaling regulates mammalian life span. *Science*, 326, 140-144

Steinbeck, C., *et al.* (2009). New open drug activity data at EBI. *Chem. Cent. J.*, 3, 62

Sternberg, M., *et al.* (2009). Protein evolution – Sequence, structure and systems. *Biochemist*, 31, 52-52

Talavera, D., *et al.* (2009). WSSas: A web service for the annotation of functional residues through structural homologues. *Bioinformatics*, 25, 1192-1194

Talavera, D., *et al.* (2009). The (non) malignancy of cancerous amino acidic substitutions. *Proteins*, in press

Thornton, J. (2009). Annotations for all by all – The BioSapiens network. *Genome Biol.*, 10, 401

Watson, J.D. & Thornton, J.M. (2009). Protein function prediction from structure in structural genomics and its contribution to the study of health and disease. In 'NATO Security through Science Series C: Environmental Security', Sussman, J. (ed), 201-215, Springer-Verlag

Wieser, D. & Niranjana, M. (2009). Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores. *In Silico Biol.*, 9, 89-103

Section 4

Support in 2009

147

Outreach and Training
Industry Support
Systems and Networking
External Services Team

149
159
163
165



Cath Brooksbank

Head of Outreach and Training

*PhD in Biochemistry, University of Cambridge, 1993.
Assistant Editor then Editor, Elsevier Trends, Cambridge
and London, UK, 1993–2000.
Associate Editor then Editor, Nature Reviews, London,
2000–2002.
At EMBL-EBI since 2002.*



Nick Goldman

*Research and
Training Coordinator*



Outreach and Training

149

INTRODUCTION

The Outreach and Training team (OTT) exists to distil the outputs of the EBI to a wide range of audiences, either by acting as a source of information and news, or by providing training on the EBI's data resources to researchers. These responsibilities involve 1) communicating the scientific mission and activities of the EBI to a wide range of audiences and 2) coordinating and operating the bioinformatics training programme and other training activities. As the EBI grows, so does the breadth and number of outreach and training-associated activities. We have strengthened the team this year by creating two new posts, one in training (eLearning content developer) and one in outreach (scientific outreach officer); our previous scientific training officer (Vicky Schneider) and scientific outreach officer (Louisa Wright) have taken on new roles, assuming managerial responsibility for the training programme and outreach programme, respectively. Our Logistics and Admin sub-team (Janet Copeland, Holly Foster and Alison Barker) has assumed responsibility for running industry programme workshops in addition to other events. This has resulted in a post being opened for a third workshops and exhibitions organiser. Finally, our scientific training officer, James Watson, has assumed formal responsibility for managing the EBI's IT training facilities.

OUTREACH

Louisa Wright, Katrina Pavelin, Cath Brooksbank

The EBI's outreach activities take many different shapes and are performed not just by the members of the team and the EBI's Outreach and Training representatives (OTRs, see Panel 1) but by many different sectors of the EBI: group/team leaders and PhD students promote the EMBL International PhD Programme; tutorials at conferences allow delegates to get to grips with EBI resources and point users to our other training activities; and careers fairs allow us to display the range of employment opportunities at the EBI.

One of the main roles of OTT is to coordinate and support these activities. By working closely with EBI staff members, we aim to align our messages with the EBI's key objectives, and disseminate accurate information of a consistently high standard. OTT also acts as a main point of contact for external enquiries, either from the media or other interested parties, and works with the EBI's scientists to ensure clarity and relevance of information provided in response to these queries.

Such is the variety of the EBI's outreach activities that we can only present our main activities here although we thank everyone who has contributed their time in supporting EBI's outreach in 2009.

Building relationships with the media

Louisa Wright, Katrina Pavelin, Cath Brooksbank, supported by EMBL press office

The EBI issues its own press releases with support from EMBL's Office of Information and Public Affairs (OIPA). The majority of our releases are targeted at the specialist press in accordance with our media strategy to capitalise on the EBI's reputation in the bioinformatics field. This also protects our relationship with the mainstream media as we provide them only with news relevant to a more general audience. This commitment has been strengthened this year with formalised media guidelines agreed between the EBI and EMBL press office for evaluating the media appeal of potential news features and the EBI taking responsibility for the dissemination and promotion of technical press releases.

In addition to official press releases, we also post research highlights (often linked to a high-profile publication) and smaller news items on the EBI's front page (www.ebi.ac.uk/). From September 2008–August 2009, we released three EMBL-EBI press releases, contributed to one EMBL press release, released three press releases jointly with other organisations (Biotechnology and Biological Sciences Research Council, University of Manchester and UK PubMedCentral Consortium) and issued five web-based news announcements. Details of EMBL-EBI press releases and our web-based announcements are shown below in reverse chronological order and figure 1 shows the number of press releases charted over the past ten years.

- UK leads European research programme with £10M investment in bioscience data handling capacity, 25 August 2009 (joint press release with BBSRC)
- Launch of the first standard graphical notation for biology, 7 August 2009 (press release)
- EMBL-EBI Director Janet Thornton awarded an ISCB Accomplishment by a Senior Scientist Award, 8 July (research highlight)
- EMBL-EBI and University of Manchester launch major new e-science resource, 1 July 2009 (joint press release with the University of Manchester)
- Gerard Kleywegt to head Protein Data Bank Europe, 30 June 2009 (research highlight)
- Sweden is the first country to pledge long-term funding for ELIXIR, 23 June 2009 (research highlight)
- Mapping gene expression with Gene Expression Atlas, 15 June 2009 (press release)
- Launch of new Ensembl Genomes and new-look Ensembl website means more genome power at your fingertips, 20 April 2009 (press release)
- A census of human transcription factors reveals their secrets, 10 March 2009 (research highlight)
- Re-write the textbooks: transcription is bidirectional, 26 January 2009 (press release with EMBL)
- How bacteria survive environmental change, 3 November 2008 (research highlight)
- UK PubMed Central to become the essential gateway for the UK's biomedical and health science researchers, 16 September 2008 (joint press release with the UKPMC Consortium)

A particular highlight of our media activities in 2009 was the widespread attention received by the announcement of a £10 million investment in the ELIXIR project by the BBSRC and which was covered by the local media (BBC News Look East television broadcast, BBC online news for Cambridgeshire and Business Weekly online news for the East of England), and the global and European research media (Genome Web and Cordis online news). The press release on the launch of a standard graphical notation for biology also received coverage from over 40 sources. Due to the relevance of this story to a specific facet of the scientific community, coverage was mainly from research-themed websites and blogs.

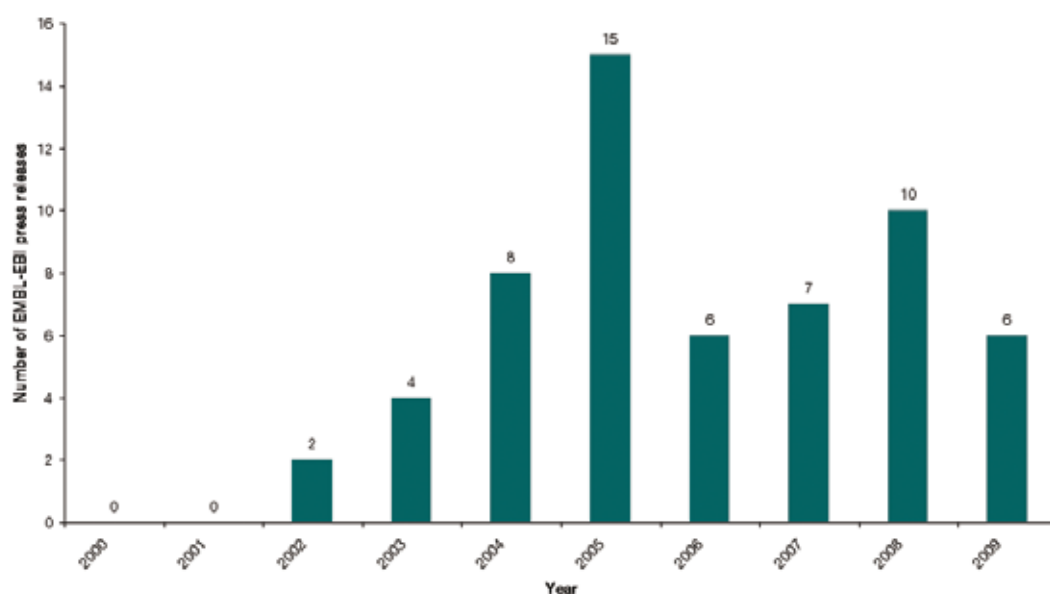


Figure 1. Number of press releases issued by EMBL-EBI each year from 2000–2009.

EBI literature

Louisa Wright and Katrina Pavelin with support from outreach representatives from each team/group

We maintain a core of promotional literature, including the 'EBI in a Nutshell' guide and resource-specific factsheets, which are regularly distributed at conferences and training events. We have added two new brochures to this core in 2009 – a brochure for our Industry Programme and a careers brochure (both available from www.ebi.ac.uk/Information/Brochures/). OTT also contributes EMBL-EBI specific information for EMBL's promotional literature, produced by OIPA.

Promoting EBI coordinated EU-funded projects

Louisa Wright and Cath Brooksbank with support from project managers

As coordinating partner of the ELIXIR preparatory phase project, the EBI is involved in the project's communication and outreach activities. We promote the ELIXIR project in presentations by EBI staff, by dissemination of literature at conferences and events and through news announcements. In addition to ELIXIR-focused press releases for the scientific and general media, we have also provided content for several specialist publications as editorials and features including:

- an article for the GARNet newsletter, reaching the UK Arabidopsis and plant science research community;
- an editorial on the progress of ELIXIR and project architecture published in September 2009 in the eStrategies magazine, which has a Europe-wide readership of 39,000;
- a viewpoint article for Research Europe, introducing ELIXIR and the importance of long-term data storage for researchers. Research Europe has a wide audience including universities, policy makers and the private sector;
- a shared editorial in Moderngov, with LifeWatch, an ESFRI (European Strategy Forum on Research Infrastructures) infrastructure for biodiversity research. The magazine has a readership of over 9,000 consisting of key decision makers in UK central and local government. The article was featured in three special editions for the 2009 UK government autumn party conferences;
- an article in the research review of Parliament Magazine, with a special focus on research infrastructures (November 2008) reaching a readership of 12,000 in the European Parliament, European Commission and throughout the EU.
- a feature in the European Companion, one of the premier reference tools for the European Union. The book is used by European policy decision makers to gain immediately accessible information on their colleagues in the European Parliament and Commission, plus information about the make up and workings of the European Union.

Exhibiting at conferences

Louisa Wright, Katrina Pavelin, Janet Copeland, Cath Brooksbank, James Watson, Alison Barker, Holly Foster

This year, the EBI has exhibited at seven conferences (figure 2), predominantly promoting the EBI's resources, training activities and career opportunities. In addition, we have collaborated with OIPA to represent EMBL at exhibitions and careers fairs where a combined presence was beneficial. OTT usually represents the EBI in this capacity but the EBI's Outreach and Training representatives have also played a major role in providing on-stand support at these events and in promoting the EBI at conferences not exhibited at by either the EBI or EMBL. This year we exhibited at the following events:

- Nature Source Event careers fair, London, September 2009
- European Crystallographic Meeting (ECM), Istanbul, August 2009
- Intelligent Systems for Molecular Biology (ISMB)/ European Conference on Computational Biology (ECCB), Stockholm, July 2009
- Federation of the Societies of Biochemistry and Molecular Biology (FEBS), Prague, July 2009 (with EMBL)
- European Proteomics Association (EuPA), Stockholm, June 2009
- MIT European Careers Fair, Boston, January 2009
- Plant and Animal Genomes (PAG), San Diego, January 2009
- HUGO Human Genome Meeting, Hyderabad, September 2008
- ECCB, Sardinia, September 2008

Promotion of training

Louisa Wright, Alison Barker, Janet Copeland, Vicky Schneider, James Watson, Katrina Pavelin, Cath Brooksbank

The Outreach and Training team promotes EMBL and EBI courses by creating promotional slides, posters, flyers and e-mail alerts, directing these promotional tools to appropriate target audiences, and supporting external promotion; for example, we provide regular contributions to EMBL's new course and conference newsletter. In collaboration with the Wellcome Trust Sanger Institute and the Wellcome Trust Course and Conference programmes, we have also launched a new website, www.hinxton.org, which serves as a one-stop shop for information on all scientific events on campus.

Promotion of the EMBL International PhD Programme is supported by the Outreach and Training team, including the provision of summary slides to EBI personnel for inclusion in their presentations, maintaining accurate programme information on the EBI's web pages in liaison with the EMBL Graduate Office and distribution of the PhD programme brochure at conferences and events.

Open day

James Watson, Alison Barker, Katrina Pavelin, Louisa Wright, Vicky Schneider, Cath Brooksbank

We now hold two open days each year, in March and November. The November event was introduced to more effectively support recruitment for the EMBL International PhD Programme. The open days combine lectures with a demonstration session involving a selection of the EBI's main bioinformatics resources. The lectures provide an overview of the EBI's activities, an introduction to the EMBL International PhD Programme, and an insight to our research, while the demonstration session gives the students the opportunity to explore the databases and tools of most relevance to their work.

Public engagement

Louisa Wright and volunteers from the EBI in collaboration with the Wellcome Trust Sanger Institute

In previous years the EBI, in partnership with the Wellcome Trust Sanger Institute, has been a regular contributor to the 'Biology Zone' at Cambridge Science Festival's 'Science on Saturday' event. In 2008 the Biology Zone received more than 2,500 visitors over six hours and, due to the popularity of the event overwhelming the venue's capacity restrictions, there was a queue to enter the building for most of the day. To engage those waiting to enter the Biology Zone, we took a different approach in 2009, utilising our most mobile and popular hands-on activities to entertain and inform families and individuals throughout the day as they waited in the queue.

In 2009, the Biology Zone was estimated to receive over 2,000 visitors. EBI and Sanger Institute volunteers presented a range of hands-on activities including:

- **What's Your Number?:** an activity which saw four physical traits surveyed and discussed with visitors, providing an opening for discussion of genetics and variation across populations;

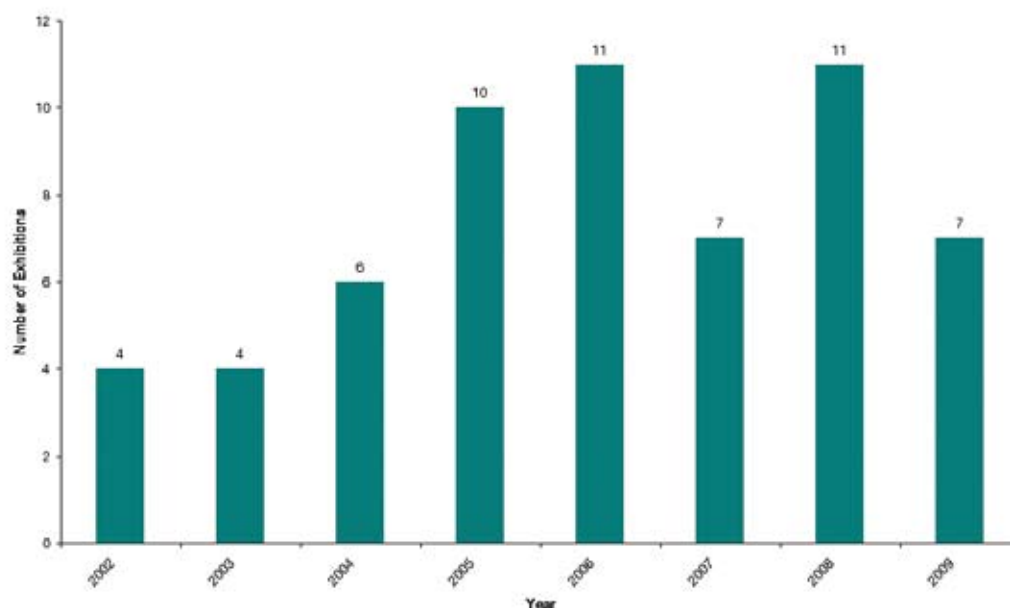


Figure 2. Conferences at which EMBL-EBI has exhibited, 2002–2009.

- **guessing genome sizes:** participants ranked the genome sizes of different organisms (including mosquito, nematode worm, zebrafish and human);
- **origami DNA:** participants folded their own DNA helix;
- **sequence bracelets:** participants threaded a double-stranded bracelet of red, green, yellow, or blue beads according to short lengths of sequence from a range of organisms, including hissing cockroach, chimpanzee, flesh-eating bacteria and human (based in the café area);
- **VoxPops:** due to the success of the 2008 pilot VoxPops project involving film-making and interviewing visitors in the Biology Zone, we repeated this in 2009, offering visitors the opportunity to comment on personal genome sequencing and the directions that genome research should take. Filming training and direction was provided by Jonathan Sanderson, a freelance film producer specialising in science-themed projects. The footage gathered during the day was used to produce a VoxPop video.

Feedback from festival visitors and event coordinators was extremely positive, with the festival organisers reporting a significant drop in the number of complaints relating to waiting time. As a measure of impact, the 'What's Your Number' activity surveyed 434 people through the course of the day, indicating that just one of our activities was successful in engaging over one fifth of the Biology Zone's visitors.

