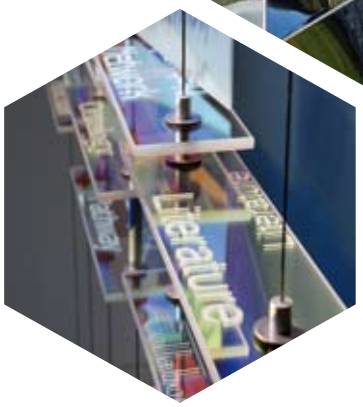


European Bioinformatics Institute

Annual Scientific Report 2008

EMBL-EBI

Annual Scientific Report 2008

European Bioinformatics Institute

EMBL-EBI

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 89902 0, Fax +49 (0)40 89902 104

EMBL Monterotondo
Adriano Buzzati-Traverso Campus
Via Ramarini, 32
00015 Monterotondo (Rome)
Italy
Tel. +39 06 90091285, Fax +39 06 90091272

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 Scientific Outreach Officer

Contents

Section 1

Introduction

Foreword

Welcome to EMBL-EBI's 2008 Annual Scientific Report.

Biological research has broken new ground in 2008: enabled by impressive advances in DNA sequencing technology, life scientists are now generating petabytes of data on a daily basis. EMBL-EBI's mission, to provide bioinformatics services, research, training and industrial support, has therefore never been more important.

New DNA sequencing methods will drive a second revolution in biology, with impacts not just in basic biological research, but for personalised medicine and for measuring the biodiversity of the planet. With the start of the 1000 Genomes Project, the fundamental nature of human variation will be revealed. To address this, the EBI launched the European Genotype Archive (EGA) during 2008 to store individual genomes; the EBI has also become the custodian of the associated Trace Archive, which holds the raw sequence data. Both will play a major role in understanding human variation.

The maturing fields of systems and chemical biology have highlighted the importance of the 'small' molecules of life and their interactions, which also relate closely to the design of novel therapeutics and new approaches to crop management. To this end, we are developing a suite of open source chemistry resources. As well as building on ChEBI, a database of small biomolecules, we are creating ChEMBL, which will hold information on structure-activity relationships.

Research at EBI is diverse and flourishing, producing both exciting discoveries and the development of powerful new tools to handle the flood of data. Highlights this year include a new phylogenetic-aware method for multiple sequence alignments, sophisticated tools to handle tiling array and image data, and new discoveries in the control of gene expression and cellular division. We have also recruited a new research group leader, to champion the increasingly important RNA field.

Our services are being used ever more frequently by scientists in Europe and worldwide. We provide simple web access for at least 300,000 independent users every month, but increasingly scientists are using programmatic access to large amounts of data through web services technologies. Web services now account for close to a million jobs per month. The computational and storage needs of all these projects are tremendous, and this year has seen an enormous growth in both the computer power and the storage needed to fulfil our commitments.

As the data resources develop, our commitment to providing training for users increases. With a growing team of trainers, we have run many new workshops and courses this year, both at Hinxton and throughout Europe. The main focus of our training is to empower biologists to make the most of their data, but we also hold more technical workshops for computational specialists. Our Industry Programme has also grown, invigorated by the new computational chemistry developments at the EBI and bringing new ideas both for services and workshops.

With increased funding from EMBL this year, we have been able to consolidate part of Europe's core set of data resources, but longer-term funding is still precarious. To address this, the preparatory phase of the ELIXIR project is in full swing. The project aims to develop a plan 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. This infrastructure will increasingly become the life-blood of life science research in Europe, allowing scientists to combine information from many sources to understand and model complex biological systems and their interactions with the environment. It is notable that all six of the other European Research Infrastructure projects in the biomedical sciences are looking to ELIXIR to support and interact with their own informatics efforts.

All our efforts at the EBI rely on extensive interaction 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.

Janet Thornton, Director
Graham Cameron, Associate Director

EMBL-EBI in 2008

SERVICES

Things seldom slow down for the service projects of the EBI, and 2008 was no exception. New technology in the lab has caused a huge surge in the rate of DNA sequencing, and has again revolutionised genomics and transcriptomics (See Facts and Figures, Figure 1 on page 13). The European Nucleotide Archive has accepted 10Tb of next-generation sequence data, and ArrayExpress now takes data from ultra high-throughput sequencing (UHTS) experiments.

This has challenged our compute and storage systems, with our total disk space reaching 2.5 petabytes in the course of the year. This has been accompanied by a steady increase in the need for compute power, and by the end of the year the EBI's computer facility included close to 7,000 cores. To accommodate these increases within the available space and without saturating the power supply to the campus we have had to include more compact 'blade' systems. There have also been significant changes in the 'flavour' of our technology, with Linux-based database servers finding favour, and Apple Macs increasingly being the desktop machine of choice.

The sheer volume of data stored creates substantial anxiety about back-up and disaster recovery systems, and the EBI is addressing this through a recently signed agreement to utilise a remote 'replication' site ten kilometres from the campus. Naturally, all of this imposes new demands on our network capacity, and we have now ordered a 10GB/second connection to London. Geography has worked in our favour and influenced our choice of location for our remote data replication centre, which sits on the route of this new network connection.

Aside from dealing with the sheer quantity of data, our nucleotide sequence efforts have been consolidated and refreshed with the imminent launch of Ensembl Genomes as a natural home for the DNA data from all organisms that have completed 'reference genomes'. This will greatly enhance the utility of those data by exploiting the software of the Ensembl system.

On the protein sequence side, a major development has been the provision of a first draft of the complete human proteome available in UniProtKB/Swiss-Prot. The InterPro resource, which is an integrated resource for protein families and domains, has grown to over 16,500 entries, providing annotations for almost 80% of proteins in the UniProt Knowledgebase. More than a third of these contain additional sequences representing isoforms generated by alternative splicing, alternative promoter usage and/or alternative translation initiation, resulting in close to 34,000 human protein sequences. Approximately 46,000 single amino acid polymorphisms (SAPs), mostly disease-linked, are also described, as well as 60,000 post-translational modifications (PTMs).

For protein structures held at the EBI as part of the Worldwide Protein Data Bank (wwPDB), the Macromolecular Structure Database team has integrated the experimental data derived by 3D cryo-electron microscopy and electron tomography techniques, and derived the molecular biological assemblies of structures. Increasingly, in the era of systems biology, knowledge about the properties of molecules is not enough. We must understand their behaviour to understand biological processes. In 2008 the IntAct database of molecular interactions has grown to cover 170,000 interactions.

The increasing relevance of the EBI's biomolecular data to the world of medicine has caused us to substantially invest in our provision of 'small' molecule data. The ChEBI (Chemical Entities of Biological Interest) collection has been enhanced by adding information on drug names and cross-references to patent information. The Wellcome Trust has awarded £4.7 million (€5.8 million) to the EBI 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 publicly listed company Galapagos NV to the public domain. This has allowed us to create a new chemogenomics team, which will develop interfaces that provide open access to the data and provide the critical link between the small molecules and their targets, which is essential for translating biological knowledge into new therapies.

All of these data, however, deal with molecules and processes rather than organisms. Perhaps the most interesting challenges of the year stem from the ability to study genome variation and its phenotypic consequences. Armed with robust human genome reference data and cheap, rapid sequencing technologies, we are at last able to look in detail at the genomes of single individuals. To cope with this we have launched the European Genotype Archive, which will be a major step towards the realisation of the medical promises of genome sequencing. This development requires the EBI, which

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.

Referenced 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

Löytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320, 1632-1635

Mancera, E., *et al*. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454, 479-485

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

has never before dealt with data identifiable to individuals, to tackle a new range of ethical and confidentiality issues and implement secure systems to support them.

The 1000 Genomes Project, an ambitious global effort to sequence the genomes of at least 1,000 people to create the most detailed and medically useful catalogue to date of human genetic variation, was launched. The EBI is making the data swiftly available to the worldwide scientific community through freely available public databases. To help achieve this, new funding from the Wellcome Trust has enabled the EBI to take guardianship of Europe's archive of raw DNA sequence data. The Trace Archive is one of the world's largest databases of biological information, and was originally created by the Wellcome Trust Sanger Institute. EMBL-EBI will maintain this important archive of DNA sequence in collaboration with long-term collaborator the US National Institute of Biotechnology Information (NCBI). The challenge of dealing with sequence data on the scale required by the 1000 Genomes Project is enormous: in a single weekend earlier this year the project deposited more data in the Trace Archive than the EBI had previously stored. The EBI and NCBI are collecting and analysing sequence generated by the other consortium partners, which include the Wellcome Trust Sanger Institute, the Beijing Genomics Institute, Shenzhen, China, the USA's National Human Genome Research Institute Large-Scale Sequencing Network and the Max Planck Institute for Molecular Genetics in Berlin.

A never-ending challenge arises in supporting the exploitation of our different information sources as a combined resource. For example, the EB-eye search engine allows users to traverse links between the databases, utilising nearly 400 million cross-references in the databases that we serve. Some of the most important of these are connections to the scientific literature – there are about 70 million citations in our databases – and indeed our literature services have been brought into the fold of the EBI's core activities. In 2008, enhancements to the CiteXplore literature resource included the ability to explore 'citing' and 'cited by' articles for some 60 million papers, and the provision of a web services applications programming interface to its capabilities. The EBI's role as a partner in the UK PubMed Central project took a step forward with the award of a Wellcome Trust grant to the EBI which will, among other things, develop the links between the scientific literature and the biomolecular databases.

As ever, the demand for these services burgeons, with web hits to the EBI site now averaging about 3.5 million per day. Particularly pleasing developments include the increasing uptake of our web services applications programming interfaces. We have long sought protocols to allow rich exploitation of our data by our users' own software. At almost a million compute jobs per month, we see signs of success.

RESEARCH

Research at the EBI is undertaken not only by dedicated research groups, but also by several service teams that incorporate a small research and development component. In recent years bioinformatics research has diversified and matured from basic sequence and structure analysis to include all aspects of biology, encompassing regulation, evolution, variation, systems modelling and even new methods to extract information from the literature. The eight specialist research groups complement 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. Many exciting new discoveries and tools have been developed this year and below are a few highlights:

Nick Goldman's group developed a new, phylogenetically aware method of sequence comparison (Löytynoja & Goldman, 2008). Unlike previously existing sequence comparison tools, the new method can invoke information about the history of insertions and deletions. The results of sequence comparisons using the new tool challenge our understanding of how evolution happens and suggest that sequence turnover is much more common than was previously assumed.

Wolfgang Huber's group, in collaboration with the group of Lars Steinmetz at EMBL Heidelberg, contributed to the most precise map of genetic recombination ever (Mancera *et al*., 2008). Genetic recombination, the process by which sexually reproducing organisms shuffle their genetic material when producing germ cells, leads to offspring with a new genetic make-up and influences the course of evolution. The study sheds light on fundamental questions about genetic shuffling and has implications for the tracking of disease genes and their inheritance.

The groups of Nick Luscombe and Paul Bertone at EMBL-EBI, collaborating with researchers from

the lab of Asifa Akhtar at EMBL Heidelberg, have uncovered the mechanism by which sex chromosomes are regulated in fruit flies to ensure that females don't produce twice as much protein from their two X chromosomes as males produce from their lone X chromosome (Kind *et al.*, 2008). They have discovered that a transcriptional regulator called MOF binds differently to male and female X chromosomes, enhancing the successful transcription of genes from male X chromosomes.

In a systems biology approach, Nicolas Le Novère's group has developed a new allosteric model for proteins in the postsynaptic density. This includes a complex model for calcium binding to calmodulin, revealing more about the complex molecular processes in the brain (Stefan *et al.*, 2008).

As part of their research, several groups have developed specialist databases, which are made available through the EBI website. These include the BioModels database of mathematical models of biological interest and the MACiE database of enzyme mechanisms. In addition, new methods to help researchers extract the knowledge hidden in the literature are being developed by the Rebholz-Schuhmann group, and will ultimately feed into the EBI's service provision.

NEW GROUP AND TEAM LEADER APPOINTMENTS

As the demands on the EBI have grown throughout 2008, we have been fortunate to recruit excellent scientists to lead newly established teams, supported by both EMBL and by external funds.

Anton Enright joined the EBI as a research group leader from the Wellcome Trust Sanger Institute. Anton's research focuses on determining the functions of regulatory RNAs and he is also interested in the analysis of biological networks, protein–protein interactions, clustering algorithms and vitalisation techniques.

Paul Flicek was appointed from within the EBI to lead the vertebrate genomics team. Paul leads the parts of the Ensembl project dealing with variation, inter-genome comparisons and functional genomics, as well as developing resources for human genetic variation.

Paul Kersey was appointed from within the EBI to lead the Ensembl genomes team, which is broadening the taxonomic coverage of Ensembl to include five taxonomic domains – metazoa, protists, bacteria, plants and fungi.

Gerard Klejwegt, currently at Uppsala University, has been appointed as the new team leader for the Macromolecular Structure Database team. Gerard will join the EBI formally in June 2009. In the interim he will be working closely with current team leader Kim Henrick to ensure a smooth transition and secure future funding for the team's activities.

John Overington joined the EBI from BioFocus DPI to lead the new chemogenomics team, which will establish a world-class data resource for drug discovery at the EBI, providing for the first time open access to target–ligand structure-activity data, to facilitate translational research.

Christoph Steinbeck joined the EBI from the University of Cologne to lead the Chemoinformatics and Metabolism teams. These conduct research and create community resources for chemoinformatics and metabolism research, including ChEBI and IntEnz. Christoph also contributes to the Chemistry Development Kit, an open source library of tools for structural chemo- and bioinformatics.

TRAINING

Training is a priority for the EBI and we have continued to invest heavily in providing training for the EBI's users. During 2008 EBI personnel have contributed to more than 330 training and training-related events around the world, reaching an estimated 30,000 scientists. We continue to make the most of synergies with the Wellcome Trust Course and Conference Programme; we ran our first joint EBI-Wellcome Trust course (proteomics bioinformatics) in November 2007. This was repeated in 2008 and we have more joint courses and conferences scheduled for 2009 and 2010.

GLOBAL CONTEXT

The EBI only operates because of multiple collaborations with scientists in Europe and worldwide (Figure 1). International exchange of data, such as occurs with the nucleotide sequences and protein structures, allows worldwide access and sharing of all our data. It is the best way to avoid duplication of effort and to share the cost of development and curation, whilst ensuring open access and stability. Progress towards sharing data in some of the newer resources is ongoing. European networks of excellence, often coordinated by the EMBL-EBI, also provide crucial links to research scientists throughout Europe.

The strengthening of collaborations on campus is of particular note this year; the Hinxton IT group is working on developing joint solutions to the IT challenges posed by the flood of data, including replication, thereby safeguarding our biological data against disasters. The Hinxton Sequencing Forum meets regularly to discuss campus-wide matters relating to sequencing data; and the campus course and conference group meets to exploit synergies in our scientific events. We now also hold monthly joint scientific seminars with the Wellcome Trust Sanger Institute.

Our links to industry through the EBI's Industry Programme provide valuable guidance for developing our resources and also for identifying upcoming areas of interest. In turn, our industry partners benefit from training, advice, interactions and exposure to new developments at the EBI. This year the Industry Programme has grown, and several projects of strategic importance to industry are underway, ranging in focus from the 'Druggable Genome' to plant pathogens.

FUNDING

In 2008 EMBL-EBI was privileged to receive an increase in our EMBL funding as part of the current five-year indicative scheme (See Figure 6a, Facts and Figures, page 17). These funds have largely been used to support key staff responsible for the core databases and to consolidate these databases, thereby improving their long-term stability and decreasing their reliance on external episodic grant funding. In addition we are in contract negotiations with the European Commission for SLING, an integrating activity that will enable us, alongside the Swiss Institute of Bioinformatics, the BRENDA Database and the European Patent Office, to make significant improvements to Europe's core data resources over the coming three years and to provide training for users throughout the European Union.

Despite this progress, the long-term funding for the data resources remains precarious (Figure 6b, page 17). The rapid increase in data, combined with new data types and the increasing importance of these data for medicine and agriculture, means that we must work with scientists and funding bodies throughout Europe to build a model for a sustainable infrastructure for biological information in the future. The preparatory phase of ELIXIR, an EU-funded project to agree upon the future bioinformatics infrastructure for Europe, is now well underway. We hope that we will be able to draft a memorandum of understanding in the coming year for agreement between EU member states, which will pave the way towards a more stable footing for Europe's biological data resources in the future.

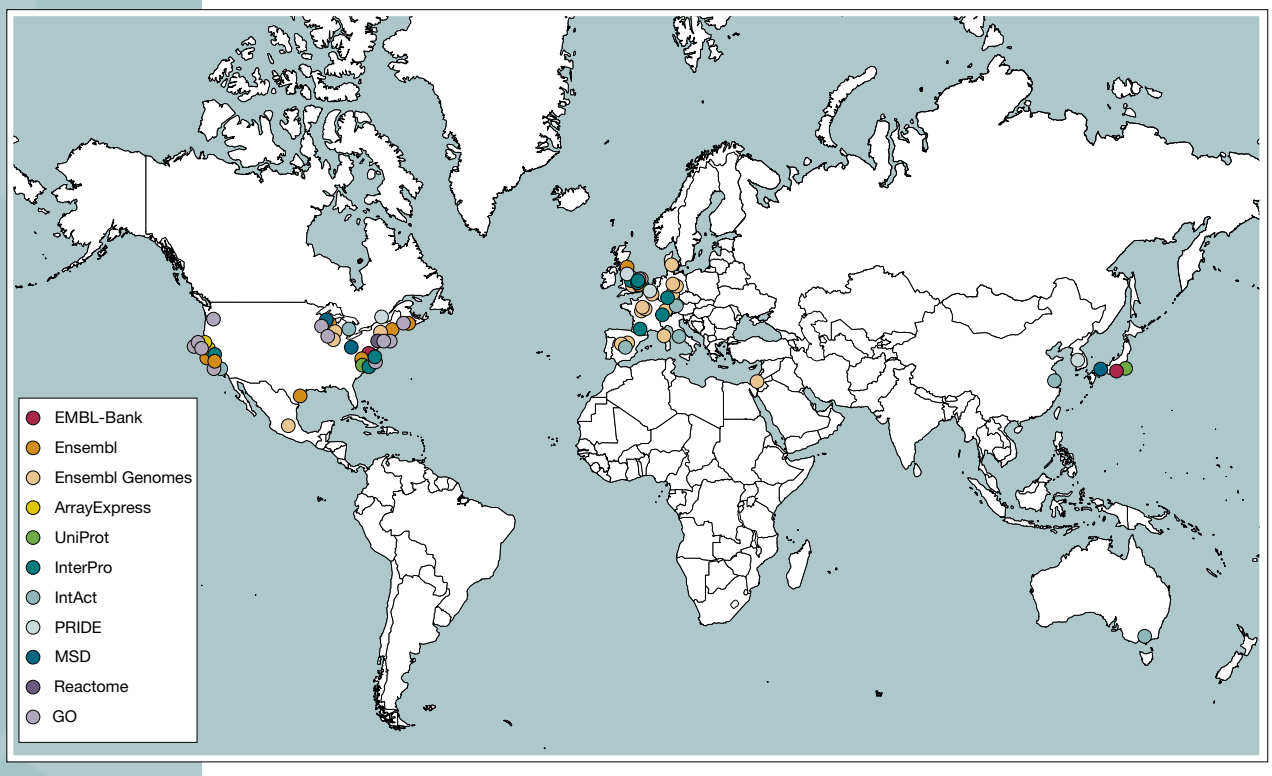


Figure 1. *Map showing locations of EMBL-EBI's collaborators for the major databases.*

Facts and Figures

SERVICES

- The EBI continues to host the major core biomolecular resources of Europe – collecting, archiving and distributing data throughout Europe and beyond. From September 2007 to August 2008 all our core data resources grew significantly (Figure 1). We have received and processed more than 1.8 x 10^9 bases compared with 1.56 x 10^9 bases in 2007. A newly established archive for next-generation sequencing data brings the total number of nucleotides we are responsible for to 1.17 x 10^{12} bases. We have processed 2.1 million UniParc entries (1.1 million in 2007), 115,000 microarray hybridisations (28,000 in 2007), 5,649 macromolecular structures (9,263 in 2007) and 12 new eukaryotic genomes in Ensembl (10 in 2007).
- The services continued to be well used during 2008 (see report on External Services, page 37). By August 2008 there were on average 2,709,000 requests per day (2,329,000 in 2007) – 3,241,000 if Ensembl is included (Figure 2).