London International Youth Science Forum visit

Louisa Wright with support from Cath Brooksbank, Sandra Orchard, Phil Jones and Bert Overduin

We were involved in the 50th London International Youth Science Forum (LIYSF), a two-week programme of scientific topics and visits for over 300 young scientists (17–21 years old) from more than 40 countries. OTT hosted a group of 24 students at EMBL-EBI as one of the visits to a scientific organisation. This provided the students with the opportunity to learn about bioinformatics and how researchers can use the resources to solve biological questions. To illustrate this concept in practice, a hands-on activity was developed for the group based around sequence homology and protein structure of a H1N1 virus component. The visit also included a tour of the Wellcome Trust Sanger Institute sequencing centre for the group to witness the technology behind next-generation sequencing.

TRAINING

Vicky Schneider, James Watson, Victoria McKenna, Cath Brooksbank

Our team coordinates EMBL-EBI's user-training programme, which aims to provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators in EMBL member states and beyond. Our objective is to provide trainees with the necessary skills to efficiently and effectively use Europe's core biological data resources. Our trainees span wet lab experimentalists and computational scientists. We are also beginning to see an increasing demand for training from clinical researchers. We have worked hard over the past two and a half years to develop a programme that caters for our full range of users. We have also expanded our programme to trainers and educators so that organisations receiving training from us will be able to realise long-term benefits from our training programmes. We aim to support Europe's community of bioinformatics trainers by widening the use of Europe's bioinformatics resources, and have started several projects in this direction.

The EMBL-EBI user-training programme is organised under the umbrella of EICAT, the EMBL International Centre for Advanced Training, which coordinates training activities for scientists at different levels. Our internal training is undertaken through EMBL's General Training and Development Programme, which is coordinated by EMBL's Personnel team.

Types of training

Our user-training programme is based on three different elements:

- **hands-on user training courses:** held in our purpose-built IT training suite, the hands-on courses are structured to suit all levels, from familiarising end users with EBI resources and tools to instructing power users how to programmatically access EBI core resources.
- **the Bioinformatics Roadshow:** the EBI takes training to users by creating and delivering training programmes tailored to the needs of the users of Europe's main data resources;
- **the EBI eLearning project:** we are developing a web-based portal that will serve both end users and trainers.

All three training elements have developed substantially during the past year and demand for our training programme continues to increase. As awareness of our training programme grows, we have become increasingly conscious of the diverse training needs of different research fields (e.g. plant bioinformatics, medical research) and different levels of experience (e.g. masters and PhD students, postdocs, senior scientists).

The user-training programme is only possible thanks to the Outreach and Training representatives (OTRs) from the various groups at the EBI. A core group of OTRs contributes substantially both to in-house and to off-site courses, and the total number of EBI staff members involved in training numbers more than 60 – 15% of the EBI's staff base. The expertise of our trainers is showcased in a new trainer gallery: www.ebi.ac.uk/training/trainers/. An alphabetical list of OTRs who have contributed to the past year's hands-on courses and Roadshows can be found in Panel 1.

Hands-on user training: trainees come to us

James Watson, Vicky Schneider, Alison Barker, Janet Copeland, Cath Brooksbank

The hands-on user training programme has run 21 courses since September 2007. Up to August 2008 the average attendance per course was 20 participants; this has now increased to an average of 29 participants per course (numbers ranged from a minimum of 15 to a maximum of 36). This year the hands-on programme trained more than 240 people from all over the world, with half of the trainees travelling from outside the UK (figure 3a). Although this means that half of our attendees work in the UK, the nationalities we reach are wide ranging (46 different nationalities this year) with 33% of all trainees being UK nationals. This can be seen in figure 3b.

The hands-on user training courses from September 2008 to August 2009 can be categorised as follows:

Focused courses on a specific topic or field:

- Interactions and Pathways: towards a whole system perspective
- Sequence to Genes: genome informatics of microorganisms
- Joint EMBRACE-EBI workshop: understanding protein structures
- Transcriptomics
- Joint EBI-Wellcome Trust Proteomics workshop
- Programmatic access to biological databases (Java or Perl).

Generic courses that cover many EBI data resources and tools, providing a general overview to the trainees:

- A two-day dip into the EBI's data resources: understanding your data
- A walk through EBI bioinformatics resources

In this way we can provide courses tailored to experimental researchers and computational biologists, as well as supporting training on the latest advances and techniques in molecular, biomedical and bioinformatics research.

Details on forthcoming courses can be found at www.ebi.ac.uk/training/hands-on/. A full list of previous courses is available at www.ebi.ac.uk/training/hands-on/previous.html.

Our training web pages have undergone major reorganisation in response to user feedback, with the support of the External Services team. This is a continually evolving process and the pages will be developed further to incorporate all training events open to external users.

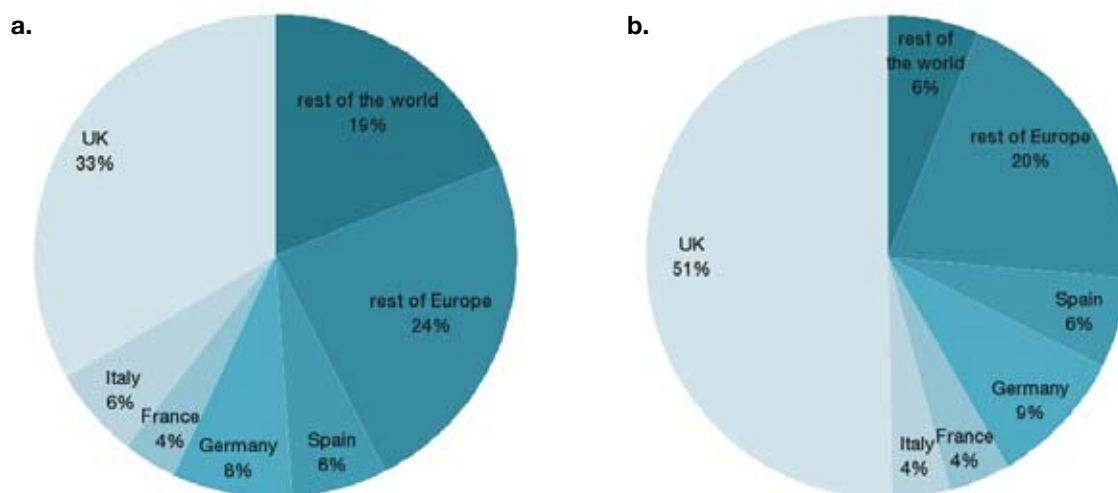


Figure 3. Breakdown of trainees according to a) the country they worked in at the time of attending a hands-on course, and b) their nationality (September 2008-August 2009).

We have continued to respond to feedback from trainees and trainers, which is collected using online evaluation forms covering all areas of the course: content, administration, trainer quality and the materials provided. The general feedback in all these areas has been positive: 92.5% of those completing the evaluation form rated the courses 'useful' or 'extremely useful' to their work and 91.3% of them would recommend our courses to their colleagues. We have trialled providing a CD containing the course materials, in addition to or instead of a printed training manual. In response to our trainers' requests, we now make all training materials publically available online.

Over the past year we have streamlined our procedures for maintenance, management and coordination of the IT training room, including health and safety (James Watson) and the implementation of a new registration system consistent with that used by the EMBL Course and Conference Office (Janet Copeland). OTT is increasingly involved in supporting events other than hands-on courses. Future efforts will define a structure to provide a consistent level of support to anyone organising an event in the IT training room, which is now reaching capacity with an average occupancy of 82% each month. We are especially grateful to the hard work and support received from the EBI systems team for helping us to maintain our hardware and update our software.

We have started to investigate a new promotional procedure in consultation with the EMBL Course and Conference Office. This allows us to reach and target specific users to make them aware of courses potentially relevant to them. We are extremely grateful to all those, both within and beyond the EBI, who have actively supported us by helping to advertise the hands-on programme.

Roadshows: we travel to the trainees

Vicky Schneider, Janet Copeland, James Watson, Cath Brooksbank

The successful completion of our training deliverables for the FELICS Integrating Infrastructure Initiative, and new financial support through the EU-funded SLING Integrating Action, have brought some changes to the roadshow programme. The SLING grant continues our partnerships from FELICS, with the EBI working in collaboration with the Swiss Institute of Bioinformatics (SIB), the European Patent Office (EPO) and the BRENDA database at the University of Braunschweig. SLING provides funding for our trainers to travel to host institutes, enabling them to host a roadshow with minimal costs. This allows us to reach out to areas of Europe we were unable to target before. We have defined a strict set of eligibility criteria (see www.ebi.ac.uk/training/roadshow/sling_new.html) to ensure that our SLING funds are targeted to researchers in the newer EU member states and those who are only just beginning to make use of Europe's core data resources. The underlying concepts and aims of the roadshows remain the same: to combine presentations and hands-on practical sessions to guide users through selected databases and tools by expert trainers. The roadshow web pages have also been redesigned to provide a full list of past, ongoing and future roadshows (www.ebi.ac.uk/training/roadshow/).

The demand for roadshows remains strong: we have run 18 roadshows from September 2008 to August 2009 and are already taking bookings for roadshows in 2011. Although our priorities lie with SLING roadshows, where possible we are continuing to train researchers beyond SLING's target zones. In these cases the costs associated with trainer travel are borne by the host institution.

Roadshow logistics and administration are managed by OTT's Logistics and Admin sub-team. The scientific programme for each roadshow is coordinated by the training programme project leader Vicky Schneider with support from scientific training officer James Watson.

eLearning

Vicky Schneider, Cath Brooksbank, James Watson, Victoria McKenna

The eLearning pilot project has proved to be an excellent starting point to acquire a complete understanding of what it will entail to build, launch and maintain a comprehensive elearning portal for EMBL-EBI. This has led to the creation of a new position devoted specifically to this challenge, filled in June 2009 by our eLearning content developer Victoria McKenna.

Through the pilot project we had the chance to gather essential feedback from the trainers who built our pilot training materials, the trainees using the courses and the technical staff involved with the installation and maintenance of the existing platform. After a thorough survey of the e-learning platforms available and the extent to which they meet the EBI's requirements for such a platform, we have now begun to test how well these meet the requirements of trainees, trainers and our web team. The next challenge will be to consolidate the disparate e-learning resources currently available in different parts of the EMBL-EBI website, and to develop a robust means of adding and maintaining courses of a consistently high standard.

Monitoring user training in 2009

James Watson, Holly Foster, Vicky Schneider, EBI's Administration Team

Over 138 EMBL-EBI personnel have participated in 280 unique training-related events during 2008–2009 consisting of: 110 demonstrations/hands-on training events, 73 posters and 220 presentations. Including the audiences reached through demonstrations and talks at conferences, we have reached more than 24,000 people. Audiences are predominantly PhD students, postdoctoral researchers and other academics, but there has also been a substantial amount of outreach to industrial researchers. The tracking of all training-related events remains a non-trivial and substantially time-consuming task; therefore we will be reviewing the systems used to collect these statistics and improve future data collection.

Staff and student education

Vicky Schneider, James Watson

Our trainers: in March 2009 we piloted our first 'user training feedback forum', where a selected group of OTRs together with the training programme project leader, scientific training officer and freelance training expert Frances Scott spent a day discussing training techniques and strategies. As a result of this forum, OTT has facilitated the creation of a shared space for trainers to exchange materials and ideas, and all the delegates came away with a shared understanding of how to get the most out of interacting with their trainees.

EMBL EBI PhD student training: OTT continues to provide 'Primers for Predocs', a series of events designed to familiarise our new PhD students with the data resources and tools available to them at the EBI. In 2009, 'Primers for predocs' comprised a week of lectures and practical sessions. Feedback from this event has influenced the next series, which will focus more on accessing data programmatically. OTT also supports the EBI's second year PhD students in the organisation of, and teaching for, an annual bioinformatics course run specifically for the EMBL PhD students.

ELIXIR training

Vicky Schneider, Cath Brooksbank, James Watson, Bren Vaughan, in collaboration with the ELIXIR training committee

We have been involved in developing the bioinformatics user-training strategy for ELIXIR (www.elixir-europe.org). The stakeholder consultation phase of ELIXIR has now concluded and the full report, containing the ELIXIR training committee's recommendations and the rationale behind them, is available at www.elixir-europe.org/files/documents/reports/WP11-Training_Strategy_Committee_Report.pdf. Initial research for the report involved gathering information from trainers and trainees throughout Europe on regularly held bioinformatics user-training courses. With technical support from the EBI's External Services team we created an online database for collecting this data. One of its outputs is a map that displays where the courses are held, and another is a calendar of user-training courses available in Europe, see www.elixir-europe.org/page.php?page=user_training_map and www.elixir-europe.org/page.php?page=calendar. Although the stakeholder consultation phase is now complete, we will continue to maintain the database as a service to the bioinformatics user community.

Course funding and support

Vicky Schneider, James Watson, Cath Brooksbank

We continue to seek external funding for our user-training courses; this enables us to invite external trainers and to subsidise our course fees, making our courses accessible to early-stage researchers from around the world. We are delighted to have received EMBO funding for two new practical courses in 2010: one on systems biology and one on structural bioinformatics. We have now run three joint EBI-Wellcome Trust Proteomics workshops and OTT will be leading future applications for similar courses in the future.

The Experimental Network for Functional Integration (ENFIN) project has sponsored several of our hands-on courses and will continue to support future courses.

EBI training collaborations

Vicky Schneider, James Watson, Cath Brooksbank

We are working with OIPA, EICAT and the EMBL Course and Conference Office to pool resources and share expertise whenever this is practical. We have developed similarly strong relationships with the Wellcome Trust Course and Conference Programmes. A new campus-wide portal (www.hinxton.org) was launched and now provides a directory of scientific events on the Wellcome Trust Genome Campus. We continue our fruitful established collaborations with the University of Cambridge, the Gulbenkian Institute, ICGEB and other organisations that hold regular bioinformatics training courses with EBI trainers regularly contributing to these events.

FUTURE PROJECTS AND GOALS

Our future goals revolve around increasing accessibility to Europe's most widely used data resources, for an increasingly diversifying target audience. We are optimistic that ELIXIR, Europe's nascent infrastructure for biological data, will provide a stable framework for access to high-quality bioinformatics training for Europe's life scientists; new users will come from all the other ESFRI biomedical science infrastructures and we are already beginning to explore new ways to meet their training needs. For example, in collaboration with the other ESFRI biomedical science infrastructures we have just completed the final stages of contract negotiations to create EMTRAIN, a new network funded by the Innovative Medicines Initiative to support the training needs of medicines research. SLING, ELIXIR and EMTRAIN share a major challenge – that of making training materials available to all in ways that transcend the ever-more complex landscape of data types and formats. We envisage that many of the data management and integration problems familiar to the EBI's services teams, which are gradually being overcome with the careful use of controlled vocabularies and standards, will also have to be solved for the training and education community. We are fortunate to have a foot in both camps.

Another challenge is building sustainability into our training programmes. We plan to achieve this by supporting others to train on our behalf, and have begun pilot projects both with our previous roadshow hosts and with a small community of undergraduate and master's degree lecturers. Supporting our own trainers, and providing a development plan for new ones, is an important part of this goal.

Finally, the planning phase of ELIXIR has made us acutely aware of our need to raise ongoing support, on behalf of the entire life science community, for the EBI and for ELIXIR. This will involve building relationships with new target groups and developing new ways to reach them.

Team Members

Outreach Programme Project Leader

Louisa Wright

Scientific Outreach Officer

Katrina Pavelin*

Training Programme Project Leader

Vicky Schneider

Scientific Training Officer

James Watson

eLearning Content Developer

Victoria McKenna*

Workshops and Exhibitions Organisers

Janet Copeland
Holly Foster*

Team Secretary

Alison Barker

* Indicates part of the year only

Panel 1

EMBL-EBI Outreach and Training representatives

These members of personnel are embedded within the service teams and research groups.

Rafael Alcántara (ChEBI)

Jeff Almeida-King (Ensembl Genomes)

Bruno Aranda (IntAct)

Richard Côté (OLS, PICR)

David Croft (Reactome)

Paula De Matos (ChEBI)

Jennifer Deegan (GO)

Emily Dimmer (GOA)

Ibrahim Emam (ArrayExpress, Atlas)

Anton Enright (RNA resources)

Nicholas Furnham (CSA)

Janna Hastings (ChEBI)

Gemma Holliday (MACiE)

Alan Horne (programmatic access)

Sarah Hunter (InterPro)

Rachael Huntley (GOA)

Bijay Jassal (Reactome)

Andy Jenkinson (programmatic access)

Rafael Jimenez (EnCore, DAS)

Phil Jones (PRIDE)

Steven Jupe (Reactome)

Misha Kapushesky (ArrayExpress, Atlas)

Samuel Kerrien (IntAct)

Paul Kersey (Ensembl Genomes)

Michael Kleen (UniProt web services)

Roman Laskowski (PDBsum,

ProFunc, Druggability portal and

Thornton group resources)

Duncan Legge (UniProt)

Jane Lomax (GO/GOA)

Rodrigo Lopez (EBI search tools)

Tim Massingham (next generation

sequencing)

Jennifer McDowall (InterPro, UniProt)

Johanna McEntyre (Literature Services)

Hamish McWilliam (EBI search tools, sequence searching)

Anika Oellrich (EBI text mining tools)

Tom Oldfield (PDBe)

Sandra Orchard (UniProt, InterPro, IntAct)

Bert Overduin (Ensembl)

John Overington (ChEMBL)

Samuel Patient (UniProt web services)

Mari Luisa Pellegrini-Calace (membrane proteins)

Sharmila Pillai (CiteXplore)

Syed Asad Rahman (SMSD)

Dietrich Rebholz-Schuhmann (EBI

text mining tools)

Florian Reisinger (EnCore, EnVision)

Peter Rice (EMBOSS)

Gabriella Rustici (ArrayExpress, Atlas, Bioconductor)

Gaurav Sahni (PDBe)

Vicky Schneider (EBI overview and tools)

Michael Schuster (Ensembl)

Sanchayita Sen (PDBe)

Giulietta Spudich (Ensembl)

Jawahar Swaminathan (PDBe)

Glen Van Ginkel (PDBe)

Sameer Velankar (PDBe)

Juan Vizcaino (PRIDE)

James Watson (Tempura, PDBsum, ProFunc)

Andy Yates (Ensembl programmatic access)

Dominic Clark

Industry Programme Manager

PhD Medical Informatics, 1988.

Senior Research Fellow, Biomedical Informatics Unit,

ICRF, 1987–1995.

UK Bioinformatics Manager, GlaxoWellcome R&D Ltd,

1995–1999.

Vice-President Informatics, Pharmagene, 1999–2001.

Managing Consultant, Sagentia Ltd, 2001–2009.

At EMBL-EBI since 2006.



John Overington

*EMBL-EBI research
lead for industrial
interactions*



Industry Support

159

For the last twelve years the Industry Programme has been an important and integral part of EMBL-EBI, providing ongoing and regular contact with a distinct group of users. The last two years in particular have seen an expansion in the size of the programme, with companies in different sectors, such as healthcare and consumer goods, beginning to access and exploit our bimolecular data. Industry represents at least 20% of our usage, and it is clear that EMBL-EBI serves a major role in distributing data to fuel the life science industries.

A more detailed brochure describing the activities and benefit of the EMBL-EBI Industry Programme can be downloaded from www.ebi.ac.uk/Information/Brochures.

THE EMBL-EBI INDUSTRY PROGRAMME

The Industry Programme was set up in 1996 and is now well established as a subscription-funded programme for larger companies. There are currently 16 members (Panel 1).

At our regular meetings with industry partners, we provide updates on activities at EMBL-EBI, to ensure that partners are kept well informed of progress and future developments. In addition, we provide customised hands-on training so that these users can best exploit the publicly held data for the discovery of new therapeutics, vaccines, consumer products and agrochemicals. However, the interaction fostered by the programme is two-way and we equally benefit from the insight and expertise of our industrial members, which helps to define and shape the services we provide.

Most recently, the influence of the programme partners has been critical in establishing cheminformatics and chemogenomics resources at EMBL-EBI, stressing the need to link the biological data of life with the chemistry of the small molecules and drugs used for intervention. These new resources help to address the 'druggability' of potential targets and the integration of different types of data, including the literature, to obtain a systems perspective.

Going forward we see our interactions with the industry partners growing even stronger, as the flood of data continues to rise and the need of industry to outsource and utilise public resources stimulates precompetitive research collaborations, increased use of open source software and standards development.

Through efforts such as the Innovative Medicines Initiative (IMI; <http://imi.europa.eu>) and the Pistoia Alliance (www.pistoiaalliance.org), we are keen to support and encourage this transition. The IMI is a joint initiative to foster pre-competitive collaboration within the pharmaceutical industry co-sponsored by EFPIA (European Federation of Pharmaceutical Industries and Associations), and the Pistoia Alliance is a not-for-profit organisation established by life science companies. In addition to these initiatives, industrial involvement in ELIXIR, the emerging pan-European infrastructure for biological information, is essential for its success. It is our view that by working closely together with industry we can empower research and enable better translation of research discoveries into new advances in medicine, health, and agriculture for the benefit of society.

Industry Programme workshops

The Industry Programme continues its coordination and organisation of high-quality facilitated workshops and symposia, providing expert-level presentations and strategic discussion opportunities for members. Workshops typically include key opinion leaders and stakeholders from the research and industry communities. The subject areas for the workshops are prioritised by the member companies. The workshops organised in 2009 are listed in table 1.

Dates	Title
2–3 February 2009	Virtual Physiological Human: anatomy and modelling
30 March – 1 April 2009	Semantic enrichment of scientific literature
20–21 April 2009	Chemistry Development Kit (CDK)
5–6 May 2009	Biological networks: reconstruction, analysis and modelling
9 June 2009	Industry workshop with OBO (Open Biomedical Ontologies) Foundry coordinators
10–11 June 2009	Genotype to phenotype: challenges and resources
10 July 2009	DrugEBllity
23–24 September 2009	Epigenetics: technology, tools and applications of epigenetic data
29–30 October 2009	New sequencing technologies, informatics resources and application in metagenomics
23–24 November 2009	Toxicogenomics & toxicoinformatics

Table 1. *Industry Programme workshops in 2009.*

Strategy meetings

The quarterly strategy meetings at EMBL-EBI are the principal forum for communication and prioritisation, allowing members to:

- review the current status and future directions of the programme activities;
- receive updates on EMBL-EBI's activities and strategy;
- review priorities for workshops and pre-competitive areas.

These meetings allow the programme to remain aligned to the industry partners' requirements and also provide time for detailed interactions between EMBL-EBI staff and representatives from member companies.

Pre-competitive activities

Through sponsorship (grants) from a number of the member organisations we have initiated some small scale development projects which will potentially benefit all EMBL-EBI users. We are also working closely with the Innovative Medicines Initiative and the Pistoia Alliance to provide an open foundation of data standards, ontologies and web services to streamline the Pharmaceutical Drug Discovery workflow (chemistry, biological screening, logistics) through common business terms, relationships and processes.

Within the IMI call 1, we are partners on the eTox project (integrating bioinformatics and chemoinformatics approaches for the development of expert systems allowing the *in silico* prediction of toxicities) through the ChEMBL group (page 61), and the EMTRAIN project (establishment of a network to facilitate and coordinate European training and education relevant for stakeholders of medicines research and development, www.emtrain.eu) as part of the development of the EBI's training programme in the field of medicines research.

During 2010, we anticipate that the programme's involvement in both initiatives will continue as we are involved in applicant consortia for new knowledge management within the IMI second call, and work closely with the Pistoia Alliance in the area of semantic information integration.

Building upon the 2008 Druggability Portal pilot, responsibility for this activity has now transitioned to the ChEMBL group who are also progressing a number of pre-competitive data sharing activities.

SUPPORT FOR SMES

SMEs have access to EMBL-EBI training, services and support. In addition, EMBL-EBI has established an annual series of information workshops designed to allow SMEs to effectively use the freely available data resources, tools and services provided by EMBL-EBI and its collaborators (including the European Patent Office).

The third Annual Information Workshop on European Bioinformatics Resources for SMEs was held in Vienna on 3–4 September 2009 with kind support from Austria Wirtschaftsservice (AWS) and presentations from EMBL-EBI, the European Patent Office, AWS and SMEs based in Austria.

The workshop agenda built upon earlier experience and included small molecule resources (ChEMBL and ChEBI), proteomics services, literature services, web services and the patent searching services providing both by EMBL-EBI and the European Patent Office.

Full details of the meeting and the presentations can be accessed from www.enfin.org/page.php?page=sme_meeting_2009. Due to the popularity of this event it has been decided to continue organising such events in 2010. An appropriate location and host organisation will be identified through discussion with members of the Council of European BioRegions (www.cebr.net/NewsStory.aspx?id=26).

SMEs wishing to receive information about future events designed for SMEs, should subscribe to the information service via the link on www.ebi.ac.uk/industry/SME.

Panel 1

EMBL-EBI's Industry Programme partners in 2009

AstraZeneca
Bayer Schering Pharma AG
Boehringer Ingelheim Pharma GmbH
& Co. KG

Eli Lilly & Company
Galderma
GlaxoSmithKline
F. Hoffmann-La Roche
Johnson & Johnson Pharmaceutical
Research & Development
Merck Serono S.A.
Nestlé Research Centre

Orion Pharma
Philips Research
Pfizer Ltd
Sanofi-Aventis Recherche &
Développement
Syngenta Limited
Unilever



Petteri Jokinen

*MSc in Computer Science
1990, Helsinki University.
At EMBL-EBI since 1996.*

Systems and Networking

163

INTRODUCTION

The Systems and Networking team manages the EMBL-EBI IT infrastructure. This includes compute and database servers, storage, desktop systems and networking, as well as managing our campus connection. An important task is supporting EMBL-EBI users in their daily activities. The team works closely with all project groups maintaining and planning their specific infrastructures. The IT environment holds five petabytes of disk storage (figure 1) and consists of more than 9,000 CPU cores (figure 2) and. Looking ahead we expect these figures to double again in 2010.

EXTERNAL DATA CENTRE

In order to reduce the risk of losing data, we have set up an external replication site. We clone our data once a day to this new site via a dedicated 10GB/s connection. This has also given us more room in the campus data centre for further equipment.

NETWORKING

Our primary internet connection is now via London and we have an active backup link to Cambridge. The primary connection is 10GB/s and the backup link is 1GB/s. We have also undertaken major work in our internal network, increasing the scalability and speed. Systems were also involved in the EBI refurbishment, which included a complete network re-wiring of the main building and a move to Voice over Internet Protocol (VoIP) telephones.

NEW MACHINES

We have reduced the number of farms by combining research and production farms into one, called the EBI farm. We have added a significant number of new machines (approximately 350) and are planning to further increase the computational power by the end of the year. We are currently in the process of evaluating large memory (512GB) machines and are expecting to obtain at least five of these machines by the end of the 2009.

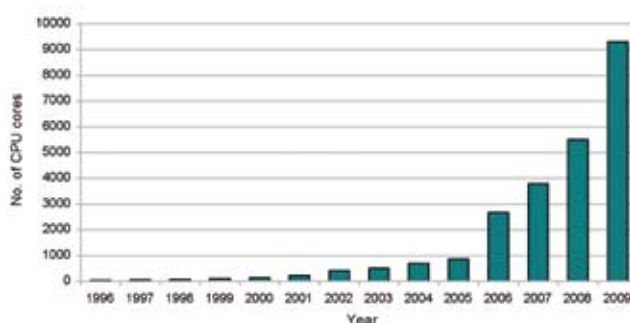
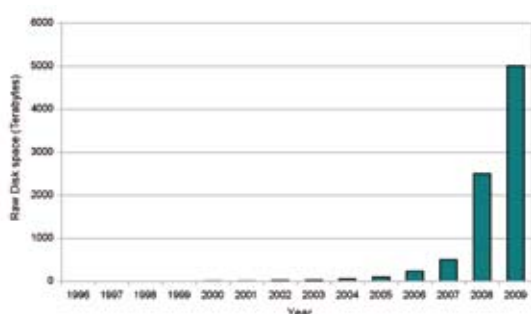


Figure 1 (left). Disk space; Figure 2 (right). CPU cores.

STORAGE

Once again the storage has more than doubled this year (figure 1). As well as expanding our storage we are continuously trying to improve the underlying architecture. This year we have placed emphasis on consolidating to fewer storage solutions thus making the management and use of the underlying architecture simpler for the inevitable future growth.

DATABASE SERVER INFRASTRUCTURE

We have installed more than twenty new database servers, including Linux-based Oracle Real Application Clusters. We have installed a new SAN infrastructure that spans the on-site data centre. This can be used for both Oracle and MySQL database servers. We have significantly increased the Oracle backup space.

DESKTOPS

The number of users in EMBL-EBI has exceeded 400 this year. We have continued to automate as many processes as possible with notable successes in Windows application deployment and desktop provision across all operating systems. We have increased the support we provide to Mac OS X and Linux desktop users by expanding the desktop team with two new specialist positions in these areas.

ONGOING PROJECTS

- In 2008 we had a new project to build a scalable trace archive and we accomplished this successfully. This year we have doubled the size of this archive and completely re-written the software that is needed for projects (e.g. 1000 Genomes Project) that store data to this archive.
- Another significant project this year has been to completely replace the infrastructure that is used by the PDBe team.
- On the desktop side we have migrated our Windows infrastructure to VMware, providing fault tolerance and increased availability as well as reducing data centre costs by reducing the physical infrastructure.

NEW PROJECTS

- We have two experimental projects that have developed this year; Hadoop farm and an Amazon compatible EBI cloud. Hadoop is a software framework that supports data-intensive distributed applications. It enables applications to work with thousands of nodes and petabytes of data. The EBI cloud is implemented using open source software. The purpose of these projects is to evaluate new technologies that can be used in future EBI projects.
- We are in the process of building a scalable and resilient email set-up.
- We have begun an exciting new project that involves building scalable infrastructures including the use of several external data centres.

SUMMARY

This year's growth has been enormous and subsequently so has the team's workload. There has been a large quantity of equipment to install and maintain as well as an ever-increasing number of people to support. We do our best to deploy high-grade solutions.

Team Members

Server and Networking team

Jonathan Barker
Elizabeth Beresford*
Gianluca Busiello*
Gavin Kelman
Manuela Menchi
Pravin Patel

Radoslaw Ryckowski
Michal Wieczorek*

Desktop team

William Barber
Richard Boyce
Karen Briggs
Andy Cafferkey
John Livingstone

Software Engineer

Ville Silventoinen

Technical Administrator

Carolina Bejar

** Indicates part of the year only*



Rodrigo Lopez

*Vet. Med Degree 1984, Oslo Vet. Høyskole.
NASAS Cand. Scient. Molecular Toxicology and Informatics 1987,
University of Oslo.
At EMBL-EBI since 1995.*

External Services Team

165

INTRODUCTION

The External Services (ES) team focuses on three major areas: web development, web administration and services frameworks. Web development is mainly concerned with the development and maintenance of websites and core publishing frameworks for both the EBI as well as various EU-funded projects. Web administration focuses on the provision of robust and stable service architectures for the EBI's databases and services. Services frameworks comprise tools for generating core services, such as the EBI search engine, EB-eye, and applications for the maintenance of core analytical tools, especially in the domain of nucleotide and protein sequence analysis.

WEB DEVELOPMENT

The team is responsible for the main EBI web portal as well as several Wellcome Trust, BBSRC and EU-funded projects websites. These include: 1000 Genomes Project, BioCatalogue, BioSapiens, DVGA, European Genome-phenome Archive (EGA), ELIXIR, EMBRACE, ENA, ENFIN, FELICS, SLING, IMPACT, INSDC, MIBBI and SYMBIOmatics. We are also responsible for the management of shared portals within the context of the Hinxton Sequence Forum (HSF) initiative and also run the 2can training and support website in collaboration with the Outreach and Training team and the general EBI staff.

The current publishing frameworks were revised during late 2008 and a final review of the existing PHP-based Content Management System was made. The major result of this revision was to consolidate all the CMSs in used and identify a system that will scale and provide higher degrees of freedom between users and reduce maintenance overheads. To this end, the group is currently working on a migration plan that uses DRUPAL (www.drupal.org) as the core system for both www.ebi.ac.uk as well as all the EU project portals.

The web developers, in collaboration with the University of Manchester, launched the BioCatalogue, a curated catalogue of life science web services (www.biocatalogue.org), during ISMB 2009 in Stockholm, Sweden. This project, funded by the BBSRC, focuses on capturing, annotating and cataloguing web services in bioinformatics from around the globe. At present, the catalogue contains more than 1,000 SOAP and REST services (DAS services are in progress), and counts more than 100 registered users. About one third of these services are provided by EMBL-EBI and partners, in particular from the EMBRACE project.

Monitoring, improving and supporting use of EBI services

Monitoring and reporting the activity of the various web portals is also a responsibility of the team. The web administrators have recently launched a new usage reporting system based on the popular AWSTATS package. This new framework completely replaces the old Analogue-based system we have been using since 2002. It comprises a MySQL back-end, and a comprehensive web and FTP logs analysis infrastructure. A simple interface is available at present that allows users to browse and review reports daily, monthly and annually. A user interface that will allow complex queries on the statistics in the database is under development.

There are more than 200 services running on a web farm, which currently comprises some 30 machines. The web farm is divided into two major sections. One is dedicated to traditional Apache documents and cgi-based application serving, while the second is entirely dedicated to Apache Tomcat Java-based services, which serve most of the EBI's databases. During the past year significant investment has taken place to provide robust Java-based services. We have created an improved Tomcat cluster software architecture for increased reliability and ease of use. This has been

achieved by building on experience with earlier and existing production systems and by incorporating greater flexibility when software upgrades to Tomcat, Java and database drivers are required. To ensure the smooth running of the web-based operations running on this architecture, we have also produced extensive documentation on its intended use, promote good practices and provide guidance on troubleshooting techniques. To enhance communication with our users, we have migrated to the RT ticket system, which supersedes the no-longer-maintained Jitterbug.