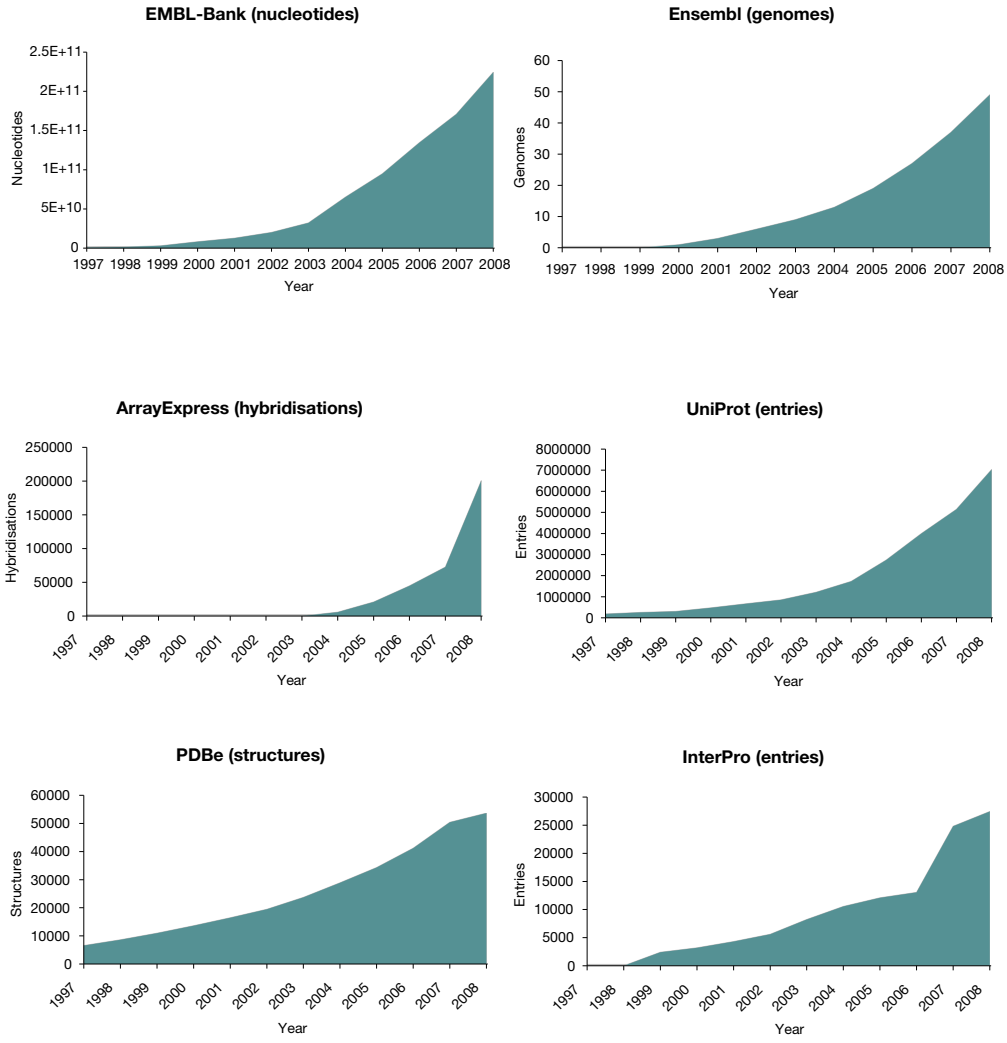


Figure 1. *Growth of EMBL-EBI's core data resources, 1997–2008 (or from launch to 2008 if launch was after 1997).*

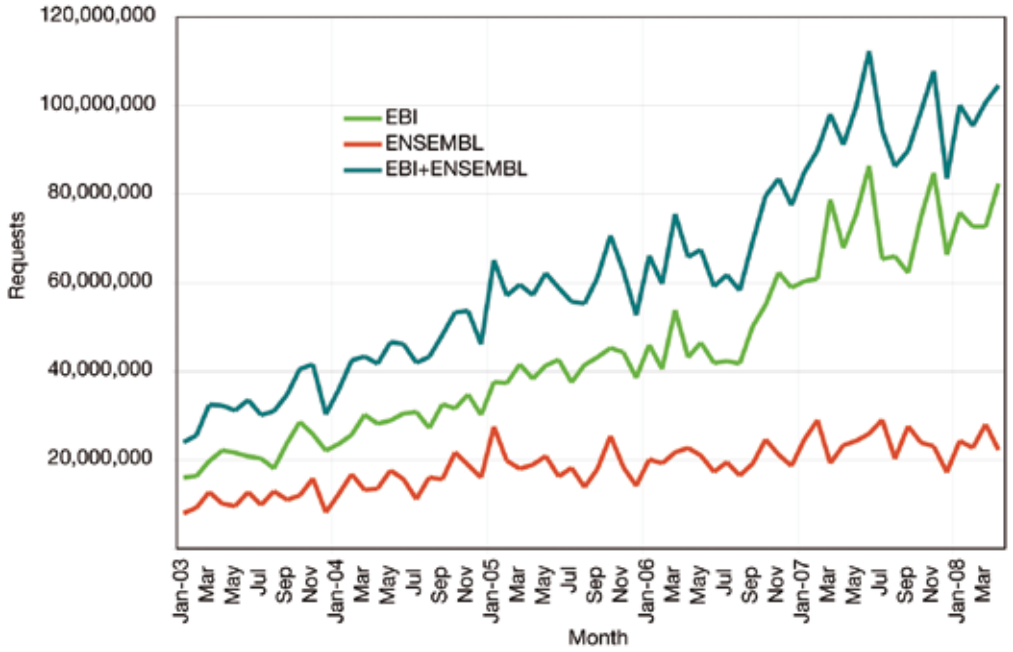


Figure 2. *A plot of the web requests received by the EBI and Ensembl from January 2003 to March 2008.*

RESEARCH

- The EBI published 232 papers between September 2007 and August 2008 (compared with 156 in 2007), 78 from research groups (53 in 2007).
- The research group leaders have successfully applied for support for their research, totalling €2 million over the next 2–5 years and including new funding from the UK research councils.
- The number of pre-doctoral students at EBI is currently 35 (39 in 2007). This includes eight PhD students who started their studies in October 2008.
- For the first time, we were successful in our application for BBSRC quota studentships; the result of this is that we will be able to host one new BBSRC-funded pre-doctoral student per year from 2010 to 2013.

OUTREACH AND TRAINING

- We have taken part in 330 training events throughout 2008 (compared with 146 in 2007), reaching over 30,000 participants.
- In its first year, the EBI's Hands-on User Training Programme organised eleven courses and trained 220 researchers, an average of 20 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 FELICS Integrated Infrastructure Initiative, is in high demand. We have run 21 roadshows during 2008, taking the total to 34 roadshows in 14 different countries since the programme first began.

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 13% (Figure 4a) from 347 at the end of 2007 to 392 in October 2008 (these figures exclude visitors), and retains its cosmopolitan flavour: we currently have personnel (including long-term visitors) from 47 countries (compared with 46 in 2007; Figure 4b). During 2008 we welcomed 56 long-term visitors (>1 month's visit; compared with 11 visitors in 2007), including 16 Marie-Curie students (16 in 2007).

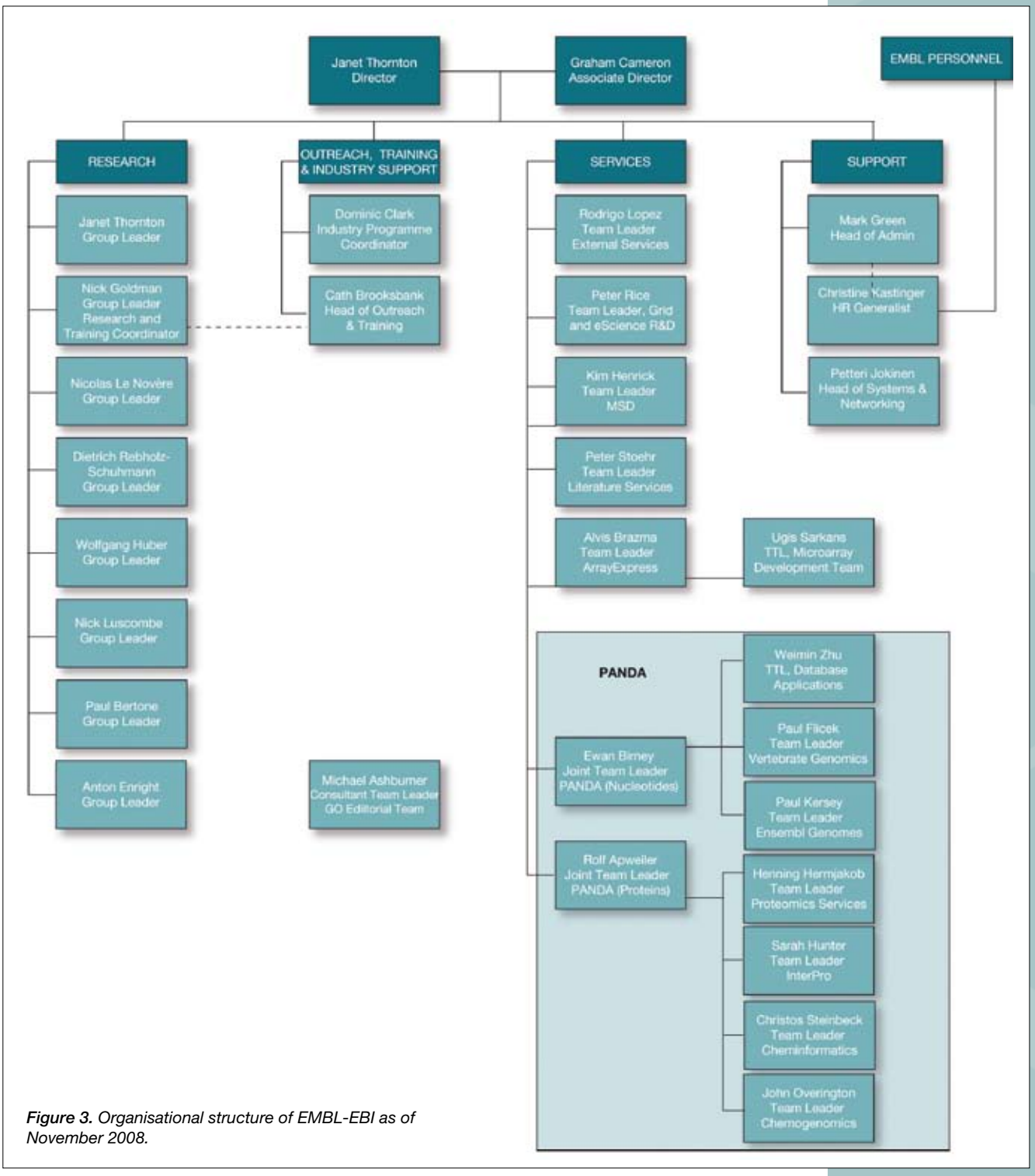


Figure 3. *Organisational structure of EMBL-EBI as of November 2008.*

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. 89% of our publications during 2008 involved collaborations with external colleagues (compared with 61% in 2007) at 808 different institutes (440 in 2007).