- During the reporting period, several new services have appeared in the web framework and some are undergoing heavy overhauling. These include:
- Gene Expression Atlas (www.ebi.ac.uk/gxa) from the Microarray group led by Alvis Brazma;
- PDBe team services (www.ebi.ac.uk/pdbe/docs/Services.html), led by Gerard Kleywegt (formerly by Kim Henrick);
- Ensembl Genomes (www.ensemblgenomes.org) from Paul Kersey's group;
- 1000 Genomes Project (www.1000genomes.org) from Paul Flicek's group;
- European Genome-phenome Archive (www.ebi.ac.uk/ega), Paul Flicek's group.

WEB AND FTP STATISTICS

During the reporting period, traffic to EMBL-EBI using both the web and FTP services, continues to grow. Table 1 shows web traffic on www.ebi.ac.uk/, excluding www.ensembl.org which is administered by the Wellcome Trust Sanger Institute. Table 2 shows traffic on <ftp://ftp.ebi.ac.uk>.

Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth(Gb)
Sep-08	253,738	535,815	40,446,668	78,510,761	1,335
Oct-08	351,199	730,988	36,678,223	83,580,948	2,303
Nov-08	330,734	681,377	29,264,905	76,560,727	1,510
Dec-08	247,260	513,642	24,445,052	66,206,357	1,238
Jan-09	268,801	563,334	30,816,478	73,917,327	1,414
Feb-09	275,177	580,883	84,420,621	125,018,308	2,372
Mar-09	311,730	677,916	39,854,717	89,651,978	2,928
Apr-09	278,231	600,623	31,777,257	74,890,131	1,938
May-09	293,526	627,089	48,827,681	89,993,882	2,814
Jun-09	244,843	555,490	48,048,960	89,606,204	2,422
Jul-09	221,843	525,201	54,524,619	94,048,315	3,154
Aug-09	204,980	477,382	48,284,510	80,534,307	2,416
Total	3,282,062	7,069,740	517,389,691	1,022,519,245	25,844

Table 1. Web traffic on www.ebi.ac.uk (excluding www.ensembl.org).

Month	Unique visitors	Number of visits	Hits	Bandwidth
Sep-08	11,919	32,215	4,673,101	18,354
Oct-08	17,298	37,501	6,503,866	13,451
Nov-08	15,152	33,662	5,212,835	12,975
Dec-08	14,644	30,500	4,300,000	17,675
Jan-09	13,002	29,823	5,272,167	14,787
Feb-09	13,092	31,269	5,098,968	14,212
Mar-09	13,820	34,687	7,147,368	18,893
Apr-09	12,648	30,896	4,498,930	12,521
May-09	11,954	29,924	4,352,014	12,894
Jun-09	11,429	28,899	4,087,803	17,379
Jul-09	10,217	28,003	2,084,152	11,697
Aug-09	9,200	25,297	3,691,711	13,741
Total	154,375	372,676	56,922,915	178,578

Table 2. FTP traffic for <ftp://ftp.ebi.ac.uk>.

EBI > Tools > Similarity & Homology > NCBI-BLAST

NCBI-BLAST Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

Alignments

Selection: |

in **fasta** format

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
<input checked="" type="checkbox"/> 1	SP:BRCA2_HUMAN	Breast cancer type 2 susceptibility protein OS=Homo sapiens GN=BRCA2 PE=1 SV=2 <i>Cross-references and related information in:</i> <ul style="list-style-type: none"> Gene Expression Nucleotide Sequences Genomes Ontologies Molecular Interactions Protein Families Literature Macromolecular Structures Protein Sequences Reactions & Pathways 	3418	17672	100.0	100.0	0.0
<input checked="" type="checkbox"/> 2	SP:BRCA2_FELCA	Breast cancer type 2 susceptibility protein homolog OS=Felis catus GN=BRCA2 PE=2 SV=2 <i>Cross-references and related information in:</i> <ul style="list-style-type: none"> Nucleotide Sequences Ontologies Protein Families Literature 	3372	11271	67.0	78.0	0.0
<input checked="" type="checkbox"/> 3	SP:BRCA2_RAT	Breast cancer type 2 susceptibility protein homolog OS=Rattus norvegicus GN=Brca2 PE=1 SV=1 <i>Cross-references and related information in:</i> <ul style="list-style-type: none"> Gene Expression Nucleotide Sequences Genomes Ontologies Protein Families Literature Macromolecular Structures Protein Sequences 	3343	8866	56.0	70.0	0.0

Figure 1. NCBI BLAST results summary table showing additional links obtained via the EB-eye search engine.

WEB APPLICATIONS AND WEB SERVICES

During ISMB 2009, the ES team launched a new version of the dispatcher bioinformatics application framework. This service comprises two modules that 1) provide access control to the computational resources and 2) generate web, SOAP and REST interfaces to consume the services, respectively. The new version of this framework, which is entirely Java based and driven by XML configuration files, represents a single point of maintenance for all core sequence analysis services, such as NCBI BLAST, FASTA, PSI-SEARCH, ClustalW2, Muscle, T-Coffee and recently PRANK, from the Goldman group at the EBI. Figure 1 show a screenshot of an NCBI BLAST search summary table that, for each

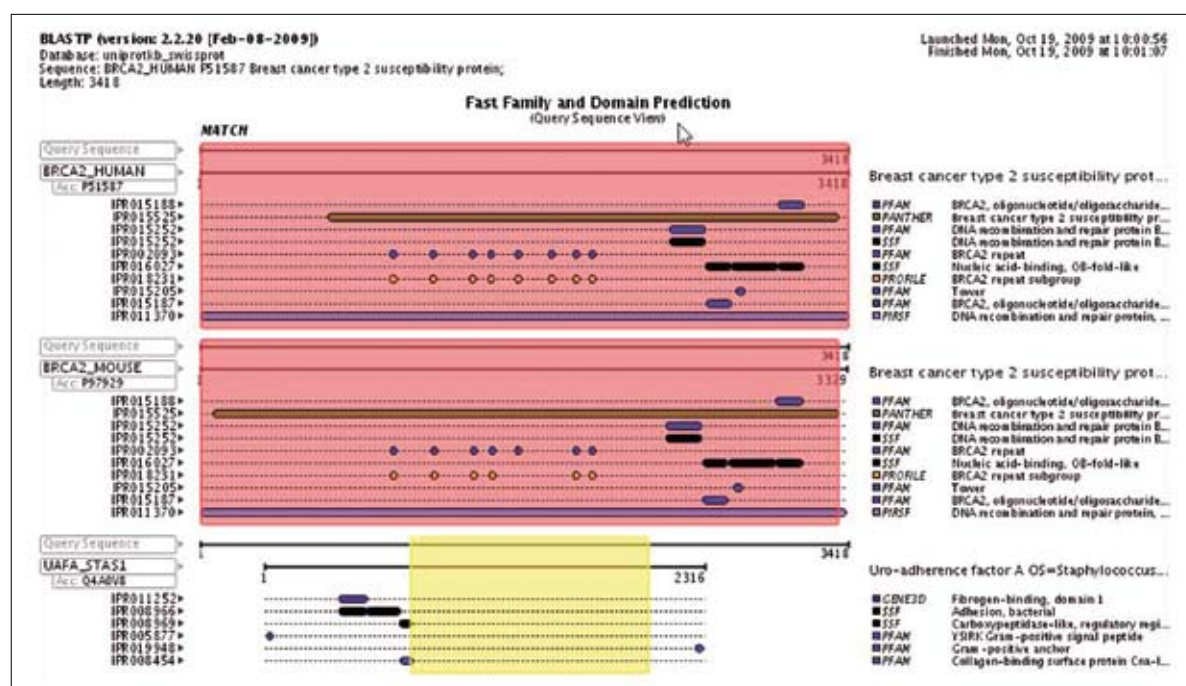


Figure 2. NCBI BLAST results displaying family and domain predictions (according to InterPro). The coloured boxes depict the alignment of each BLAST hit and indicate overlapping functional domains, incomplete coverage or absent domains.

BLAST hit, provides additional links to the data resources indexed in the EB-eye search engine. This feature allows the user to see whether the BLAST hits have additional annotations in other sequence, functional domains, structure, gene expression, ontology and literature databases. Figure 2 shows the functional family and domain architecture in the BLAST database hits. The main access URLs for these new services are: www.ebi.ac.uk/Tools/sss for sequence similarity search applications and www.ebi.ac.uk/Tools/msa for multiple sequence alignment tools.

COURSES AND CONFERENCES

Various members of the group have been engaged in training activities during 2008–2009, both at the EBI and elsewhere. These are listed below:

27 June – 2 July 2009	ISMB/ECCB, Stockholm, Sweden
27–29 April 2009	EBI hands-on training: Programmatic access to biological databases (Java)
16–18 March 2009	EBI hands-on training: From Sequence to Genes: genome informatics of microorganisms
23–24 February 2009	FELICS Bioinformatics Roadshow: Nicosia, Cyprus
18–19 February 2009	FELICS Bioinformatics Roadshow: John Innes Centre, Norwich, UK
9–13 February 2009	EBI PhD training: Primers for predocs
24–27 November 2008	EBI hands-on training: Programmatic access in Java: webservices & work flows
22 September 2008	ECCB Tutorial: Interoperability of bioinformatics software and databases
12 September 2008	FELICS Bioinformatics Roadshow: Munich, Germany
8–11 September 2008	EBI hands-on training: Programmatic access in Perl: webservices & work flows
5 September 2008	FELICS Bioinformatics Roadshow: The Hague, The Netherlands
28 August 2008	EMBRACE: Web Services in Systems Biology
30 July – 1 August 2008	III Congreso Colombiano de Biotecnología

Team Members

Software Engineers

Mickael Goujon
Robert Langlois
Hamish McWilliam
Eric Nzuobontane
Silvano Squizzato

Franck Valentin
Weizhong Li

Web Developers

Asif Kibria
Thomas Laurent
Gulam Patel

Stephen Robinson
Brendan Vaughan

Web Systems Administrators

Jenny Martin
Dietmar Sturmayer*

** Indicates part of the year only*

Section 5

Facts and Figures

169

Services and Research	171
Publications	177
Major Database Collaborations	185
Scientific Advisory Boards	187
External Seminar Speakers	189



Services and Research

SERVICES

- The EBI continues to host the major core biomolecular resources of Europe – collecting, archiving and distributing data throughout Europe and beyond. The services continued to be well used during 2009 (see report on External Services, page 165). By September 2009 there were on average 2,881,000 requests per day (2,709,000 in 2008) – 3,898,000 if Ensembl is included (figure 1).
- From September 2008 to August 2009 all our core data resources grew significantly (figure 2). EMBL-Bank has received and processed more than 2.4×10^{10} bases compared with 1.8×10^9 bases in 2008. In total, the European Nucleotide Archive now contains 8.9×10^{12} bases. In 2009 we have processed 3.6 million UniParc entries (2.1 million in 2008); 60,646 microarray hybridisations (115,000 in 2008); 7,174 macromolecular structures (5,649 in 2008) and eight new eukaryotic genomes in Ensembl (12 in 2008). Ensembl now holds 57 eukaryotic genomes and the newly launched Ensembl Genomes resource holds 147 non-vertebrate genomes.

RESEARCH

- The EBI published 218 papers between September 2008 and August 2009 (compared with 232 in 2008), 97 from research groups (78 in 2008).
- The research group leaders have successfully applied for support for their research, totalling €1.3 million over the next 2–4 years.

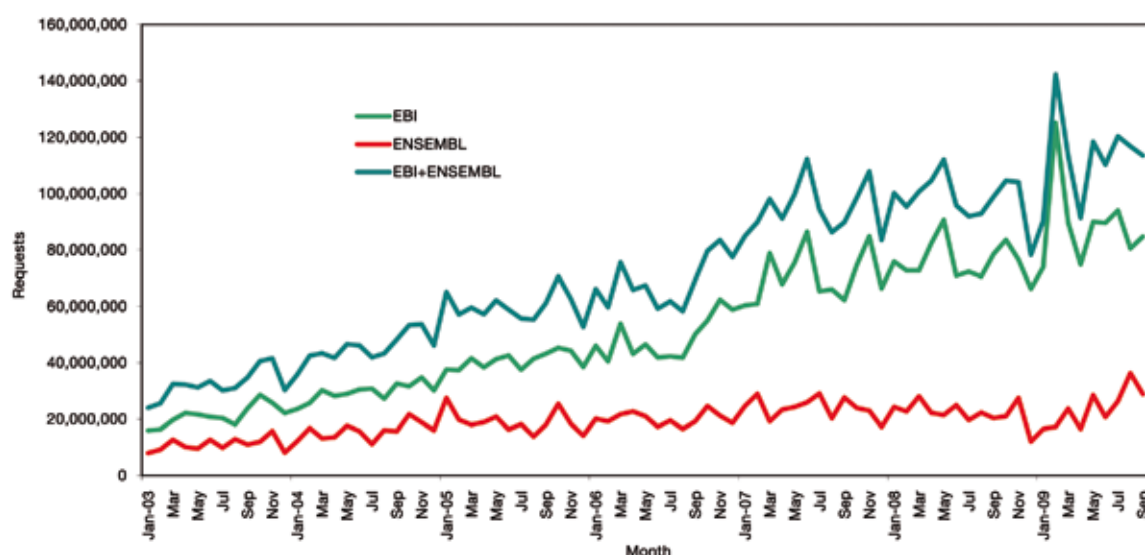


Figure 1. A plot of the web requests received by the EBI and Ensembl from January 2003 to September 2009.

- The number of pre-doctoral students at EBI is currently 34 (35 in 2008). This includes seven PhD students who started their studies in October 2009. Eleven students submitted their theses in 2009.

OUTREACH AND TRAINING

- We have taken part in 280 training events throughout 2009 (compared with 330 in 2008), reaching over 24,000 participants.
- In its third year, the EBI's hands-on user training programme organised thirteen courses and trained over 240 researchers, an average of 29 delegates per course.
- The EBI continues to provide extensive training for users of its services off-site, and our Bioinformatics Roadshow programme, run as part of the SLING Integrating Action, is in high demand. We have run 18 roadshows during 2009 and already have 15 scheduled for 2010.

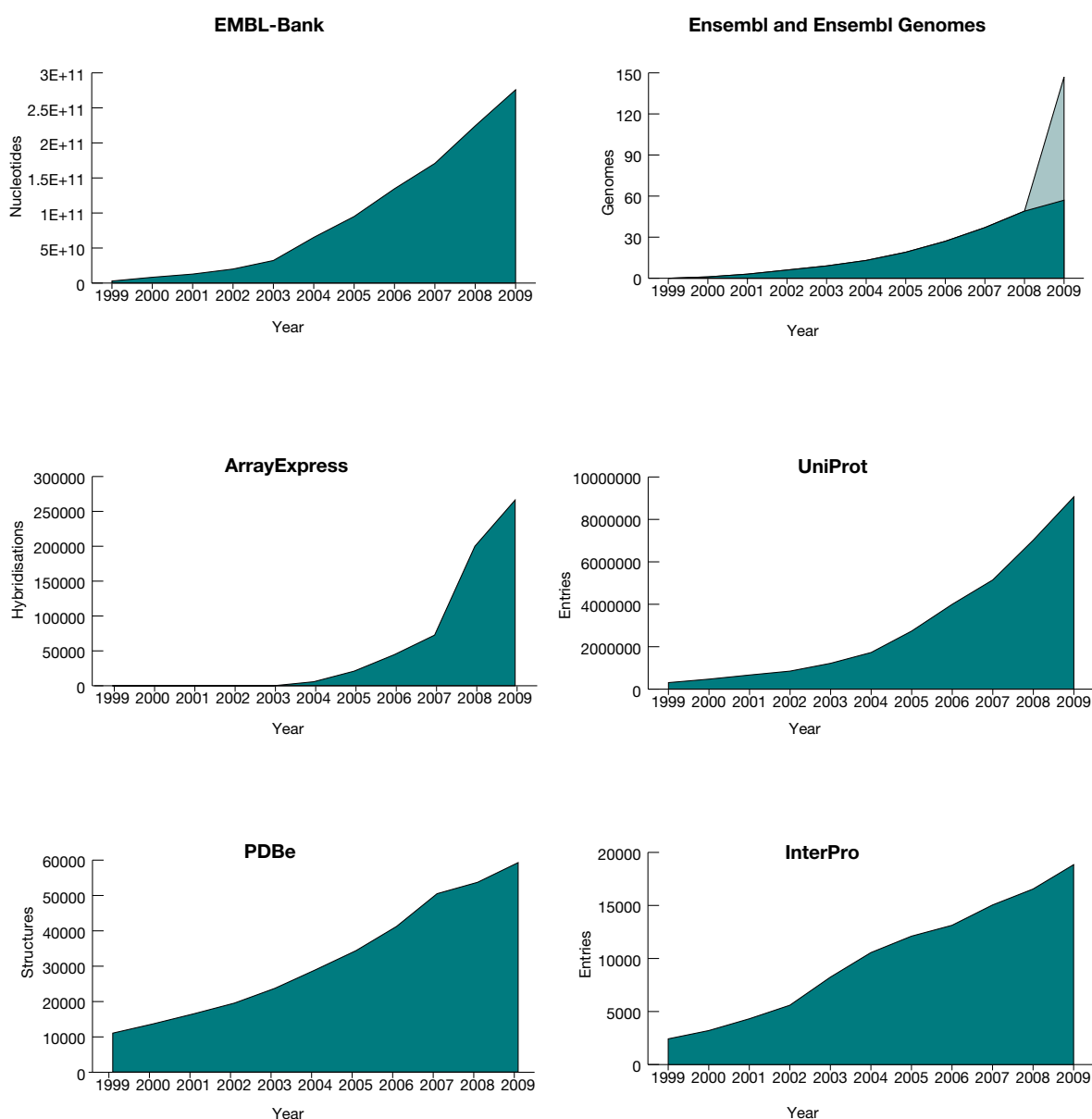


Figure 2. Growth of EMBL-EBI's core data resources, 1999–2009 (or from launch to 2009 if launch was after 1997).

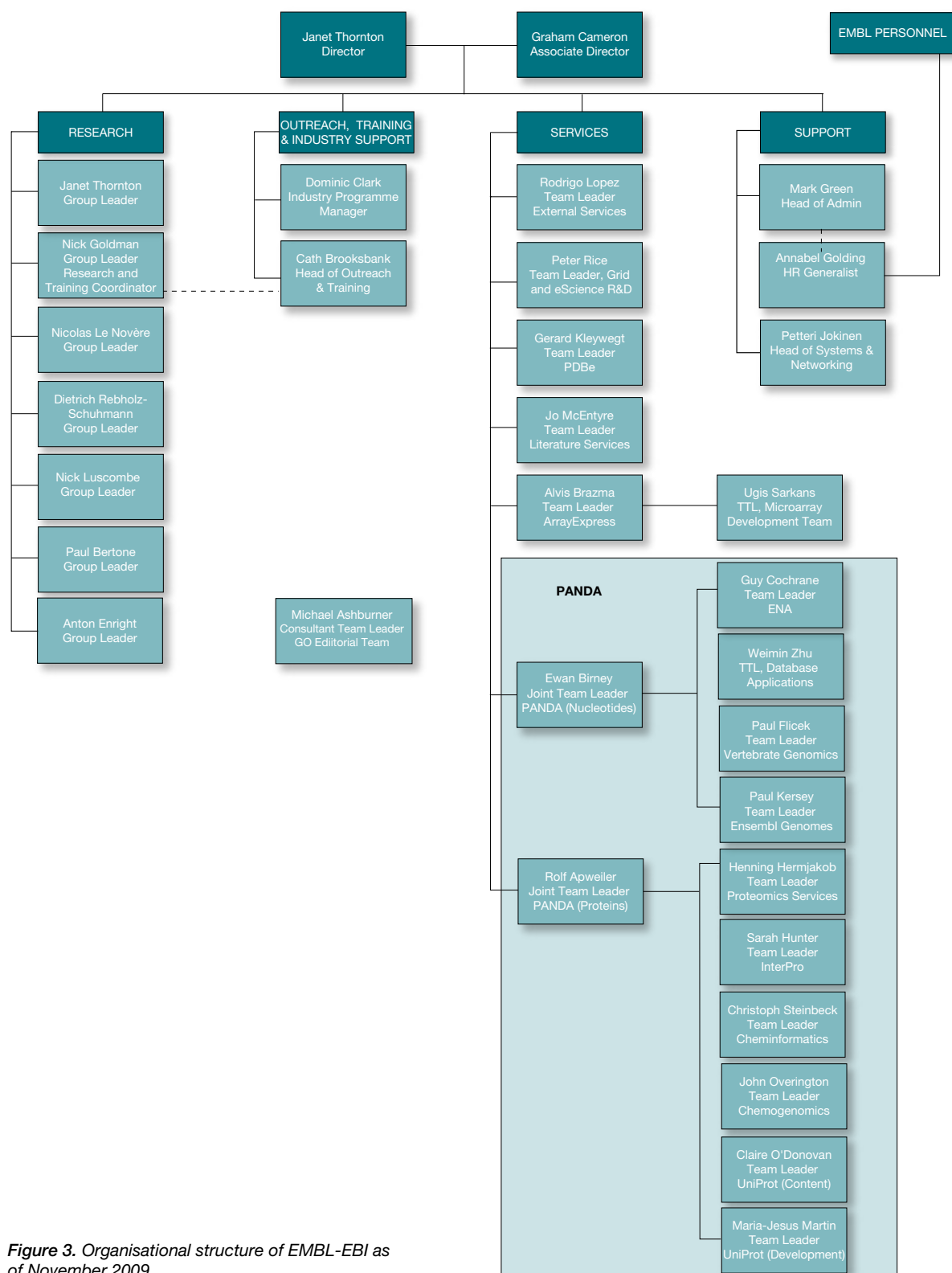


Figure 3. Organisational structure of EMBL-EBI as of November 2009.

STAFF

- Our organisational structure (figure 3) continues to reflect the four parts of our mission, with divisions for services, research, outreach and training, and support.
- The number of EBI members of personnel has grown by 3.5% (figure 4a) from 392 at the end of 2008 to 406 in October 2009 (these figures exclude visitors), and retains its cosmopolitan flavour: we currently have personnel (including long-term visitors) from 48 countries (compared with 47 in 2008; figure 4b). During 2009 we welcomed 47 long-term visitors (>1 month's visit; compared with 56 visitors in 2008).

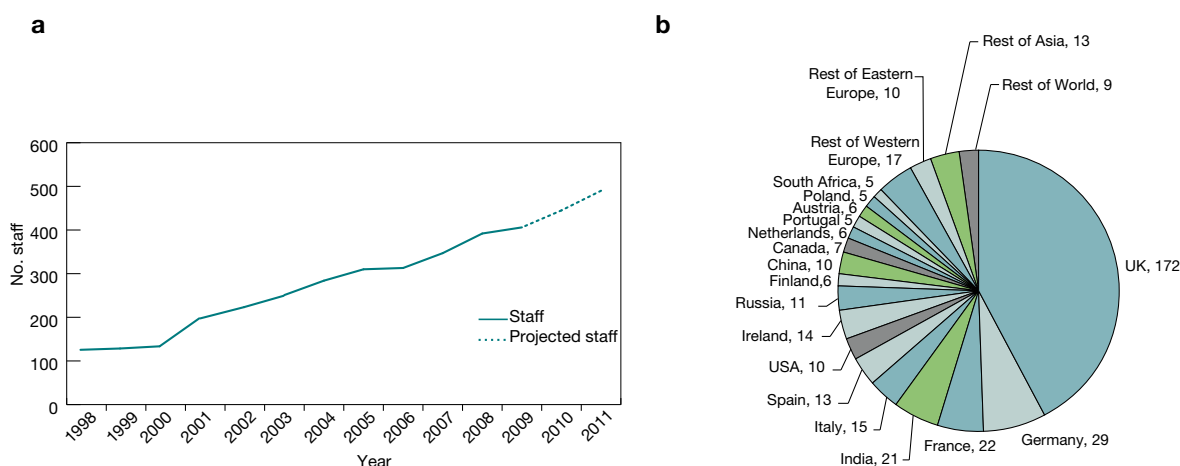


Figure 4. EMBL-EBI members of personnel. (a) Staff growth from 1998 to present, and projected staff growth. (b) Nationalities of EMBL-EBI members of personnel as of November 2009.

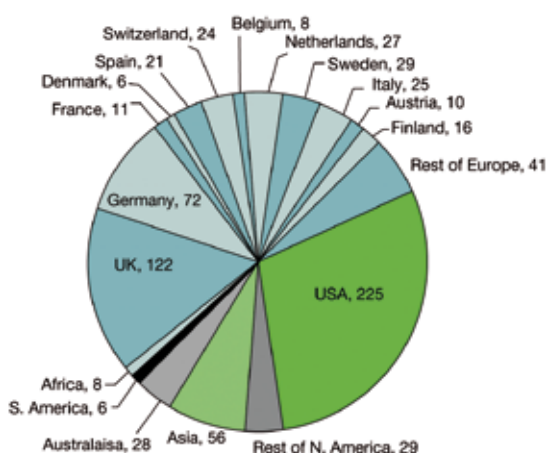
COLLABORATIONS

- Work at the EBI has continued to benefit from many collaborations (figures 5a and 5b) and almost all of our resources are funded through collaborative agreements. 80% of our publications during 2009 involved collaborations with external colleagues (compared with 89% in 2008) at 806 different institutes (808 in 2008).

FUNDING AND RESOURCE ALLOCATION

- We raised €19 million in external funding for 2009, compared with €23.8 million in 2008 (figure 6a). This excludes approximately €12 million (£10 million) from the UK research councils towards ELIXIR's compute infrastructure.
- Total internal funding to the EBI in 2009 was €20 million (€19.4 million in 2008), of which 47% (49% in 2008) was spent on salaries (figure 7). We have continued to invest in the EBI's core computing infrastructure: our spend on computing equipment was €4.2 million (compared with €5.2 million in 2008) – 21% of our total internal spend (figure 7a). We have increased our storage capacity from 2.5PB to 5PB and increased our compute power from 6,200 to more than 9,000 CPU cores during 2009.

a: publications



b: funding

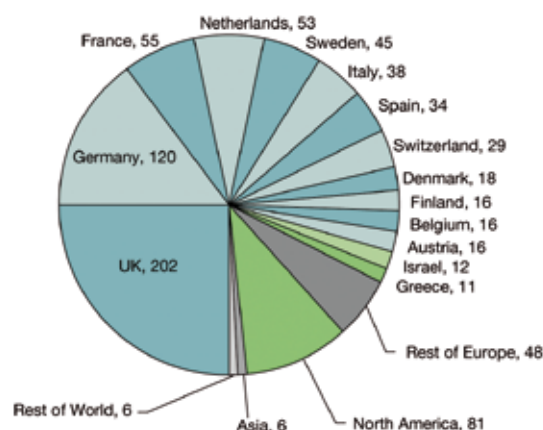


Figure 5. Collaborations as measured by (a) publications with other institutes and (b) funding shared with other institutions. Data for (a) were de-duplicated if the same institution appeared in the affiliations list of more than one paper. Data for (b) were not de-duplicated and in some cases the same institution is represented several times through different collaborations.

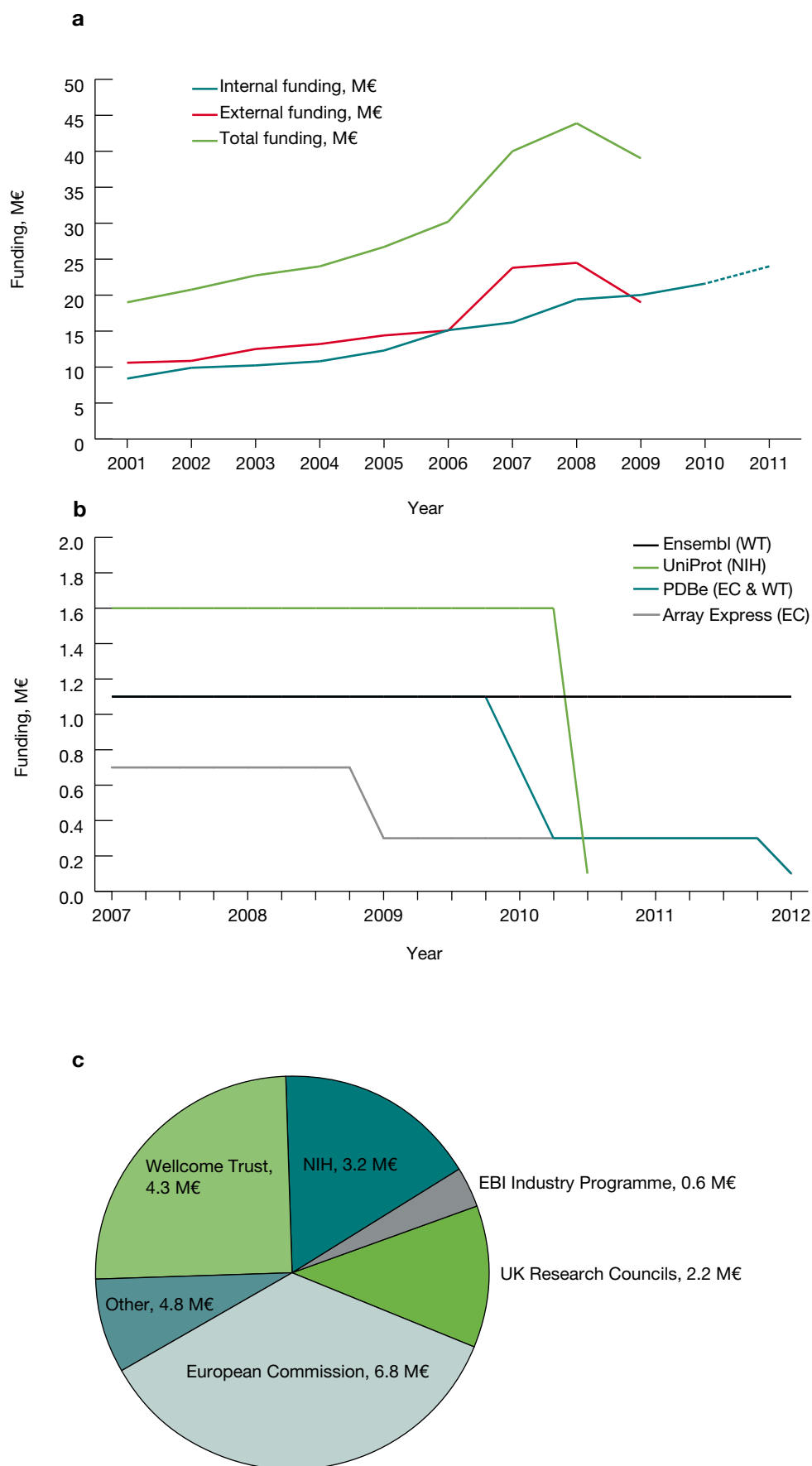
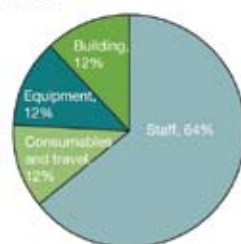


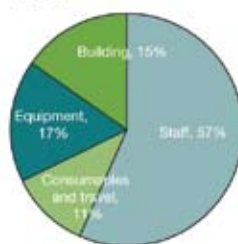
Figure 6. EMBL-EBI funding. (a) Growth of internal, external and total funds from 2001 to the present day, and agreed internal funds for 2009–2011. (b) Assured external funds for EMBL-EBI's core data resources from the present day to 2012. (c) Sources of external funding for the year (excluding funds raised towards ELIXIR) as of November 2009. The Wellcome Trust also supports us through provision of our buildings.

a: Internal spend

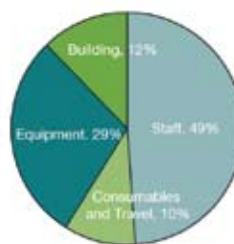
2006



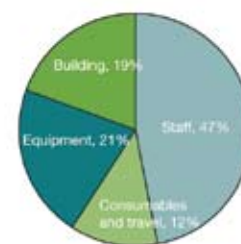
2007



2008

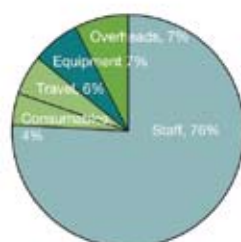


2009

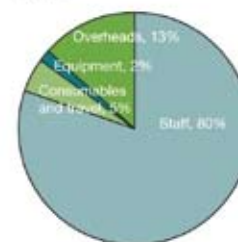


b: External spend

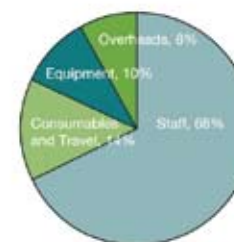
2006



2007



2008



2009

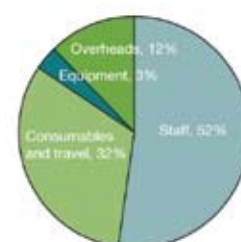


Figure 7. Breakdown of spend for 2005–2009. (a) Internal and (b) external spend.