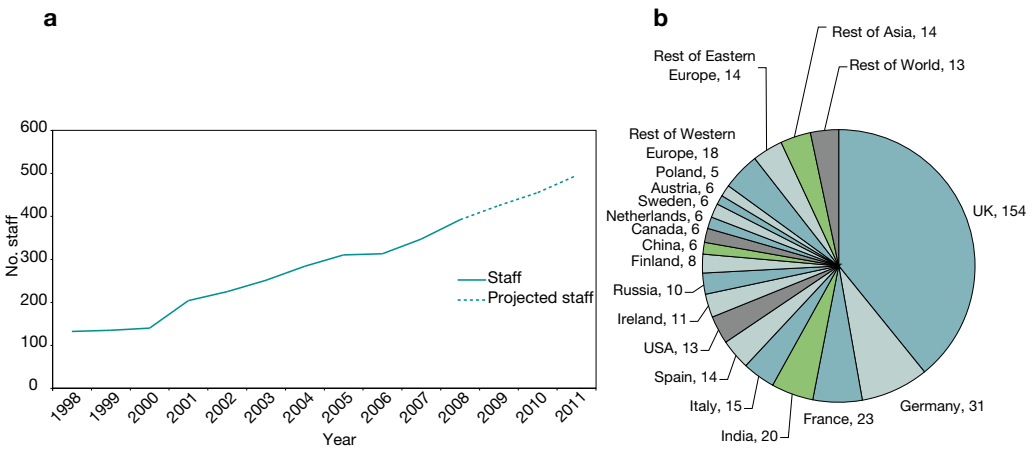


Figure 4. *EMBL-EBI members of personnel. (a) Staff growth from 1998 to present, and projected staff growth based on internal funding provided by the 2007–2011 Indicative Scheme. (b) Nationalities of EMBL-EBI members of personnel as of November 2008.*

FUNDING AND RESOURCE ALLOCATION

- We raised a total of €43.2 million in funding for 2008, compared with €31.25 million for 2007 (Figure 6a). €23.8 million of this came from external sources (Figure 6a, c).
- Total internal funding to the EBI in 2008 was €19.4 million (€16.2 million in 2007), of which 49% (57% in 2007) was spent on salaries (Figure 7). We have continued to invest in the EBI's core computing infrastructure: our spend on computing equipment was €5.2 million (compared with €3 million in 2007) – 29% of our total internal spend (Figure 7a). We have increased our storage capacity from 500 TB to 2.5 PB (a five-fold increase in storage) and increased our compute power from 3,600 to 6,200 CPU cores during 2008.

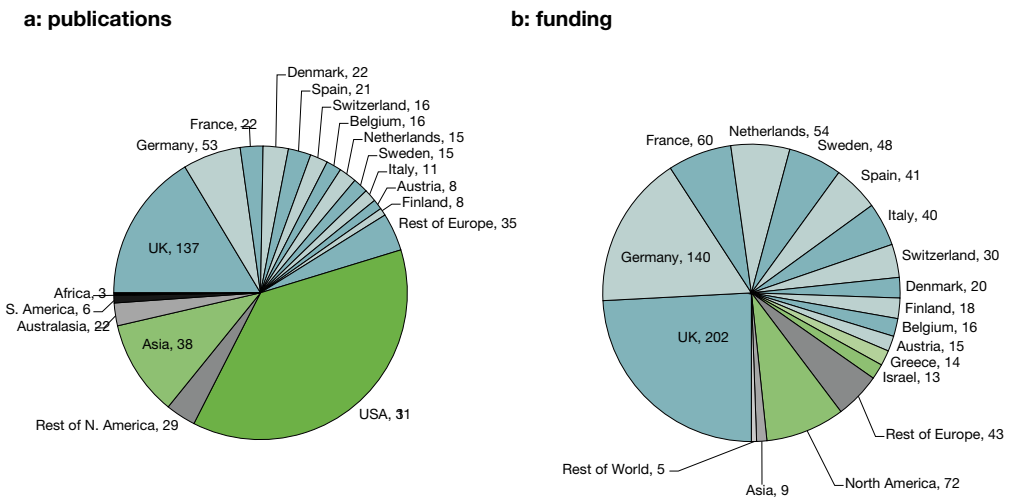


Figure 5. *Collaborations as measured by (a) publications with other institutes and (b) funding shared with other institutions. Data for (a) were de-duplicated if the same institution appeared in the affliliations list of more than one paper. Data for (b) were not de-duplicated and in some cases the same institution is represented several times through different collaborations.*

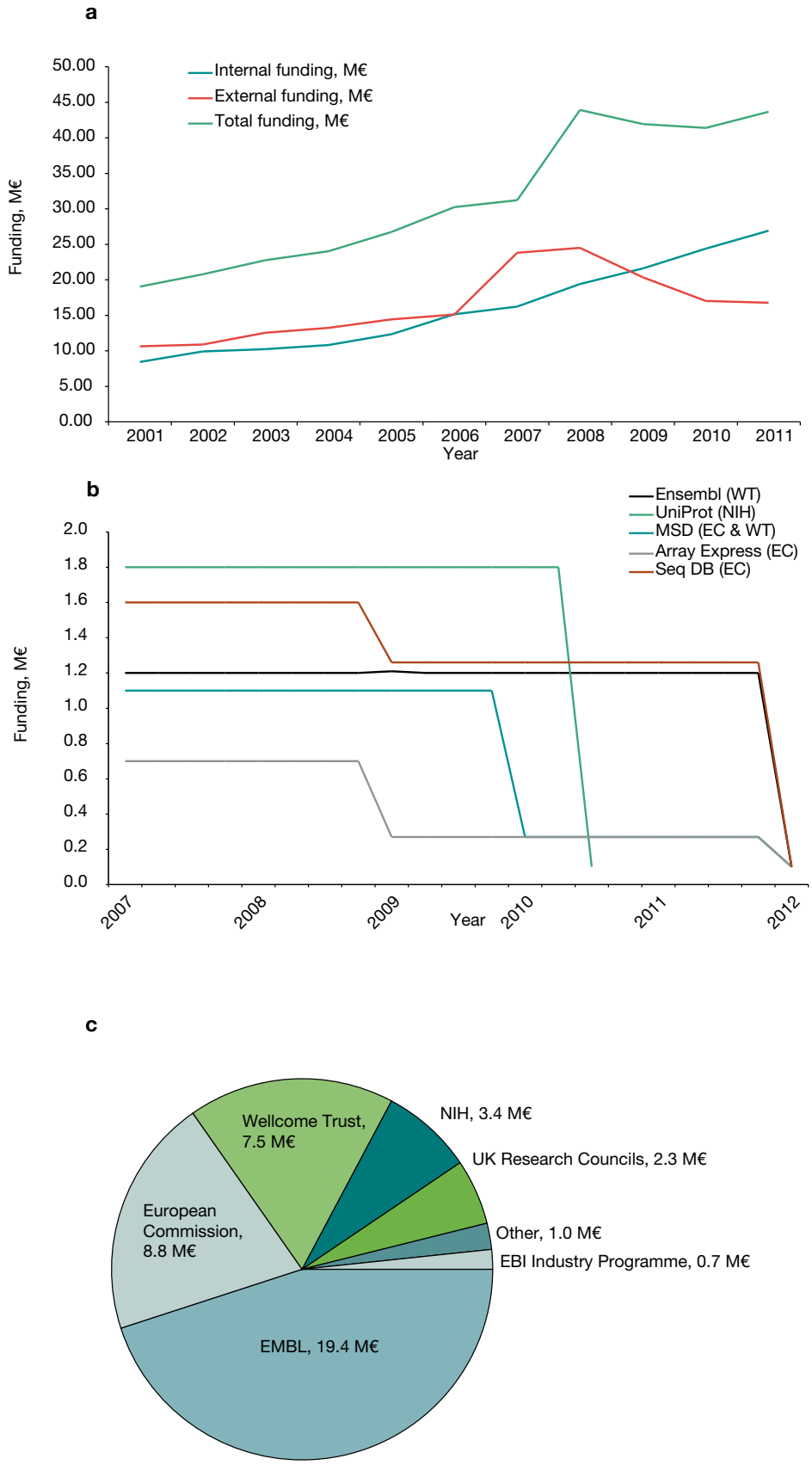


Figure 6. *EMBL-EBI funding. (a) Growth of internal, external and total funds from 2001 to the present day, and agreed internal funds for 2008–2011. (b) Assured external funds for EMBL-EBI's core data resources from the present day to 2011, showing a sharp decline in funding for most of our core resources in 2009. (c) Sources of funding for the year as of November 2008. The Wellcome Trust also supports us through provision of our buildings.*

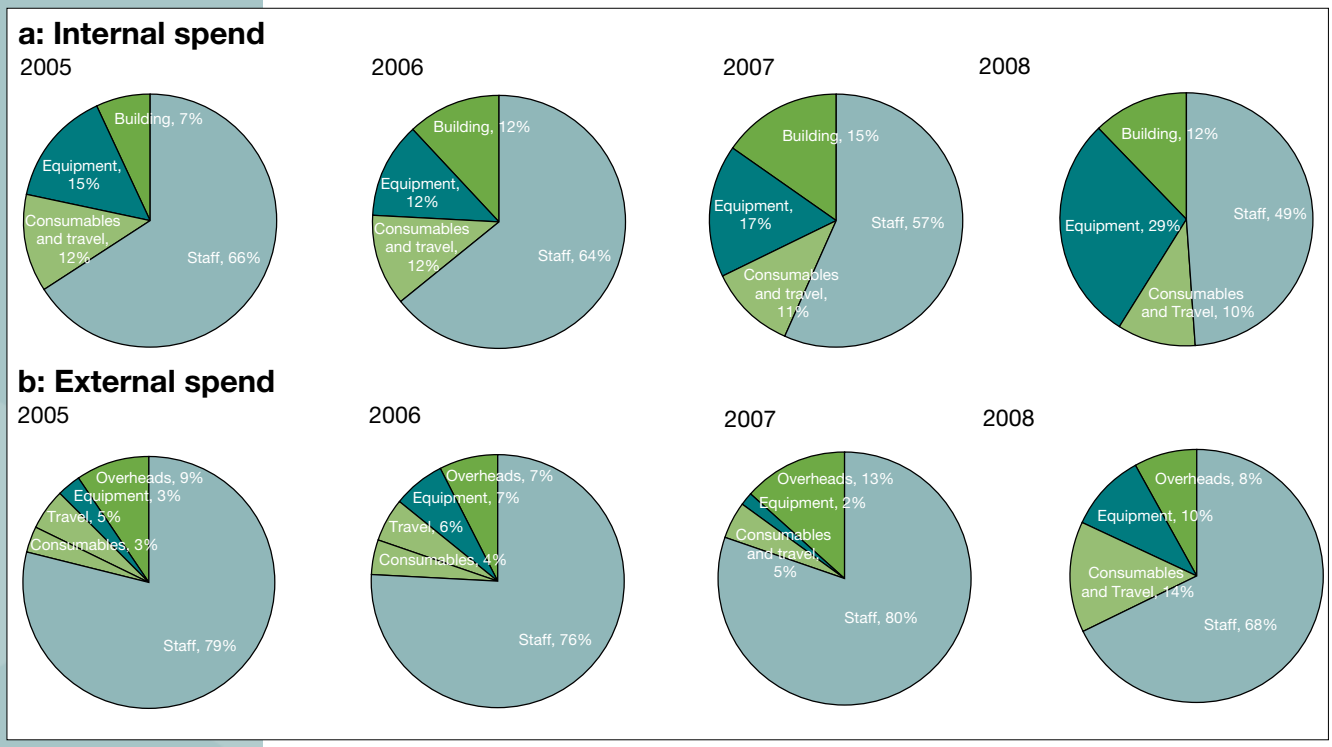


Figure 7. *Breakdown of spend for 2005–2008. (a) Internal and (b) external spend.*

Outreach and Training

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.

OUTREACH

Louisa Wright, Cath Brooksbank, Janet Copeland, Alison Barker

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: PhD students return to their own schools to promote careers in science and give talks to undergraduates to 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. 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 2008.

Owing to the variety of outreach activities performed by representatives of the EBI, one of the main roles of the OTT is to coordinate and support these activities. By working closely with EBI staff members, we aim to ensure consistency of appearance of outreach materials, accuracy of information and aligning of messages with the EBI's key objectives.

The 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.

Building relationships with the media

Louisa Wright, Cath Brooksbank, Anna-Lynn Wegener (EMBL press officer)

Owing to the leading position of the EBI within the European bioinformatics community and its relationship with collaborators and users, the EBI issues its own press releases with support from EMBL's Office of Information and Public Affairs (OIPA). In addition to official press releases, we also post research highlights (often linked to a high-profile publication from one of our research or services groups) and smaller news items on the EBI's front page (http://www.ebi.ac.uk/). From October 2007–October 2008, we released ten EMBL-EBI press releases, contributed to three EMBL press releases, released one press release jointly with the Wellcome Trust Sanger Institute and issued five web-based news announcements.

Details of these press releases and our web-based announcements are shown below in reverse chronological order:

- First version of SBGN language released today, 23 August 2008 (research highlight)
- A one-stop shop for minimal information standards, 7August 2008 (press release)
- Calcium control of molecular learning, 1 August 2008 (research highlight)
- Open access to large-scale drug discovery data, 23 July 2008 (press release)
- European Genotype Archive launched for personal genome information, 14 July 2008 (press release)
- EBI becomes new host of Europe's raw sequence information, 14 July 2008 (press release)
- Zooming in on genetic shuffling, 9 July 2008 (press release with EMBL)
- Scientists fix bugs in our understanding of evolution, 19 June (press release)
- X chromosome exposed, 30 May 2008 (press release with EMBL)
- Securing the future of Europe's biological data resources, 28 May 2008 (press release)

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

PhD, University of Cambridge, 1992.
Postdoctoral work at National Institute for Medical Research, London, and University of Cambridge.
Wellcome Trust Senior Fellow 1995–2006.
At EMBL-EBI since 2002.

Team Members

Scientific Training Officers
Vicky Schneider
James Watson*

Scientific Outreach Officer
Louisa Wright

Workshops and Exhibitions Organiser
Janet Copeland

Outreach and Training Assistant
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.

Ruth Akhtar (EMBL-Bank)
Rafael Alcantara (ChEBI)
Richard Côté (OLS, PICR)
David Croft (Reactome)
Paula De Matos (ChEBI)
Jennifer Deegan (GO)
Emily Dimmer (GOA)
Nicholas Furnham (CSA)
Janna Hastings (ChEBI)
Alan Horne (programmatic access)
Rachael Huntley (GOA)
Bijay Jassal (Reactome)
Andy Jenkinson (programmatic access)
Phil Jones (PRIDE)
Misha Kapushesky (ArrayExpress, Atlas)
Samuel Kerrien (IntAct)
Paul Kersey (Ensembl Genomes)
Michael Kleen (UniProt web services)
Eugene Krissinel (PDBe)
Eugene Kulesha (DAS)
Roman Laskowski (PDBsum, ProFunc, Druggability portal and Thornton group resources)
Rodrigo Lopez (EBI search tools)
Lennart Martens (Proteomics)
Jennifer McDowall (InterPro, UniProt)
Hamish McWilliam (EBI search tools, sequence searching)
Anika Oellrich (EBI text mining tools)
Tom Oldfield (PDBe)
Sandra Orchard (UniProt, InterPro, IntAct)
Bert Overduin (Ensembl)
Samuel Patient (UniProt web services)
Sharmila Pillai (CiteXplore)
Dietrich Rebholz-Schuhmann (EBI text mining tools)
Florian Reisinger (ENCORE, EnVision)
Peter Rice (EMBOSS)
Gabriella Rustici (ArrayExpress, Atlas, Bioconductor)
Gaurav Sahni (PDBe)
Esther Schmidt (Reactome)
Sanchayita Sen (PDBe)
Giulietta Spudich (Ensembl)
Peter Sterk (Integr8)
Peter Stoehr (Literature resources)
Glen Van Ginkel (PDBe)

- Platypus genetic blueprint reveals the early history of mammals, 7 May 2008 (press release)
- Sanger Institute and European Bioinformatics Institute top list of most influential UK research, 1 May 2008 (press release with Wellcome Trust Sanger Institute)
- Strength of European rat genetics research highlighted in special focus issue of *Nature Genetics*, 29 April 2008 (research highlight)
- An unexpected way to cause leukaemia, 8 April 2008 (press release with EMBL)
- International consortium announces the 1000 Genomes Project, 22 January 2008 (press release)
- Analysing large-scale proteomics projects with latent semantic indexing, 22 January 2008 (research highlight)
- ArrayExpress database doubles in size to 100,000 hybridisations, 11 December 2007 (press release)
- The East Wing and a new dawn for the EMBL-EBI, 23 October 2007 (press release)
- Coverage and error models of protein–protein interaction data by directed graph analysis, 22 October 2007 (research highlight)

The number of press releases issued in this period has shown an increase from the previous two years (Figure 1). The majority of our releases are targeted at the specialist press as part of our media strategy to capitalise on the EBI's reputation in the bioinformatics field and also protect our relationship with the mainstream media by only providing them with news relevant to a more general audience. A particular highlight of our media activities in 2008 was the worldwide media coverage received by the completion of the platypus genome (May 2008). The sequencing effort involved several international laboratories and so the coordination of the media announcement also involved wide-scale collaboration. Several press releases were issued simultaneously and all resulted in high-profile coverage. In particular, the EMBL-EBI issued release led to a radio interview with Ewan Birney on BBC Radio Cambridgeshire and links to examples of the print coverage received are given on the EMBL press office's coverage archive webpage at http://www.embl.org/aboutus/news/EMBLinpress/index.html.

Press release coverage and media impact are monitored using several tools. We use a regular electronic media monitoring service that returns details of press coverage based on the use of keywords, information provided from AlphaGalileo and EurekAlert – online media resources that we and EMBL OIPA post our press releases to, and Google Alerts, which also returns results based on selected keywords. In 2008, we have also worked towards enhancing our relationship with the specialist

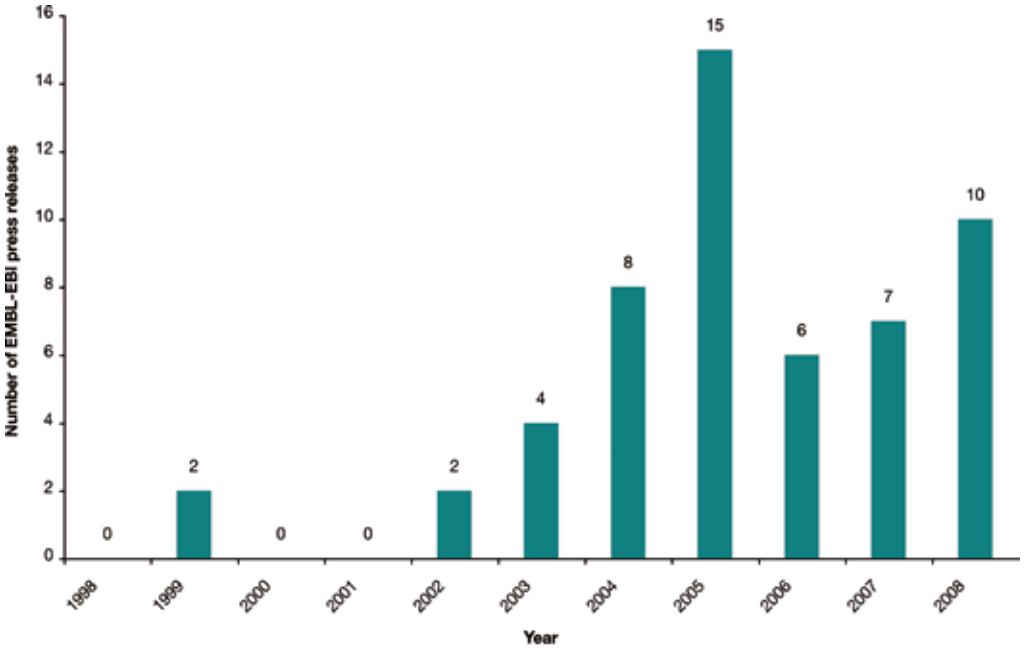


Figure 1. *Number of press releases issued by EMBL-EBI each year from 1998 to 2008.*

media. We initiated this by requesting feedback from our media contacts on their perception of the quality of EBI's media communications (press releases and mode of contact). The returned feedback demonstrated a positive appreciation for the quality of the EBI's press releases and the relevance of our targeting, reinforcing our strategy of carefully managing the decision of when to issue a press release and also press release distribution.

Broadcast media work in 2008

Louisa Wright and Cath Brooksbank

2008 also saw increased representation of the EBI in the broadcast media. We have hosted film crews from German and Latvian television as well as from Apple Inc.

In addition to assisting external filming projects, we commissioned our own production of an ELIXIR (European Life Sciences Infrastructure for Biological Information) promotional DVD, narrated by Janet Thornton and centred on the need for sustainable funding for Europe's biomolecular data resources. The DVD is being used as a dissemination tool and over two hundred have been sent to the national funding agencies in Europe.

EBI literature

Louisa Wright with support from outreach representatives from each team/group

Two of our most-used resources are our introductory 'EBI In a Nutshell' brochure and the range of resource fact sheets we produce. These form the core of our promotional literature and are regularly distributed at conferences and training events. The fact sheets provide an introductory-level guide to each specific resource and are routinely included in our hands-on training course manuals. In keeping with the EBI's commitment to develop its resources and improve usability, the resources evolve at a fast pace. This requires us to keep abreast of new developments affecting functionality and interface appearance and update our range of resource fact sheets when required. This year saw a revision of the fact sheets on all the EBI's core data resources.

The 'EBI In a Nutshell' brochure was also revised to include details of the new services and research groups recently formed at the EBI. Our brochures and fact sheets are downloadable from http://www.ebi.ac.uk/Information/Brochures/.

OIPA produces EMBL's promotional literature and the Outreach and Training team is responsible for contributing EMBL-EBI specific information for these. We collated and edited the EMBL-EBI section of EMBL's 'Research at a Glance' (http://www.embl.org/aboutus/news/publications/pdf/raag07-08.pdf), and our group and team leaders are interviewed to provide features for the EMBL Annual Report. The 2007–2008 report includes articles on the ELIXIR and ENCODE projects, the development of proteomics standards, and research from the group of Wolfgang Huber. Less formally, we are regular contributors to the bimonthly EMBL newsletter, EMBL*&cetera,* which is distributed to an average readership of 6,000.