Publications

EMBL-EBI publications 2008 (from September 2008)

- Aberg, K., *et al.* (2008). Support for schizophrenia susceptibility locus on chromosome 2q detected in a Swedish crosatellites and SNPs. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 147, 1238-1244
- Altman, R.B., *et al.* (2008). Text mining for biology - The way forward: Opinions from leading scientists. *Genome Biol.*, 9, S7.1-S7.15
- Andreini, C., *et al.* (2008). Metal ions in biological catalysis: From enzyme databases to general principles. *J. Biol. Inorg. Chem.*, 13, 1205-1218
- Beisswanger, E., *et al.* (2008). Gene Regulation Ontology (GRO): Design principles and use cases. *Stud. Health Technol. Inform.*, 136, 9-14
- Cerri, D., *et al.* (2008). Towards Knowledge in the Cloud. In 'OTM 2008 Workshops including SEMELS', Meersman, R., *et al.*, (eds)
- Chatr-Aryamontri, A., *et al.* (2008). MINT and IntAct contribute to the Second BioCreative challenge: Serving the text-mining community with high quality molecular interaction data. *Genome Biol.*, 9, Suppl 2, S5
- Coghlan, A., *et al.* (2008). nGASP - The nematode genome annotation assessment project. *BMC Bioinformatics*, 9, 549
- Cotton, R.G., *et al.* (2008). GENETICS. The Human Variome Project. *Science*, 322, 861-862
- Davis, A.M., *et al.* (2008). Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today*, 13, 831-841
- Eisenacher, M., *et al.* (2008). Proteomics data collection - 3rd ProDaC Workshop: April 22 nd 2008, Toledo, Spain. *Proteomics*, 8, 4163-4167
- Fernandez-Suarez, X.M., & Birney, E. (2008). Advanced genomic data mining. *PLoS Comput. Biol.*, 4, e1000121
- Goldman, N., & Yang, Z. (2008). Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. *Philos. Trans. R. Soc. B-Biol. Sci.*, 363, 3889-3892
- Guerlet, G., *et al.* (2008). Comparative models of P2X2 receptor support inter-subunit ATP-binding sites. *Biochem. Biophys. Res. Commun.*, 375, 405-409
- Helsens, K., *et al.* (2008). Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol. Cell. Proteomics*, 7, 2364-2372
- Herrgard, M.J., *et al.* (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, 26, 1155-1160
- Holland, R.C.G., *et al.* (2008). BioJava: An open-source framework for bioinformatics. *Bioinformatics*, 24, 2096-2097
- Jimenez, R.C., *et al.* (2008). Dasty2, an Ajax protein DAS client. *Bioinformatics*, 24, 2119-2121
- Jones, A.R. & Orchard, S. (2008). Minimum reporting guidelines for proteomics released by the proteomics standards initiative. *Mol. Cell. Proteomics*, 7, 2067-2068
- Kahn, R.A., *et al.* (2008). Consensus nomenclature for the human ArfGAP domain-containing proteins. *J. Cell Biol.*, 182, 1039-1044
- Kersey, P., *et al.* (2008). Building a biological space based on protein sequence similarities and biological ontologies. *Comb. Chem. High Throughput Screen.*, 11, 653-660
- Köhn, D. & Le Novère, N. (2008) SED-ML – An XML Format for the Implementation of the MIASE Guidelines CMSB 2008. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Heiner, M. & Uhrmacher, A.M. (eds), 5307, 176-190, Springer-Verlag
- Kuhn, S., *et al.* (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, 9, 400

- Le Novère, N. (2008). Multiscale modelling of neuronal signalling. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Heiner, M. & Uhrmacher, A.M. (eds), 5307, 176-190, Springer-Verlag
- Le Novère, N. (2008). Neurological disease: Are systems approaches the way forward? *Pharmacopsychiatry*, 41, S28-S31
- Loos, R. & Ogiwara, M. (2008). Time and Space Complexity for Splicing Systems. *Theory of Computing Systems*, 1-16
- Lovering, R.C., et al. (2008). Access to immunology through the gene ontology. *Immunology*, 125, 154-160
- Loytynoja, A. & Goldman, N. (2008). A model of evolution and structure for multiple sequence alignment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 3913-3919
- Maan, M.E., et al. (2008). Color polymorphism and predation in a Lake Victoria cichlid fish. *Copeia*, 621-629
- Madera, M. (2008). Profile Comparer: A program for scoring and aligning profile hidden Markov models. *Bioinformatics*, 24, 2630-2631
- O'Neill, K., et al. (2008). OntoDas – A tool for facilitating the construction of complex queries to the Gene Ontology. *BMC Bioinformatics*, 9, 437
- Orchard, S., et al. (2008). Annual Spring Meeting of the Proteomics Standards Initiative 23-25 April 2008, Toledo, Spain. *Proteomics*, 8, 4168-4172
- Pain, A., et al. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, 455, 799-803
- Paten, B., et al. (2008). Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, 18, 1814-1828
- Paten, B., et al. (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18, 1829-1843
- Rakyan, V.K., et al. (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, 18, 1518-1529
- Ramialison, M., et al. (2008). Rapid identification of PAX2/5/8 direct downstream targets in the otic vesicle by combinatorial use of bioinformatics tools. *Genome Biol.*, 9, r145
- Reeves, G.A., et al. (2008). The protein feature ontology: A tool for the unification of protein feature annotations. *Bioinformatics*, 24, 2767-2772
- Rustici, G., et al. (2008). Data storage and analysis in ArrayExpress and expression profiler. *Curr. Protoc. Bioinformatics*, unit 7.13, Suppl 23, 1-27
- Saini, H.K., et al. (2008). Annotation of mammalian primary microRNAs. *BMC Genomics*, 9, 564
- Sasaki, Y., et al. (2008). BioLexicon: A Lexical Resource for the Biology Domain. In *Third International Symposium on Semantic Mining in Biomedicine (SMBM)*
- Schmidt, A., et al. (2008). An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol Cell Proteomics*, 7, 2138-2150
- Seehausen, O., et al. (2008). Speciation through sensory drive in cichlid fish. *Nature*, 455, 620-626
- Smedley, D., et al. (2008). Solutions for data integration in functional genomics: A critical assessment and case study. *Brief. Bioinform.*, 9, 532-544
- Strombergsson, H., et al. (2008). Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *J. Chem. Inf. Model.*, 48, 2278-2288
- Taylor, M.S., et al. (2008). Rapidly evolving human promoter regions. *Nat. Genet.*, 40, 1262-1263
- UniProt Consortium (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, D190-D195
- van Dongen, S., et al. (2008). Detecting microRNA binding and siRNA off-target effects from expression data. *Nat. Methods*, 5, 1023-1025
- Vinken, M., et al. (2008). The carcinoGENOMICS project: Critical selection of model compounds for the development of omics-based in vitro carcinogenicity screening assays. *Mutat. Res./Rev. Mutat. Res.*, 659, 202-210
- Washietl, S., et al. (2008). Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.*, 24, 583-587
- Wilkins, M.R., et al. (2008). Information management for proteomics: A perspective. *Expert Rev. Proteomics*, 5, 663-678
- EMBL-EBI publications 2009 (to August 2009)**
- Adler, P., et al. (2009). Ranking genes by their co-expression to subsets of pathway members. *Ann. N. Y. Acad. Sci.*, 1158, 1-13
- Aligianni, S., et al. (2009). The fission yeast homeodomain protein Yox1p binds to MBF and confines MBF-dependent cell-cycle transcription to G1-S via negative feedback. *PLoS Genet.*, 5, 1-12
- Allen, M., et al. (2009). DNA methylation-histone modification relationships across the desmin locus in human primary cells. *BMC Mol. Biol.*, 10, 51
- Anders, S. (2009). Visualization of genomic data with the Hilbert curve. *Bioinformatics*, 25, 1231-1235
- Andreini, C., et al. (2009). Metal-MACiE: A database of metals involved in biological catalysis. *Bioinformatics*, 25, 2088-2089
- Andreini, C., et al. (2009). Structural Analysis of Metal Sites in Proteins: Non-heme Iron Sites as a Case Study. *J. Mol. Biol.*, 388, 356-380
- Apweiler, R., et al. (2009). Approaching clinical proteomics: Current state and future fields of application in fluid proteomics. *Clin. Chem. Lab. Med.*, 47, 724-744

- Apweiler, R., *et al.* (2009). The Universal Protein resource (UniProt) 2009. *Nucleic Acids Res.*, 37, D169-D174
- Baker, C.J.O. & Rebholz-Schuhmann, D. (2009). Between proteins and phenotypes: Annotation and interpretation of mutations. *BMC Bioinformatics*, 10, 11
- Barrell, D., *et al.* (2009). The GOA database in 2009 - An integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, 37, D396-D403
- Barsnes, H., *et al.* (2009). OMSSA Parser: An open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics*, 9, 3772-3774
- Barsnes, H., *et al.* (2009). PRIDE converter: Making proteomics data-sharing easy. *Nat. Biotechnol.*, 27, 598-599
- Bell, A.W., *et al.* (2009). A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods*, 6, 423-430
- Berka, K., *et al.* (2009). Representative amino acid side chain interactions in proteins. a comparison of highly accurate correlated ab initio quantum chemical and empirical potential procedures. *J. Chem. Theory Comput.*, 5, 982-992
- Berman, H., *et al.* (2009). The Worldwide Protein Data bank. In 'Structural Bioinformatics', Gu, J. & Bourne, P.E. (eds), 293-303, John Wiley & Son
- Berriman, M., *et al.* (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature*, 460, 352-358
- Bishop, C.J., *et al.* (2009). Assigning strains to bacterial species via the internet. *BMC Biol.*, 7, 3
- Blankenburg, H., *et al.* (2009). DASMI: Exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25, 1321-1328
- Blum, A., *et al.* (2009). 11 β -Hydroxysteroid dehydrogenase type 1 inhibitors with oleanan and ursan scaffolds. *Mol. Cell. Endocrinol.*, 301, 132-136
- Bonaglia, M.C., *et al.* (2009). Mosaic 22q13 deletions: Evidence for concurrent mosaic segmental isodisomy and gene conversion. *Eur. J. Hum. Genet.*, 17, 426-433
- Bourgon, R., *et al.* (2009). Array-based genotyping in *S.cerevisiae* using semi-supervised clustering. *Bioinformatics*, 25, 1056-1062
- Brazma, A. (2009). Minimum Information About a Microarray Experiment (MIAME)-successes, failures, challenges. *TheScientificWorldJournal*, 9, 420-423
- Brazma, A., *et al.* (2009). Message from the program chairs. *Proceedings of the 2009 9th IEEE International Conference on Bioinformatics and BioEngineering, BIBE 2009*
- Brazma, A., *et al.* (2009). Introduction. *J. Bioinform. Comput. Biol.*, 7, 5
- Bruce, A.W., *et al.* (2009). Functional diversity for REST (NRSF) is defined by in vivo binding affinity hierarchies at the DNA sequence level. *Genome Res.*, 19, 994-1005
- Caldas, J., *et al.* (2009). Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25, i145-i153
- Cao, H., *et al.* (2009). Comparative genomics indicates the mammalian CD33rSiglec locus evolved by an ancient large-scale inverse duplication and suggests all Siglecs share a common ancestral region. *Immunogenetics*, 61, 1-17
- Carbon, S., *et al.* (2009). AmiGO: Online Access to Ontology and Annotation Data. *Bioinformatics*, 25, 288-289
- Carlile, M., *et al.* (2009). Strand selective generation of endo-siRNAs from the Na/phosphate transporter gene *Slc34a1* in murine tissues. *Nucleic Acids Res.*, 37, 2274-2282
- Carver, T., *et al.* (2009). DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics*, 25, 119-120
- Chelliah, V., *et al.* (2009) Data Integration and Semantic Enrichment of Systems Biology Models and Simulations. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Paton, N.W., Missier, P. & Hedeler, C. (eds), 5647, 5-15, Springer-Verlag
- Chiang, T. & Scholtens, D. (2009). A general pipeline for quality and statistical assessment of protein interaction data using R and Bioconductor. *Nature Protocols*, 4, 535-546
- Cochrane, G., *et al.* (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37, D19-25
- Couto, F., *et al.* (2009). Verification of Uncurated Protein Annotations. In 'Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration', 311-325, IGI Global Publishing
- Cuff, A., *et al.* (2009). The CATH Hierarchy Revisited-Structural Divergence in Domain Superfamilies and the Continuity of Fold Space. *Structure*, 17, 1051-1062
- Cuff, A.I., *et al.* (2009). The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, 37, D310-D314
- de Jager, S.M., *et al.* (2009). Dissecting regulatory pathways of G1/S control in Arabidopsis: common and distinct targets of CYCD3;1, E2Fa and E2Fc. *Plant Mol. Biol.*, 71, 1-21
- Degtyarenko, K., *et al.* (2009). ChEBI: An open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics*, Suppl 26, unit 14.9, 1-20
- Drager, A., *et al.* (2009). SBML2LATEX: Conversion of SBML files into human-readable reports. *Bioinformatics*, 25, 1455-1456
- Durinck, S., *et al.* (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nature Protocols*, 4, 1184-1191
- Eisenacher, M., *et al.* (2009). Proteomics Data Collection -4th ProDaC workshop 15 August 2008, Amsterdam, the Netherlands. *Proteomics*, 9, 218-222

- Eisenacher, M., *et al.* (2009). Proteomics Data Collection - 5th ProDaC Workshop 4 March 2009, Kolympari, Crete, Greece. *Proteomics*, 9, 3626-3629
- Eisenacher, M., *et al.* (2009). Getting a grip on proteomics data - Proteomics Data Collection (ProDaC). *Proteomics*, 9, 3928-3933
- El-Hawari, Y., *et al.* (2009). Analysis of the substrate-binding site of human carbonyl reductases CBR1 and CBR3 by site-directed mutagenesis. *Chem. Biol. Interact.*, 178, 234-241
- Endler, L., *et al.* (2009). Designing and encoding models for synthetic biology. *J. R. Soc. Interface*, 6, S405-S417
- Feltrin, E., *et al.* (2009). Muscle research and gene ontology: New standards for improved data integration. *BMC Medical Genomics*, 2, 6
- Flieck, P. (2009). The need for speed. *Genome Biol.*, 10, 212
- Flieck, P. & Birney, E. (2009). Visualising the Epigenome, In 'Epigenomics', Ferguson-Smith, A.C., Grealley, J.M., & Martienssen, R.A. (eds), 55-66, Springer, Netherlands
- Fredman, D., *et al.* (2009). Web-based tools and approaches to study long-range gene regulation in Metazoa. *Brief Funct. Genomic. Proteomic*, 8, 231-242
- Furnham, N., *et al.* (2009). Missing in action: Enzyme functional annotations in biological databases. *Nat. Chem. Biol.*, 5, 521-525
- Gagneur, J., *et al.* (2009). Genome-wide allele- and strand-specific expression profiling. *Mol. Syst. Biol.*, 5, 1-9
- Gaudet, P., *et al.* (2009). The gene ontology's reference genome project: A unified framework for functional annotation across species. *PLoS Comput. Biol.*, 5,
- Gehlenborg, N., *et al.* (2009). Prequips - An extensible software platform for integration, visualization and analysis of LC-MS-MS proteomics data. *Bioinformatics*, 25, 682-683
- Gibson, A., *et al.* (2009). The data playground: An intuitive workflow specification environment. *Fut. Gen. Comp. Systems*, 25, 453-459
- Giglio, M.G., *et al.* (2009). Applying the Gene Ontology in microbial annotation. *Trends Microbiol.*, 17, 262-268
- Gnad, F., *et al.* (2009). MAPU 2.0: High-accuracy proteomes mapped to genomes. *Nucleic Acids Res.*, 37, D902-D906
- Haider, S., *et al.* (2009). BioMart central portal - Unified access to biological data. *Nucleic Acids Res.*, 37, W23-W27
- Haimel, M., *et al.* (2009). ProteinArchitect: Protein evolution above the sequence level. *PLoS ONE*, 4, e6176
- Harland, L. & Gaulton, A. (2009). Drug target central. *Expert Opin. Drug Discov.*, 4, 857-872
- Harmar, A.J., *et al.* (2009). IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.*, 37, D680-D685
- Harttig, U., *et al.* (2009). Owner controlled data exchange in nutrigenomic collaborations: the NuGO information network. *Genes Nutr.*, 1-10
- Holliday, G.L., *et al.* (2009). Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis. *J. Mol. Biol.*, 390, 560-577
- Huang, G.J., *et al.* (2009). High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res.*, 19, 1133-1140
- Hubbard, T.J.P., *et al.* (2009). Ensembl 2009. *Nucleic Acids Res.*, 37, D690-D697
- Hunter, S., *et al.* (2009). InterPro: The integrative protein signature database. *Nucleic Acids Res.*, 37, D211-D215
- Hwang, D., *et al.* (2009). A systems approach to prion disease. *Mol. Syst. Biol.*, 5, 1-23
- Ishimori, T., *et al.* (2009). Pipeline scheduling with input port constraints for an FPGA-based biochemical simulator. In 'Lecture Notes in Computer Science', 5453, 368-373
- Jain, E., *et al.* (2009). Infrastructure for the life sciences: Design and implementation of the UniProt website. *BMC Bioinformatics*, 10, 136
- Jimeno-Yepes, A., *et al.* (2009). Terminological cleansing for improved information retrieval based on ontological terms. In 'Proceedings of the WSDM'2009 ACM Workshop on Exploiting Semantic Annotations in Information Retrieval, ESAIR 2009', 6-14
- Johansson, S., *et al.* (2009). Probing natural killer cell education by Ly49 receptor expression analysis and computational modelling in single MHC class I mice. *PLoS ONE*, 4, e6046
- Kahraman A., *et al.* (2009). On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins*, [Epub ahead of print] PMID: 19927322
- Kaput, J., *et al.* (2009). Planning the human variome project: The Spain report. *Hum. Mutat.*, 30, 496-510
- Kathiresan, T., *et al.* (2009). A protein interaction network for the large conductance Ca²⁺-activated K⁺ channel in the mouse cochlea. *Mol. Cell. Proteomics*, 8, 1972-1987.
- Kauffmann, A., *et al.* (2009). ArrayQualityMetrics - A bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25, 415-416
- Kauffmann, A., *et al.* (2009). Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics*, 25, 2092-2094
- Kirmizis, A., *et al.* (2009). Distinct transcriptional outputs associated with mono- and dimethylated histone H3 arginine 2. *Nat. Struct. Mol. Biol.*, 16, 449-451
- Kleywegt, G.J. (2009). On vital aid: The why, what and how of validation. *Acta Crystallogr. Section D*, 65, 134-139
- Koscielny, G., *et al.* (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93, 213-220