Promoting EBI coordinated EU-funded projects

Louisa Wright and Cath Brooksbank with support from project managers

As coordinating partner of the ELIXIR project, the EBI is responsible for the project's communication and outreach activities. This has involved working closely with the project coordinator, Janet Thornton, and project manager, Andrew Lyall, to develop materials to promote awareness of ELIXIR and communicate the project's objectives, benefits and rationale. In addition to overseeing the production of the ELIXIR DVD, we have overseen the design of two ELIXIR-themed display banners for use at stakeholder meetings and conferences, and produced a project leaflet.

The Outreach and Training team also coordinated the commission of an editorial in the eStrategies magazine to promote the EU-funded projects coordinated by the EBI (ELIXIR, EMBRACE, ENFIN and BioSapiens). The eStrategies magazine has a Europe-wide readership of 39,000 and the editorial was published in September 2008.

Panel 1 continued: EMBL-EBI Outreach and Training representatives

Robert Vaughan (EMBL-Bank)
Sameer Velankar (PDBe)
Andy Yates (Ensembl programmatic access)
Vicky Schneider (EBI overview and tools)
James Watson (Tempura, PDBsum, ProFunc)

Promotional materials

Louisa Wright

We have supported the promotion of individual EBI resources and collaborations by producing display banners for use at conferences:

- ELIXIR banner – displayed at stakeholder meetings (April and November 2008) and European Conference on Computer Biology (ECCB) in Sardinia;
- PDBe banners – displayed at the International Union of Crystallographers (IUCR) meeting in Osaka, Japan;
- HGNC banner – displayed at HGM in Hyderabad, India;
- Proteomics services banner – displayed at the HUPO Congress, Amsterdam.

Exhibiting at conferences

Louisa Wright, Janet Copeland, Cath Brooksbank, Vicky Schneider, James Watson, Alison Barker

This year, the Outreach and Training team has exhibited at eleven conferences (Figure 2), predominantly promoting the EBI's resources, training activities and career opportunities (see list below). We have collaborated with OIPA to represent EMBL at exhibitions where a combined presence was beneficial. While exhibiting at scientific conferences has been one of the main mechanisms to promote the EBI and its services to experimental researchers and computational biologists, we also represented the EBI as part of an EMBL exhibition at the *Nature* Source Event careers fair (September 2008, London) in order to raise awareness of the range of career opportunities available within EMBL.

- ECCB, Sardinia, September 2008
- HUPO, Amsterdam, August 2008
- IUCR, Osaka, August 2008 (support for EMBL and wwPDB stands)
- Intelligent Systems for Molecular Biology (ISMB), Toronto, July 2008
- International Congress of Genetics, Berlin, July 2008
- FEBS, Athens, June 2008 (support for EMBL stand)
- Genomes to Systems, Manchester, March 2008
- AAAS, Boston, February 2008 (support for EMBL stand)

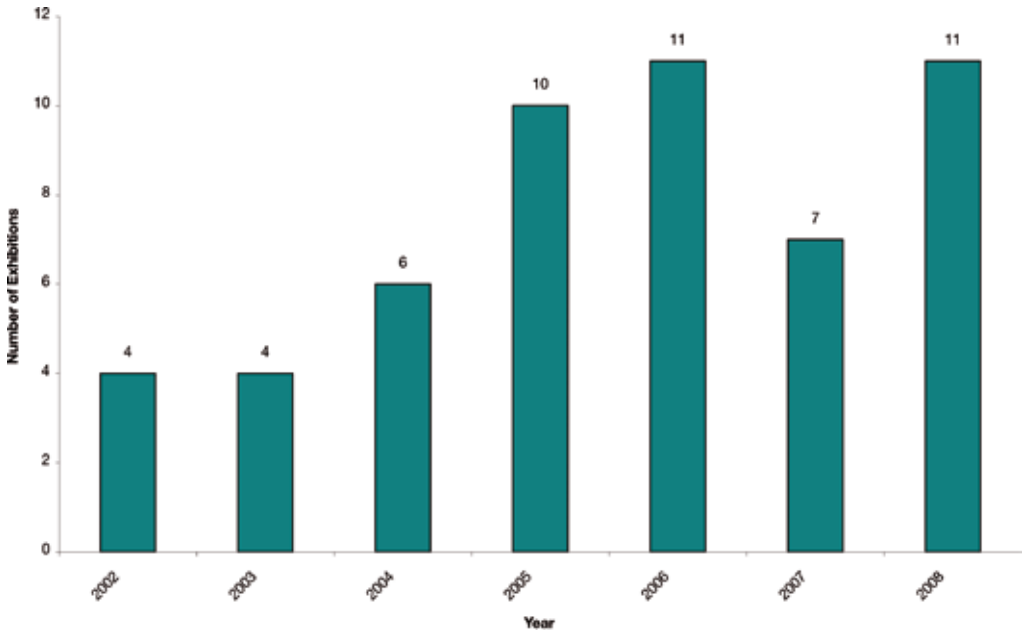


Figure 2. *Conferences at which EMBL-EBI has exhibited, 2002–2008.*

- Plant and Animal Genomes (PAG), San Diego, January 2008
- International Conference on Systems Biology (ICSB) 2007, Long Beach, October 2007 (support for EMBL stand)
- HUPO 2007, Seoul, October 2007

The EBI's Outreach and Training representatives have played a major role in providing on-stand support at these events and also in promoting the EBI at conferences not exhibited at by either the EBI or EMBL.

Promotion of training

Louisa Wright, Alison Barker, Janet Copeland, Vicky Schneider, James Watson, Cath Brooksbank

In October 2008, the Hands-on User Training Programme completed its first year of operation. The programme has been promoted through several channels: posters were mailed out to training contacts, upcoming training courses were featured on the EBI's front page, email alerts were used to promote individual courses to our training contacts; course details were posted to special interest web groups of direct relevance to the course focus; and training programme flyers and individual course flyers were handed out at conferences attended by the EBI throughout 2008.

During 2008, we also launched the first courses of the EBI's eLearning project for beta testing. The objectives of the eLearning materials are to maximise access to training on the EBI's resources and overcome financial and time restrictions that prevent users attending scheduled training events. We promoted the eLearning project by developing a promotional postcard containing information on the courses available and details of how to access them. This was distributed to stand visitors at all conferences attended by the EBI in 2008. We encouraged participation in the beta-testing phase of the project by running a competition offering entry into a prize draw when a course was completed and feedback received. An announcement of the launch of the eLearning course was also listed on the EBI's front page. Use of the eLearning tools has been steadily increasing throughout 2008 (a total of 222 users registered from July–December 2008) and we hope to see this continue as the resource is developed further in 2009.

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 webpages and distribution of the PhD programme brochure at conferences and events.

Masters open day

Vicky Scheider, James Watson, Louisa Wright, Alison Barker, Janet Copeland, Cath Brooksbank

The EBI's Masters open day has been run twice in 2008, firstly in March and secondly in November. The newly introduced November-timed event was held in order 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. The next Masters open day will be held at the EBI on 19 March 2009.

Public engagement

Cath Brooksbank, Louisa Wright and volunteers from the EBI in collaboration with the Wellcome Trust Sanger Institute

The EBI is a regular contributor to the 'Biology Zone', which forms part of the Cambridge Science Festival's 'Science on Saturday' event and we work closely with the Wellcome Trust Sanger Institute in planning and delivering our festival activities. In addition we are members of the Cambridge Genetics Education Network (CGENe; other members include the Wellcome Trust Sanger Institute, Medical Research Council and the Cambridge Genetics Knowledge Park) and liaise closely with the group to ensure a diverse and stimulating collection of activities for the festival's theme.

For this year's theme, 'The world of Science', EBI and Sanger Institute staff volunteers presented a range of hands-on activities: DNA bracelet making, jelly baby DNA models, origami DNA in addition to some computer-based activities (e.g. what's your name in DNA?). These were mainly based

within the Biology Zone which received over 2,500 visitors over six hours. Due to the popularity of the event overwhelming the venue's capacity restrictions, there is a queue to enter the building for most of the day. To overcome these restrictions, and where practical, activities were also used to entertain the public waiting to enter the Biology Zone. The materials we have developed for schools and which form part of our science festival activities are available from the EBI's training pages at www.ebi.ac.uk/training/schools/.

In addition to our usual activities, EBI and Sanger volunteers were also involved in a pilot Video and VoxPops project involving film-making and interviewing visitors in the Biology Zone. Filming training and direction was provided by Jonathan Sanderson, a freelance film producer specialising in science-themed projects. Three film crews interviewed the public on their thoughts on personal genomics, exploring issues such as impact on life decisions, security of information and access to personal genome data by researchers. The footage gathered during the day was used to produce three VoxPop videos that are available on the Wellcome Trust Sanger Institute's dedicated public engagement website http://www.yourgenome.org/feature/feature200803_ScienceWeek.shtml.

The EBI and the Sanger Institute's public engagement team have begun to plan our activities for the 2009 Science Festival. Our activities this year will focus on mobile activities that can be use for short, interactive demonstrations to stimulate further dialogue with people waiting to enter the main Biology Zone. The VoxPop-style filming will also be repeated with roaming interviews (audio recording and filming) with members of the public in the queue.

TRAINING

Vicky Schneider, James Watson, Cath Brooksbank, Alison Barker, Janet Copeland with support from OTRs

The mission of the EBI's training programme is to provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators. Our audience spans wet-lab experimentalists and computational scientists. Ultimately, we aim to extend our expertise to provide trainers with knowledge, training strategies and tools. This will enrich Europe's community of bioinformatics trainers to widen the use of Europe's bioinformatics resources, meet training needs and support the development and delivery of training materials.

The Outreach and Training team coordinates the EBI's user training programme, designed to equip users of the EBI's bioinformatics services in EMBL member states and beyond, with the knowledge they need to become confident users of Europe's core biological data resources. Of equal importance is the role of OTT in coordinating several activities aimed at training EMBL-EBI personnel. Internal training is aimed at improving the scientific and transferable skills of our personnel; external training focuses on teaching scientists to use the EBI's data resources and tools.

Training needs of EBI personnel
We support the training needs of EBI personnel by: coordinating training activities for the EBI's predoctoral students; applying for external funding for students and visitors; and by acting as a contact point for EMBL's non-scientific training and development programme, which is coordinated from EMBL's main laboratory in Heidelberg.

Industrial user training
We interact closely with the EBI's Industry Programme coordinator and regularly meet its members to ensure that the EBI's user training complements the Industry Programme's training events (see separate section on page 31).

User training programme
The user training programme spans three categories of training:
1) **The Bioinformatics Roadshow:** a travelling user training programme tailored to the needs of users of Europe's main data resources;

2) The **EBI Hands-on User Training programme**: a series of two- to three-day courses, held in the EBI's IT training suite, which aims to familiarise experimental researchers with the EBI's core data resources;

3) The **EBI eLearning pilot project**: a web-based portal that serves both end-users and trainers.

2008 has witnessed the first year of our in-house user training programme, the growth and beta testing of our eLearning pilot project, and the successful continuation of the Bioinformatics Roadshow.