- Krestyaninova, M., *et al.* (2009). A system for information management in BioMedical studies - SIMBioMS. *Bioinformatics*, 25, 2768-2769
- Krissinel, E. (2009). Crystal contacts as nature's docking solutions. *J. Comput. Chem.*, in press
- Kuhn, S., *et al.* (2009). Components for computer-assisted structure elucidation. *Chem. Cent. J.*, 3, 62
- Kuhn, T., *et al.* (2009). Creating chemo- and bioinformatics workflows, further developments within the CDK-Taverna Project. *Chem. Cent. J.*, 3, 42
- Laskowski, R. (2009). Integrated Servers for Structure-informed Function Prediction. In '*Protein Structure to Function with Bioinformatics*', Rigden, D.J. (ed), 251-272, Springer
- Laskowski, R. (2009). Protein Structure Databases. In '*Data Mining Techniques for the Life Sciences, Methods in Molecular Biology*', Carugo, O. & Eisenhaber, F. (eds), 609, Humana Press
- Laskowski, R. (2009). Structural Quality Assurance. In '*Structural Bioinformatics*', Gu, J. & Bourne, P.E. (eds), 341-375, John Wiley
- Laskowski, R.A. (2009). PDBsum new things. *Nucleic Acids Res.*, 37, D355-359
- Laskowski, R.A., *et al.* (2009). The structural basis of allosteric regulation in proteins. *FEBS Lett.*, 583, 1692-1698
- Laskowski, R.A., *et al.* (2009). The fine details of evolution. *Biochem. Soc. Trans.*, 37, 723-726
- Laughton, C.A., *et al.* (2009). COCO: A simple tool to enrich the representation of conformational variability in NMR structures. *Proteins: Structure, Function and Bioinformatics*, 75, 206-216
- Lawson, D., *et al.* (2009). VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res.*, 37, D583-587
- Le Novère, N., *et al.* (2009). The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27, 735-741
- Lefever, S., *et al.* (2009). RDML: Structured language and reporting guidelines for real-time quantitative PCR data. *Nucleic Acids Res.*, 37, 2065-2069
- Lewis, M.A., *et al.* (2009). An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat. Genet.*, 41, 614-618
- Loewenstein, Y., *et al.* (2009). Protein function annotation by homology-based inference. *Genome Biol.*, 10, 207
- Lovering, R.C., *et al.* (2009). Improvements to cardiovascular Gene Ontology. *Atherosclerosis*, 205, 9-14
- Loytynoja, A. & Goldman, N. (2009). Evolution. Uniting alignments and trees. *Science*, 324, 1528-1529
- Magalhaes, I.S., *et al.* (2009). Divergent selection and phenotypic plasticity during incipient speciation in Lake Victoria cichlid fish. *Journal of Evolutionary Biology*, 22, 260-274
- Martens, L. & Apweiler, R. (2009). Algorithms and databases. *Methods Mol. Biol.*, 564, 245-259
- Matthews, L., *et al.* (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37, D619-D622
- McWilliam, H., *et al.* (2009). Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res.*, 37, W6-W10
- Megy, K., *et al.* (2009). Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. *Infection, Genetics and Evolution*, 9, 308-313
- Mehta, A. & Orchard, S. (2009). Nucleoside diphosphate kinase (NDPK, NM23, AWD): recent regulatory advances in endocytosis, metastasis, psoriasis, insulin release, fetal erythroid lineage and heart failure; translational medicine exemplified. *Mol. Cell. Biochem.*, 329, 1-13
- Minh, B.Q., *et al.* (2009). Budgeted phylogenetic diversity on circular split systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 22-29
- Mishima, Y., *et al.* (2009). Zebrafish miR-1 and miR-133 shape muscle gene expression and regulate sarcomeric actin organization. *Genes Dev*, 23, 619-632
- Morley, R.H., *et al.* (2009). A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. *Proc. Natl Acad. Sci. USA*, 106, 3829-3834
- Motallebipour, M., *et al.* (2009). Novel genes in cell cycle control and lipid metabolism with dynamically regulated binding sites for sterol regulatory element-binding protein 1 and RNA polymerase II in HepG2 cells detected by chromatin immunoprecipitation with microarray detection. *FEBS J.*, 276, 1878-1890
- Nagel, K., *et al.* (2009). Annotation of protein residues based on a literature analysis: Cross-validation against UniProtKb. *BMC Bioinformatics*, 10, Suppl 8, S4
- Nobeli, I., *et al.* (2009). Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.*, 27, 157-167
- O'Connor, M.N., *et al.* (2009). Functional genomics in zebrafish permits rapid characterization of novel platelet membrane proteins. *Blood*, 113, 4754-4762
- Orchard, S. (2009). Ending the "publish and vanish" culture: How the data standardization process will assist in data harvesting. *J. Proteome Res.*, 8, 3219
- Orchard, S., *et al.* (2009). Second Joint HUPO Publication and Proteomics Standards Initiative Workshop. *Proteomics*, 9, 4426 - 4428
- Orchard, S. *et al.* (2009). Annual Spring Meeting of the Proteomics Standards Initiative. *Proteomics*, 9 4429 - 4432
- Orchard, S., *et al.* (2009). Managing the data explosion: A report on the HUPO-PSI workshop august 2008, Amsterdam, the Netherlands. *Proteomics*, 9, 499-501
- Orchard, S. & Ping, P. (2009). HUPO world congress publication committee meeting. *Proteomics*, 9, 502-503

- Orchard, S. & Taylor, C.F. (2009). Debunking minimum information myths: one hat need not fit all. *N. Biotechnol*, 25, 171-172
- Overington, J. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput. Aided Mol. Des.*, 23, 195-198
- Parkinson, H., *et al.* (2009). ArrayExpress update –from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, 37, D868-872
- Paten, B., *et al.* (2009). Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, 25, 295-301
- Pellegrini-Calace, M., *et al.* (2009). PoreWalker: A novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput. Biol.*, 5, e1000440
- Persson, B., *et al.* (2009). The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.*, 178, 94-98
- Pezik, P., *et al.* (2009). Using Biomedical Terminological Resources for Information Retrieval. In 'In Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration', IGI Global Publishing
- Prokopenko, I., *et al.* (2009). Variants in MTNR1B influence fasting glucose levels. *Nat. Genet.*, 41, 77-81
- Pruitt, K.D., *et al.* (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, 19, 1316-1323
- Rahman, S.A., *et al.* (2009). Small Molecule Subgraph Detector (SMSD) toolkit. *Journal of Cheminformatics*, in press
- Rayner, T.F., *et al.* (2009). MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. *Bioinformatics*, 25, 279-280
- Read, R.J. & Kleywegt, G.J. (2009). Case-controlled structure validation. *Acta Crystallogr. Section D*, 65, 140-147
- Reeves, G.A., *et al.* (2009). Genome and proteome annotation: Organization, interpretation and integration. *J. R. Soc. Interface*, 6, 129-147
- Robinson, J., *et al.* (2009). The IMGT/HLA database. *Nucleic Acids Res.*, 37, D1013-D1017
- Rodriguez, H., *et al.* (2009). Recommendations from the 2008 International Summit on proteomics data release and sharing policy: The Amsterdam principles. *J. Proteome Res.*, 8, 3689-3692
- San Mauro, D., *et al.* (2009). Experimental design in caecilian systematics: Phylogenetic information of mitochondrial genomes and nuclear rag1. *Syst. Biol.*, 58, 425-438
- Sarkar, D., *et al.* (2009). Quality Assessment and Data Analysis for microRNA Expression Arrays. *Nucleic Acids Res.*, 37, 8
- Schober, D., *et al.* (2009). Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*, 10, 125
- Seemple, C.A.M. & Taylor, M.S. (2009). The structure of change. *Science*, 323, 347-348
- Seshasayee, A.S.N., *et al.* (2009). Principles of transcriptional regulation and evolution of the metabolic system in E. coli. *Genome Res.*, 19, 79-91
- Sieglauff, D.H., *et al.* (2009). Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc. Natl Acad. Sci. USA*, 106, 3053-3058
- Smedley, D., *et al.* (2009). BioMart - Biological queries made easy. *BMC Genomics*, 10, article 22
- Stefan, M.I., *et al.* (2009). Computing phenomenologic Adair-Klotz constants from microscopic MWC parameters. *BMC Syst. Biol.*, 3, 68
- Steinbeck, C., *et al.* (2009). New open drug activity data at EBI. *Chem. Cent. J.*, 3, 62
- Sternberg, M., *et al.* (2009). Protein evolution - Sequence, structure and systems. *Biochemist*, 31, 52-52
- Stolovitzky, G., *et al.* (2009). Annals of the New York Academy of Sciences: Preface. *Ann. N. Y. Acad. Sci.*, 1158, 9-12
- Strombergsson, H. & Kleywegt, G. J. (2009). A chemogenomics view on protein-ligand spaces. *BMC Bioinformatics*, 10, Suppl 6, S13
- Suzuki, H., *et al.* (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, 41, 553-562
- Talavera, D., *et al.* (2009). WSsas: A web service for the annotation of functional residues through structural homologues. *Bioinformatics*, 25, 1192-1194
- Talavera, D., *et al.* (2009). The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, in press
- Tan, D.J.L., *et al.* (2009). Mapping organelle proteins and protein complexes in Drosophila melanogaster. *J. Proteome Res.*, 8, 2667-2678
- Thornton, J. (2009). Annotations for all by all - The BioSapiens network. *Genome Biol.*, 10, 401
- Trieschnigg, D., *et al.* (2009). MeSH Up: Effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25, 1412-1418
- Uniprot Consortium. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, 37, D169-D174
- Vaquerizas, J.M., *et al.* (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, 10, 252-263

- Vilella, A.J., *et al.* (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19, 327-335
- Vingron, M., *et al.* (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol.*, 10, 8
- Vizcaino, J.A., *et al.* (2009). A guide to the PRIDE proteomics data repository. *Proteomics*, 9 4276-4283
- Vizcaino, J.A., *et al.* (2009). Charting online OMICS resources: A navigational chart for clinical researchers. *Proteom. Clin. Appl.*, 3, 18-29
- Vranken, W.F. & Rieping, W. (2009). Relationship between chemical shift value and accessible surface area for all amino acid atoms. *BMC Struct. Biol.*, 9, 20
- Watkins, N.A., *et al.* (2009). A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, 113, e1-9
- Watson, J.D. & Thornton, J.M. (2009). Protein function prediction from structure in structural genomics and its contribution to the study of health and disease. In 'NATO Security through Science Series C: Environmental Security', Sussman, J. (ed), 201-215, Springer Verlag
- Wieser, D. & Niranjan, M. (2009). Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores. *In Silico Biol.*, 9, 89-103
- Wolkenhauer, O., *et al.* (2009). SysBioMed report: advancing systems biology for medical applications. *IET Syst Biol*, 3, 131-136
- Wortman, J.R., *et al.* (2009). The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort. *Fungal genetics and biology*, 46, Suppl 1, S2-S13
- Xu, Z., *et al.* (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457, 1033-1037
- Yan, W.H., *et al.* (2009). Systematic comparison of the human saliva and plasma proteomes. *Proteom. Clin. Appl.*, 3, 116-134
- Ye, K., *et al.* (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25, 2865-2871
- Zhenyu, J., *et al.* (2009). Association study between gene expression and multiple relevant phenotypes with cluster analysis. In 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Pizzuti, C., Richie, M.D. & Giacobini, M. (eds), 5483, p1-12, Springer-Verlag
- Theses submitted by EBI pre-doc students in 2009**
- Kahraman, A. (2009). The geometry and physicochemistry of protein binding. PhD Thesis, University of Cambridge, Cambridge, UK*
- Kerrison, N. Transcriptome analysis of longevity in mice. PhD Thesis, University of Cambridge, Cambridge, UK
- Lawler, K. Transcriptional and post-transcriptional regulation of gene expression: computational analysis of microarray studies in fungal species. PhD Thesis, University of Cambridge, Cambridge, UK
- Meynert, A. Function and evolution of regulatory elements in vertebrates. PhD Thesis, University of Cambridge, Cambridge, UK
- Mueller, M. (2009). Integrated analysis of proteomics data to assess and improve the scope of mass spectrometry based genome annotation. PhD Thesis, University of Cambridge, Cambridge, UK*
- Nagel, K. (2009). Automatic functional annotation of predicted active sites: combining PDB and literature mining. PhD Thesis, University of Cambridge, Cambridge, UK*
- Pardi, F. (2009). Algorithms on phylogenetic trees. PhD Thesis, University of Cambridge, Cambridge, UK*
- Seshasayee, A.S.N. (2009). A computational study of bacterial gene regulation and adaptation on a genomic scale. PhD Thesis, University of Cambridge, Cambridge, UK*
- Stefan, M. (2009). On the function of calcium-regulated allosteric devices in synaptic plasticity. PhD Thesis, University of Cambridge, Cambridge, UK*
- Tolle, D. Functional consequences of lateral and vertical movement of neurotransmitter receptors. PhD Thesis, University of Cambridge, Cambridge, UK
- Zerbino, D. (2009). Genome assembly and comparison. PhD Thesis, University of Cambridge, Cambridge, UK

* awarded

Major Database Collaborations

This list shows representative collaborations for our major databases; it is not intended to be comprehensive, but rather to give a flavour for the global impact of our work.

EMBL-Bank (the International Sequence Database Collaboration; www.insdc.org/)

ENA at EMBL-EBI, Hinxton, UK

GenBank, the Trace Archive and the Sequence Read Archive at the National Center for Biotechnology Information, Bethesda, MD, USA

DDBJ at the National Institute of Genetics, Mishima, Japan

Ensembl

Here we list collaborations with the major genome centres and representative collaborations for the human, mouse, rat and chicken genomes. There are many others.

Ensembl at EMBL-EBI and the Wellcome Trust Sanger Institute, Hinxton, UK

Genome Browser at the University of California, Santa Cruz, CA, USA

Map Viewer at the National Center for Biotechnology Information, Bethesda, MD, USA

Broad Institute, Cambridge, MA, USA

Baylor College of Medicine, Houston, Texas

DOE Joint Genome Institute, Walnut Creek, CA, USA

Mouse Genome Informatics at the Jackson Laboratory, Bar Harbor, ME, USA

Rat Genome Database at the Medical College of Wisconsin, Milwaukee, WI, USA

The Roslin Institute, Midlothian, Scotland, UK

Ensembl Genomes

Central Aspergillus Data Repository, Manchester, UK

Gramene at Cold Spring Harbor Laboratory, NY, USA

VectorBase: a collaboration of EMBL-EBI; University of Notre Dame, South Bend, IL, USA; Harvard University, MA, USA; Institute of Molecular Biology and Biochemistry, Greece; University of New Mexico, NM, USA; Imperial College, London, UK

WormBase at California Institute of Technology, Pasadena, CA, USA

ArrayExpress

ArrayExpress at EMBL-EBI, Hinxton, UK

The Microarray Gene Expression Data Society, Stanford University, CA, USA

Stanford Microarray Database, Stanford University, CA, USA

UniProt (the UniProt Consortium; www.uniprot.org)

UniProt at EMBL-EBI, Hinxton, UK

UniProt at the Swiss Institute of Bioinformatics, Geneva, Switzerland

UniProt at the Protein Information Resource, Georgetown University Medical Centre, Washington, DC, USA

UniProt at the Protein Information Resource, University of Delaware, Delaware, DE, USA

InterPro

InterPro and PDBe at EMBL-EBI, Hinxton, UK
 CATH-Gene3D at University College London, UK
 HAMAP at the Swiss Institute of Bioinformatics, Geneva, Switzerland
 PIRSF at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, USA
 Pfam at the Wellcome Trust Sanger Institute, Hinxton, UK
 PRINTS at the University of Manchester, UK
 ProDom at INRA and CNRS, Toulouse, France
 PROSITE at the Swiss Institute of Bioinformatics, Geneva, Switzerland
 SMART at EMBL, Heidelberg, Germany
 SUPERFAMILY at the University of Bristol, UK
 TIGRFAMs at The Institute of Genome Research, Rockville, MD, USA
 PANTHER at SRI, Menlo Park, CA, USA
 SCOP at the Laboratory of Molecular Biology, University of Cambridge, UK

PDBe (the worldwide Protein Databank; www.wwpdb.org/)

PDBe at EMBL-EBI, Hinxton, UK
 PDBj at Osaka University, Japan
 Research Collaborator for Structural Bioinformatics, USA
 BioMagResBank, University of Wisconsin, Madison, WI, USA

IntAct (The IMEx Consortium; imex.sourceforge.net)

IntAct at EMBL-EBI, Hinxton, UK
 DIP at the University of California, Los Angeles, CA, USA
 MINT at University Tor Vergata, Rome, Italy
 MIPS at the National Research Centre for Environment and Health, Munich, Germany
 Neuroproteomics platform of National Neurosciences Facility, Melbourne, Australia
 Shanghai Institutes for Biological Sciences, Shanghai, China
 Centro Nacional de Biotecnología, Madrid, Spain

PRIDE

PRIDE at EMBL-EBI, Hinxton, UK
 Ghent University, Ghent, Belgium
 Faculty of Life Sciences, The University of Manchester, UK
 The Yonsei Proteome Research Center, Yonsei University, Seoul, Korea.

Reactome

Reactome at EMBL-EBI, Hinxton, UK
 Reactome at Cold Spring Harbor Laboratory, NY, USA
 Ontario Institute for Cancer Research, Toronto, Ontario, Canada
 New York University Medical Center, NY, USA

Gene Ontology (the Gene Ontology Consortium; www.geneontology.org/GO.consortiumlist.shtml)

The GO Editorial Office, the Gene Ontology Annotation Project and Reactome at EMBL-EBI, Hinxton, UK
 FlyBase at the University of Cambridge, UK
 Berkeley Bioinformatics and Ontology Project, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
Saccharomyces Genome Database, Stanford University, Stanford, CA, USA
 Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA
 The *Arabidopsis* Information Resource, Carnegie Institution of Washington, Stanford, CA, USA
 WormBase at California Institute of Technology, Pasadena, CA, USA
 Rat Genome Database at the Medical College of Wisconsin, Milwaukee, WI, USA
 DictyBase at Northwestern University, Chicago, Israel, USA
 GeneDB *S. pombe* and GeneDB for protozoa at the Wellcome Trust Sanger Institute, Hinxton, UK
 Reactome at Cold Spring Harbor Laboratory, NY, USA
 The J. Craig Venter Institute, Rockville, MD, USA
 Gramene at Cornell University, Ithaca, NY, USA
 The Zebrafish Information Network at the University of Oregon, Eugene, OR, USA
 British Heart Foundation, University College London, London, UK
 EcoliWiki
 Institute for Genome Sciences, University of Maryland, Baltimore, MD, USA
 Agbase, Mississippi State University, Mississippi, MS, USA
 Candida Genome Database, Stanford University, Stanford, CA, USA
 Muscle TRAIT, University of Padua, Padua, Italy
 Plant-Association Microbe Gene Ontology, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Scientific Advisory Boards

EMBL-EBI Bioinformatics Advisory Committee

Anna Tramontano, University of Rome 'La Sapienza', Rome, Italy (chair)
 Rob Cooke, GlaxoSmithKline
 Roderic Guigó, Centre de Regulació Genòmica, Barcelona, Spain
 Olli Kallioniemi, VTT Medical Biotechnology, Turku, Finland
 John Sulston, Wellcome Trust Sanger Institute, Hinxton, UK
 Mathias Uhlén, Royal Institute of Technology (KTH), Stockholm, Sweden
 Martin Vingron, Max Planck Institute for Molecular Genetics, Berlin, Germany

ArrayExpress Scientific Advisory Board

Frank Holstege, University Medical Center Utrecht, The Netherlands (chair)
 Catherine A. Ball, Stanford Microarray Database, Stanford University, USA
 Richard Durbin, Wellcome Trust Sanger Institute, Hinxton, UK
 Roderic Guigó, Centre de Regulació Genòmica, Barcelona, Spain
 Edwin Southern, University of Oxford, UK
 Christian J. Stoeckert, University of Pennsylvania, USA
 Martin Vingron, Max Planck Institute for Molecular Genetics, Berlin, Germany

BioModels Database Scientific Advisory Board

Upinder S. Bhalla, National Centre for Biological Sciences, India
 Michael Hucka, California Institute of Technology, USA
 Pedro Mendes, Manchester Centre of Integrative Systems Biology, UK
 Ion Moraru, University of Connecticut Health Center, USA
 Herbert Sauro, Washington University, USA
 Jacky L. Snoep, Stellenbosch University, South Africa

Ensembl Scientific Advisory Board

Detlev Arndt, European Molecular Biology Laboratory, Heidelberg, Germany
 Michael Ashburner, University of Cambridge, UK
 Stephan Beck, University College London, London, UK
 Allan Bradley, Wellcome Trust Sanger Institute, Hinxton, UK
 Søren Brunak, Technical University of Denmark, Lyngby, Denmark
 Michele Clamp, Broad Institute, USA
 Michael Eisen, University of California and Howard Hughes Medical Institute, Berkeley, USA
 Jim Kent, University of California Santa Cruz, USA
 Mark McCarthy, University of Oxford, UK
 Chris Ponting, University of Oxford, UK
 Nick Walton, University of Cambridge, UK

Ensembl Genomes Scientific Advisory Board

Detlev Arendt, European Molecular Biology Laboratory, Heidelberg, Germany
 Mike Bevan, John Innes Centre, Norwich, UK

Michele Clamp, Broad Institute, USA
Steve Oliver, University of Cambridge, UK
Julian Parkhill, Wellcome Trust Sanger Institute, UK
Doreen Ware, Cold Spring Harbour Laboratory, USA

European Nucleotide Archive Advisors

Antoine Danchin, CNRS, Institut Pasteur, Paris, France
Babis Savakis, University of Crete and IMBB-FORTH, Heraklion, Greece
Jean Weissenbach, Génoscope, Evry, France

Gene Ontology Scientific Advisory Board

David Botstein, Lewis-Sigler Institute, Princeton University, USA
Philip Bourne, University of California San Diego, USA
Lawrence Hunter, University of Colorado Health Sciences Center, Aurora, USA
Richard Scheuermann, University of Texas Southwestern Medical Center, Dallas, USA
Michael Schroeder, Technische Universität Dresden, Germany
Barry Smith, SUNY Buffalo, USA
Simon Tavaré, University of Southern California, Los Angeles, USA and University of Cambridge, UK
Michael Tyers, University of Edinburgh, UK

InterPro Scientific Advisory Board (joint with Pfam)

Philip Bourne, University of California San Diego, USA
Michael Galperin, National Center for Biotechnology Information, Bethesda, USA
Erik Sonnhammer, Stockholm University, Sweden
Alfonso Valencia, Spanish National Cancer Research Centre, Madrid, Spain

Reactome Scientific Advisory Board

Julie Ahringer, University of Cambridge, UK
Russ Altman, Stanford University, USA
Gary Bader, University of Toronto, Canada
Richard Belew, University of California San Diego, USA
Matt Day, Nature Publishing Group
Edda Klipp, Max-Planck Institute for Molecular Genetics, Germany
Adrian Krainer, Cold Spring Harbor Laboratory, USA
Ed Marcotte, University of Texas at Austin, USA
Mark McCarthy, Oxford University, UK
William Pearson, University of Virginia, Charlottesville, USA
Pardis Sabeti, Broad Institute, USA
David Stewart, Cold Spring Harbor Laboratory, USA

UniProt Scientific Advisory Board

Michael Ashburner, University of Cambridge, UK (chair)
Helen Berman, Rutgers University, USA
Judith Blake, The Jackson Laboratory, USA
Takashi Gojobori, National Institute of Genetics, Tokyo, Japan
Young-Ki Paik, Yonsei University, Seoul, Korea
Manuel Peitsch, Novartis Institutes for BioMedical Research, Cambridge, USA
David Searls, University of Pennsylvania, USA
Gunnar von Heijne, Stockholm University, Sweden

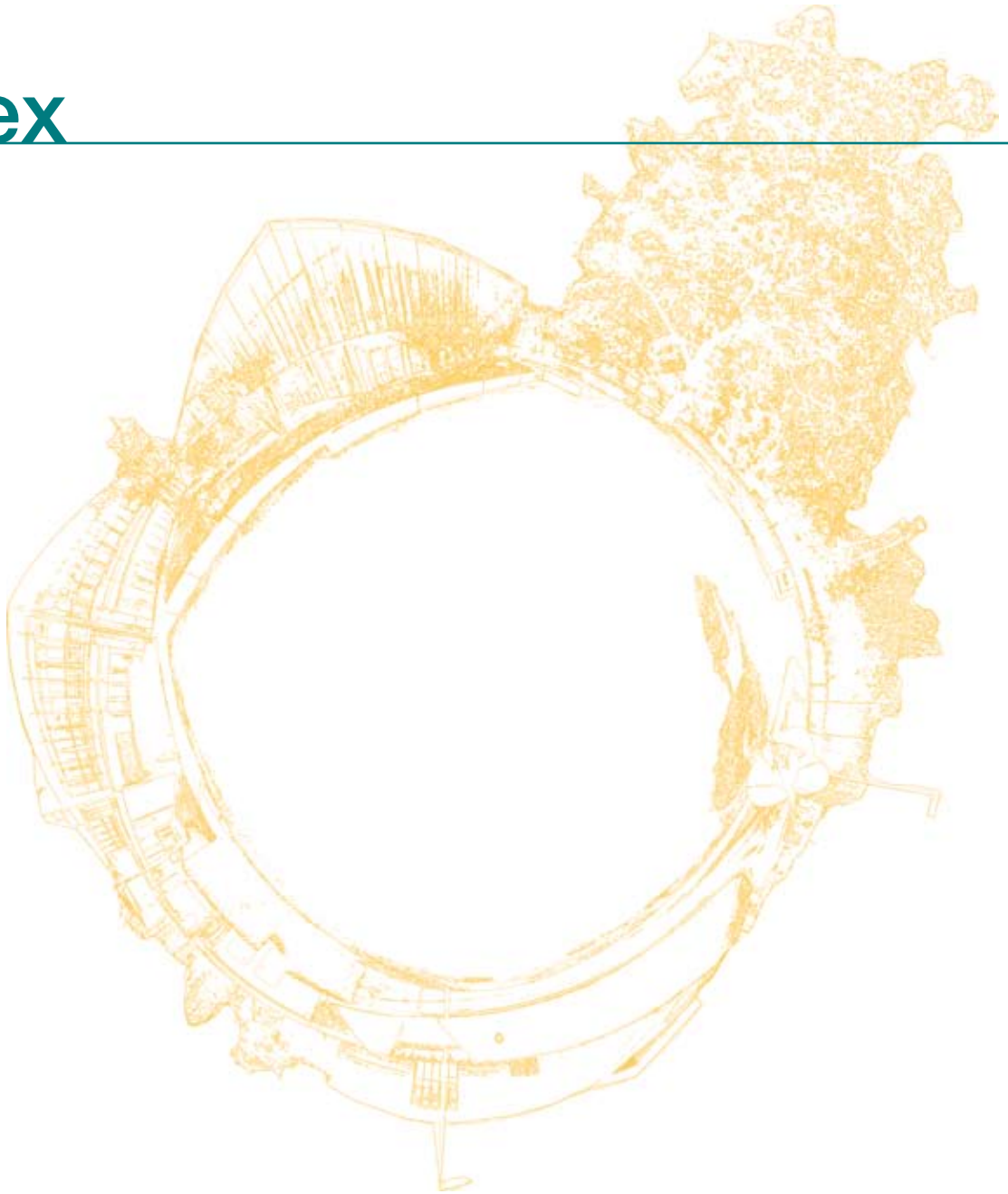
wwPDB Scientific Advisory Committee

Andreas Engel, University of Basel, Switzerland
Udo Heinemann, Max Delbrück House (Flachbau), Berlin, Germany
Ernest Laue, University of Cambridge, UK
Tomas Lundqvist, AstraZeneca R&D, Mölndal, Sweden
Andrea Mattevi, University of Pavia, Italy
Randy J. Read, University of Cambridge, UK
Helen Saibil, Birkbeck College London, UK
Michael Sattler, TUM, Munich, Germany
Torsten Schwede, Swiss Institute of Bioinformatics, University of Basel, Switzerland
Titia Sixma, Netherlands Cancer Institute, Amsterdam, The Netherlands
Keith S. Wilson, University of York, UK

External Seminar Speakers

Date	Speaker	Title
1 September 2008	Matthias Heinemann	Phenotypic bistability in <i>Escherichia coli</i> 's central carbon metabolism
15 September 2008	Melanie Stefan	An allosteric model of calmodulin explains differential activation of PP2B and CaMKII
22 September 2008	Anna Panchenko	Protein interactions and binding sites: diversity, specificity and evolution
6 October 2008	John O'Brien	Counting change carefully: robust distances among molecular sequences through Markov-induced counting processes
20 October 2008	Robert Hoffmann	Collaborative publishing and information management in the life sciences
27 October 2008	Amy Schmid	Boosting predictive accuracy in regulatory network models using novel biological information
3 November 2008	Mauno Vihinen	Pathogenic or not? And if so, then how? Analysis of the effects of mis-sense mutations by bioinformatics methods
24 November 2008	Ian Grieve	Genome-wide co-expression analysis in recombinant inbred rat strains
1 December 2008	Matthieu Louis	Sensory logic of odor perception in <i>Drosophila</i> larvae
15 December 2008	Christoph Best	Molecules in the mist: bioimage informatics for cryo-electron microscopy
18 February 2009	Jaime Prilusky and Eran Hodis	Proteopedia - a scientific 'wiki' for the intuitive communication of 3D structure and function of biomacromolecules
2 March 2009	Rune Linding	Molecular logic gates and network medicine
6 April 2009	Thomas Rattei	SIMAP - structuring the network of protein similarities
5 May 2009	Sven Nelander	Drug combinations, gene combinations, and cancer
12 May 2009	Jasmin Fisher	The executable pathway to biological networks
26 May 2009	Frank Bruggeman	The erratic nature of single cells
2 June 2009	Ulrich Stelzl	Constructing directed protein interaction networks for activated EGF/Erk signaling
16 June 2009	Diego di Bernardo	Inference and modeling of gene networks and applications to genetic diseases
23 June 2009	Martin Vingron	Transcription factor affinity prediction: methods, statistics, and delineation of tissue specific binding sites
14 July 2009	Felix Naef	Studying circadian clocks using comparative genomics and modelling
28 July 2009	David Thybert	Detecting functional potentialities in prokaryotic genomes: Application to the ROS/RNS detoxification sub-system
31 July 2009	Sheldon McKay	Challenges in comparative genome browsing

Index



A

Apweiler, Rolf 17

B

Bertone, Paul 105
 BioSapiens 14
 Birney, Ewan 17
 Brazma, Alvis 79
 Brooksbank, Cath 149

C

Cameron, Graham 7, 9
 ChEMBL Team 61
 Chemoinformatics and Metabolism 65
 Clark, Dominic 159
 Cochrane, Guy 31
 Computational biology of proteins 141
 Computational chemical biology 61
 Computational systems neurobiology 123
 Contents 3

D

Database Research and Development 71
 Developing and integrating tools for biologists 93

E

ELIXIR 7, 11–14, 150–151, 156–157
 Enright, Anton 111
 Ensembl Genomes Team 45
 European Nucleotide Archive Team 31
 Evolutionary tools for sequence analysis 117
 External Seminar Speakers 189
 External Services Team 165

F

Facts and Figures 171
 Flicek, Paul 37
 Foreword 7
 Functional genomics and analysis of small RNA function 111

G

Genome-scale analysis of regulatory systems 129
 GO Editorial Office 75
 Goldman, Nick 117, 149

H

Harris, Midori 75
 Henrick, Kim 89
 Hermjakob, Henning 51
 Highlights of 2009 9
 Hunter, Sarah 57

I

Industry Support 159
 InterPro Team 57
 Introduction 5

J

Jokinen, Petteri 163

K

Kersey, Paul 45

Kleywegt, Gerard 89

L

Le Novère, Nicolas 123

Literature Resource Development 97

Lopez, Rodrigo 165

Luscombe, Nicholas 129

M

Major Database Collaborations 185

McEntyre, Johanna 97

Microarray Informatics Team 79

Microarray Software Development Team 85

O

Outreach and Training 149

Overington, John 61, 159

P

PANDA Group 17

Protein Data Bank in Europe (PDBe) Team 89

Proteomics Services Team 51

Publications 177

R

Rebholz-Schuhmann, Dietrich 135

Research 103

Rice, Peter 93

S

Sarkans, Ugis 85

Scientific Advisory Boards 187

Semantic standardisation of the scientific literature 135

Services 15, 169

Steinbeck, Christoph 65

Stoehr, Peter 97

Support 147

Systems and Networking 163

T

Thornton, Janet 7, 9, 141

V

Vertebrate Genomics 37

Z

Zhu, Weimin 71

EMBL member states:

Austria, Belgium, Croatia, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom. Associate member state: Australia.

EMBL-EBI is a part of the European Molecular Biology Laboratory (EMBL)