All three training elements have been received with great enthusiasm and their demand continues to increase. This reflects the greater need for experimental biologists to become expert users of the EBI's bioinformatics tools and resources, as well as the interest from bioinformaticians and computational biologists in receiving tailored training on technical aspects and programmatic access to biological information.

The user training programme relies heavily on the Outreach and Training representatives from the various groups at the EBI. The number of EBI staff specifically working on training has increased in the past year. Panel 1 lists our trainers and their main training topics.

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.

Training on tour: the Bioinformatics Roadshow

Vicky Schneider, James Watson, Janet Copeland, Cath Brooksbank and OTRs

The Bioinformatics Roadshow is run under the auspices of the EU-funded FELICS Integrated Infrastructure Initiative, in collaboration with our partners the Swiss Institute of Bioinformatics (SIB) the European Patent Office (EPO) and the BRENDA database at the University of Braunschweig. The FELICS contract began in February 2006 and will finish in February 2009. During this time we have successfully expanded the roadshow model from user training for a single database (MSD) to cover the vast majority of the data resources served by FELICS. During 2008 we have had an overwhelming demand for roadshows. The success of this programme is also reflected by several requests for repeat roadshows in the same location. We have now exceeded our contracted deliverables for FELICS (18 roadshows over three years), organising a total of 34 roadshows in 14 countries. Through these roadshows we have addressed the needs of European scientists for hands-on training in the use of Europe's most widely used bioinformatics data resources and tools. Logistics and administrative aspects are managed by our Workshops and Exhibitions Organiser Janet Copeland, whilst the scientific programme for each roadshow is coordinated by Scientific Training Officers Vicky Schneider and James Watson. Figure 3 shows a summary of the roadshows organised by OTT since 2006, as well as those scheduled for the near future.

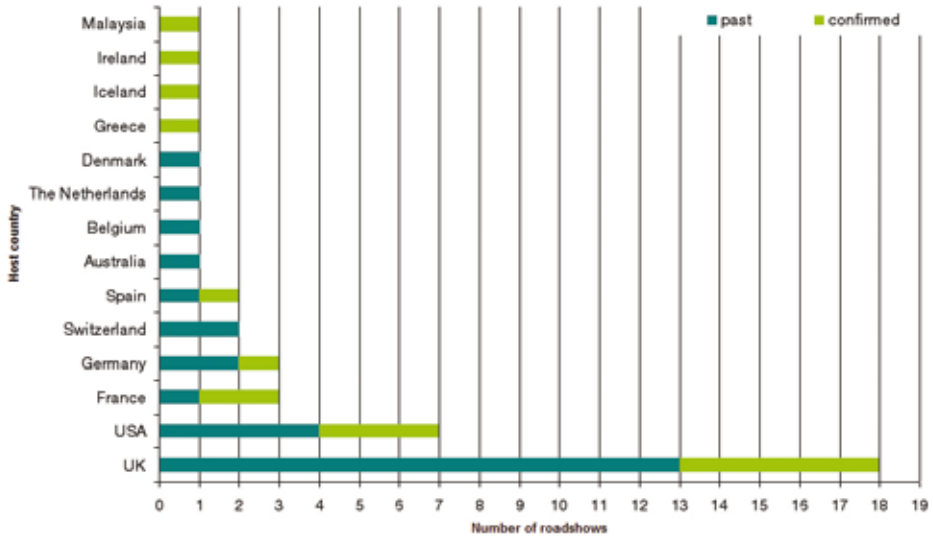


Figure 3. *Summary of roadshow locations by country. Figures represent roadshows organised by the Outreach and Training team.*

Roadshows typically last two days and involve an average of three trainers, depending on the variety of resources and tools covered. Each roadshow is self-funded, so the host covers the travel and subsistence costs of trainers, and provides suitably equipped training rooms that include computing facilities. The programme is designed together with the hosts, with input from the trainers, to guarantee that the learning needs of the audience will be met (for example, programmes are often tailored to the host institute's interests). To date, the majority of training has been undertaken by EBI trainers;

SIB has also organised several roadshows, and for the first time we have run joint roadshows both with SIB and EPO.

We have also implemented systems for monitoring and improving the roadshow programme by:

- collecting feedback from our named contact at each host institute, as well as collating and analysing trainee feedback using a standardised evaluation form;
- establishing a centrally-managed registration system for roadshows. This provides the trainers with prior information about the trainees' areas of interest, allowing them to shape each roadshow accordingly. It also helps us to monitor how many people we're reaching;
- creating generic templates for the production of promotional literature, and have developed templates for the production of training materials;
- making plans to invest in more trainers to meet demands for training on specific topics.

Figure 4 shows the main areas covered by the past roadshows. Roadshows planned for 2009 show an increasing demand for training in transcriptomics resources and sequence searching tools.

The average number of attendees per roadshow is 37 participants, from which we have gather feed-back at the end of each roadshow using an online evaluation form. Much of the information gathered is qualitative, but on the questions where we can derive a quantitative score by asking trainees to rate us from one to five (from poor to outstanding), our average score is four or more.

The Hands-on User Training Programme

Vicky Schneider, James Watson, Alison Barker, Janet Copeland, Cath Brooksbank

Since the launch of the Hands-on User Training Programme in September 2007, we have run eleven courses (September 2007 to August 2008) with an average attendance of 20 participants per course. As might be expected from a programme that is run from the UK, the majority of our trainees are UK based, with about one third travelling from outside the UK (Figure 5).

The majority of our courses focus on a specific topic (e.g. transcriptomics, proteomics, genomics, structures, interactions and pathways), and familiarise trainees with a range of EBI data resources and tools relevant to each topic. Two broader courses ('A dip into the EBI's data resources' and 'A walk through the EBI's data resources') offer a general introduction to all of the EBI's core data resources.

The aim of the programme is to provide experimental researchers with the knowledge they need to identify the most relevant data resources for their areas of work and gain the most from them. This concept has been extended to the computational biology community through courses on programmatic access and data integration. A new group of 'programmers who train' has arisen from these

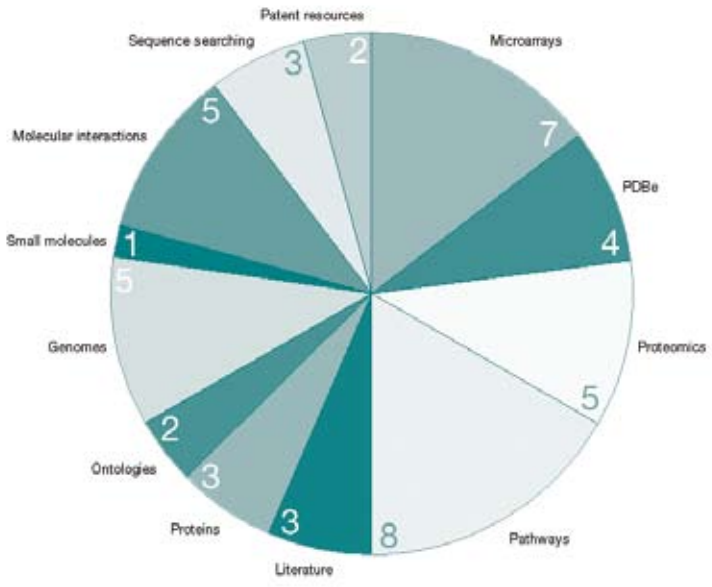


Figure 4. *Main subject areas covered by the twelve roadshow events in 2008. Numbers show the number of times each topic was included in the training programme.*

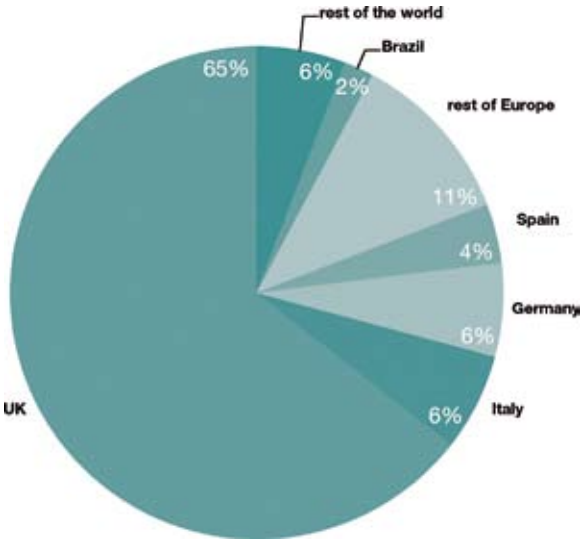


Figure 5. *Origins of participants attending hands-on training courses in 2008.*

activities. These courses have also created a nucleus of communication, the benefits of which may extend beyond training to the development of the data resources. Details of forthcoming courses can be found at http://www.ebi.ac.uk/training/handson/. Previous courses are listed at http://www.ebi.ac.uk/training/handson/previous.html.

The coordination of the hands-on programme involves every member of the team, and in our first year of operation we have worked extremely hard to listen to feedback from trainees and trainers, and improve our courses in response to their suggestions. We now provide a printed training manual as well as online training materials, have implemented a new course registration system that brings us in line with the EMBL Course and Conference Office, and are in the process of streamlining our promotional procedures. We are especially grateful to all those, both within and beyond the EBI, who have actively supported us by helping to advertise the hands-on programme.

We use online evaluation forms to collect feedback from our trainees. The feedback is very positive and attendees are particularly pleased with the overall course organisation, materials provided as well as scientific content and knowledge of the trainers.

ELEARNING PILOT PROJECT: NO MORE SPACE BARRIERS

Cath Brooksbank, Vicky Schneider, James Watson, OTRs with some work contracted to The Consultants-E

In July 2008 we began to beta test our eLearning platform, which currently contains the following courses:

- EBI and EB-eye;
- Sequence searching;
- Patent searching;
- Ensembl (author: Giulietta Spudich);
- Transcriptomics (authors: Gabriela Rustici and Eleanor Williams).

User surveys collect information about users' experiences with the course structure and training elements, scientific content, clarity and usability. This feedback will be used to improve the eLearning platform in future. Currently, most courses have six main training elements that are designed to appeal to different learning styles: video tutorials; printable (PDF) tutorials; key concepts quizzes; reflective tasks; and a glossary of terms.

Given the dynamic nature of bioinformatics resources and tools, one of our major challenges is producing materials that do not immediately become outdated. For this reason we have rationalised the production of videos, limiting them conceptual topics rather than demonstrations of how to use the data resources.

MONITORING USER TRAINING IN 2008

James Watson, Vicky Schneider, EMBL-EBI outreach and training representatives, EBI's Administration Team

EMBL-EBI personnel have participated in 330 training-related events throughout 2008. Through a tracking system developed by James Watson in collaboration with the administration team, we have substantially improved collection and tracking of all training-related activities at the EBI.

Besides the EBI-wide training programme discussed above, EBI staff participate in a large number of other training events. Over the past year more than 130 staff members have actively participated in training or training-related activities consisting of: 102 demonstrations/hands-on training events, 60 posters and 193 presentations. If we consider the audiences reached through demonstrations and talks at conferences, we have reached more than 30,000 people. Audiences are predominantly pre-docs, postdocs and academics, but there has also been a substantial amount of outreach to industrial researchers. Tracking of all training-related events remains a non-trivial task that we will continue to streamline in the future. A general overview (incomplete but likely to be representative) of the resources covered by EBI staff in different training events is shown in Figure 6.

A COLLABORATIVE APPROACH TO BIOINFORMATICS USER TRAINING

Vicky Schneider, James Watson, EMBL-EBI outreach and training representatives, Cath Brooksbank

The EBI's user training activities are part of a complex landscape that includes many other training initiatives. We are developing strong relationships, both within and beyond EMBL, so that we can make the most of our synergies with other training programmes.

EMBL-wide collaboration

We meet regularly with training representatives from the other EMBL sites to make the most of opportunities to work together on training and communication projects; we are working with OIPA, EICAT and the Course and Conference Office to pool resources and share expertise whenever this is practical. We are currently working on an improved conference registration system and are developing systems for promoting courses and conferences together.

Campus-wide collaboration

We meet regularly with our colleagues on campus who run the Wellcome Trust Advanced Courses and Wellcome Trust Scientific Conference Programme to discuss opportunities for working together. One important result of these discussions is the development of a new portal, www.hinxton.org, which provides a one-stop shop for scientists who want to find out about all the courses and conferences on the Wellcome Trust Genome Campus. www.hinxton.org will launch early in December 2008.

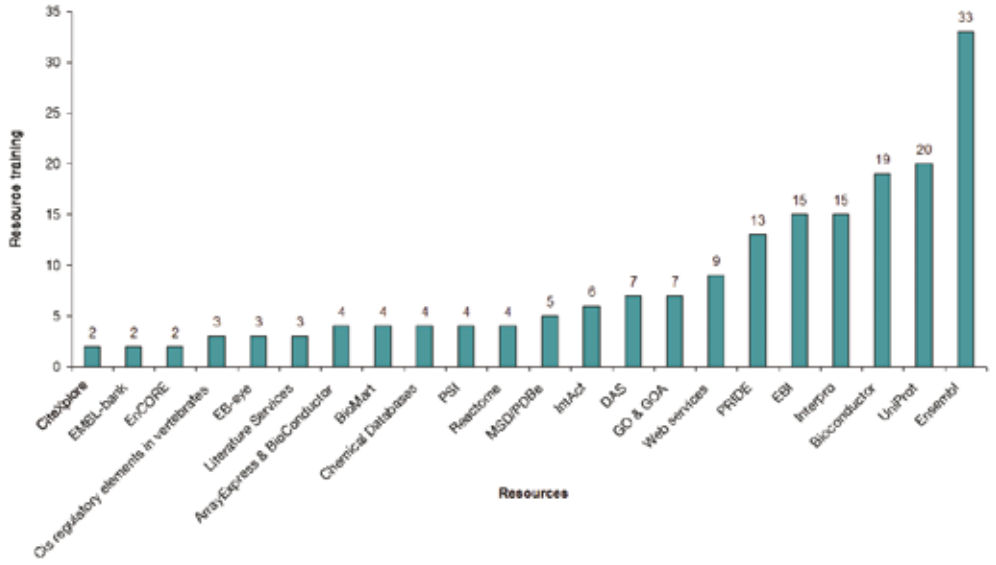


Figure 6. Number of times each data resource has been a component of a training event in 2008 (data collected from trainer reports).

Several jointly organised courses and conferences are now in the pipeline, and we are working on holding related events back-to-back for the benefit of scientists who have to travel substantial distances to reach the campus.

EMBO courses
EMBO Practical Courses from individual research groups (Alvis Brazma, Kim Henrick) are held at the EBI and we are exploring further synergies with EMBO.

Other universities and institutes Europe-wide
We have established collaborations with the University of Cambridge, the Gulbenkian Institute, ICGEB and other organisations that hold regular bioinformatics training courses; and EBI trainers regularly contribute to these events.

A pan-European infrastructure for bioinformatics user training

We are also involved in developing a bioinformatics user training strategy for Europe through ELIXIR (www.elixir-europe.org). Anna Tramontano and Cath Brooksbank co-chair the training strategy workpackage, which is project managed by Vicky Schneider. In consultation with a training strategy committee, we are in the process of developing a training infrastructure that will improve accessibility to bioinformatics training for life scientists throughout the European Union. We are also involved in testing the effectiveness of different training methods, with a particular focus on e-learning.

STAFF AND STUDENT EDUCATION

Vicky Schneider, Nick Goldman, Cath Brooksbank

Fundraising

The Outreach and Training team coordinates several activities aimed at training EMBL-EBI personnel. We have successfully applied for BBSRC Case studentships. We also continue to provide training-related information for other EMBL-EBI staff members and industry partners, as well as advice on applications for externally-funded studentships.

EMBL-EBI PhD student training

The EBI's pre-doctoral students have a core of training-related events in which they are obliged to participate:

Bioinformatics course: a three-day course covering a variety of bioinformatics topics is designed for students enrolled in the EMBL PhD programme. These are primarily experimental biologists who require an introduction to some of the approaches and tools used for analysing sequences, structures and expression data. The EMBL-EBI pre-doctoral students organise and teach on this course, providing them with valuable experience in event organisation and training. The course organisers exchanged many ideas with the Outreach and Training team this year, and we plan to remain involved in this event in future years.

'Primers for Predocs' series: an annual series of seminars for all first-year EMBL-EBI PhD students. Representatives of each EMBL-EBI group introduce students to their group's activities using lectures, hands-on exercises and open discussions. We used the feedback from 2007 to improve the series in 2008. The pre-doctoral students work in pairs to answer some questions posed by experts from the data resources and then present their findings. These presentations are followed by talks and, where possible, hands-on tutorials. Feedback collected from this series has stimulate us to change the format from a weekly series to an intensive five-day course for 2009.

Non-Scientific Training and Development Programme

Cath Brooksbank, Vicky Schneider, Alison Barker, in collaboration with Rebecca West (EMBL Heidelberg)

The EMBL Non-Scientific Training and Development Programme for all EMBL-EBI members of personnel (www.ebi.ac.uk/training/internal/) has been a major success. Several courses are run in the IT training suite and OTT members support the programme by helping to promote it within the EBI, as well as communicating specific training needs and requests to EMBL.

FUTURE PROJECTS AND GOALS

Our future goals all revolve around increasing accessibility to Europe's most widely used data resources. In the immediate future, a new EU-funded Integrating Action, SLING, will help us to serve those EU countries that do not have the resources to host a roadshow. We hope to be able to provide end-user training and user support for the data resources made available through this project, through fully-funded training roadshows. Contract negotiations are currently in progress. Looking further into the future, 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. The EBI's services form a significant subset of the bioinformatics resources and tools for both SLING and ELIXIR; the EBI's own outreach and training activities will therefore provide important contributions to, and test cases for, these wider efforts. ELearning modules covering UniProt, EMBL-Bank, Reactome, PRIDE, InterPro and MSD/PDBe are being prepared for launch in 2009. Moreover, the arrival of scientific training officer Rosemary Wilson at EMBL Hamburg has initiated a series of training collaborations, including the development of joint e-learning materials.

Industry Support

The fourth element of the mission of the EMBL-EBI is to help disseminate cutting-edge technologies to industry. Industry users already comprise a significant proportion of the EBI services user-base and will continue to grow. In addition to access to online services and the generic training courses that are provided by the Outreach and Training team (see page 19), EMBL-EBI provides further support to industry through two specific programmes: the EMBL-EBI Industry Programme and SME support forum.

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 15 members (see Panel 1 overleaf).

Scope

The statement detailing the scope of the Industry Programme was recently revised. The newly refined programme will:

- provide a forum for interaction between the EBI and our users in industry;
- provide training for our commercial users;
- inform 'industrial users' of EBI's status and future plans;
- feed industry requirements into the EBI's planning;
- provide a neutral meeting place for inter-company interactions on bioinformatics;
- coordinate workshops on topics decided by the programme – gathering expert speakers (industrial and academic);
- initiate 'special projects' at the EBI with targeted collaborative funding;
- liaise as appropriate with other industry initiatives.

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 2008 are listed in Table 1.

Strategy meetings

The quarterly 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;
- be updated 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' priorities and also provide time for detailed interactions between EMBL-EBI staff and representatives from member companies.

Pre-competitive activities

EMBL-EBI's bioinformatics resources are designed to be relevant to researchers from a spectrum of different research areas. Therefore, we are currently working with the Industry Programme members and key external research groups to develop new data resources and analysis systems to assist researchers in interpreting genomic data in a pharmaceutical/agricultural context and integrating this with chemical information. The Druggability Portal pilot project has shown good progress during the year and will now be integrated into future chemogenomics initiatives.

Dominic Clark

Industry Programme Coordinator

PhD Medical Informatics,1988.
Imperial Cancer Research Fund 1987–1995.
UK Bioinformatics Manager, GlaxoWellcome R&D Ltd, 1995–1999.
Vice-President Informatics, Pharmagene, 1999–2001.
Managing Consultant and Bioinformatics Specialist, Sagentia Ltd, 2001.
At EMBL-EBI (secondment) since 2006.

Panel 1. EMBL-EBI's Industry Partners in 2008

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 KGaA
Nestlé Research Centre
Philips Research
Pfizer Ltd (from January 2008)
Syngenta Limited
Sanofi-Aventis Recherche & Développement
Unilever

Highlights of 2008

The key highlights of 2008 are:

- the greater engagement by Industry Programme members in workshops. This reflects increased emphasis on orchestration of pre-competitive elements in the workshops combined with expert level speakers;
- greater emphasis on interdisciplinary links between bioinformatics and the fields of chemistry and medicine;
- the Druggability Portal project pilot has successfully demonstrated how pre-competitive projects can be initiated within the EMBL-EBI;
- greater dialogue between the programme members and the scientific community with EMBL-EBI acting as broker.

During 2009 we anticipate further broadened and deepening of workshop topics and the possible creation of further pre-competitive initiatives.

Training collaborations

A number of PhD studentships are currently being supported by members of the Industry Programme.

THE EMBL-EBI SME SUPPORT FORUM

EMBL-EBI launched its support forum for Small-to-Medium Enterprises (SMEs) in 2003 through a grant from the UK Department of Trade and Industry (DTI) under the 'Harnessing Genomics' initiative. Several different types of company, including bioinformatics service providers, drug discovery companies, platform technology companies, and those providing integration services joined the SME support forum, each having different support needs. Since the end of the period of DTI funding in March 2007, the forum has received no sustained external funding. The focus has therefore switched to annual information meetings funded through outreach elements of EU grants.

The second Annual Forum for SMEs was held in Berlin from 27–28 October 2008 and included speakers from EMBL-EBI and collaborators. Full details of the meeting and the presentations can be access from http://www.enfin.org/page.php?page=sme_meeting_2008. Due to the popularity of this format, it has been decided to continue organising such events in 2009 with meetings planned at alternative centres in Europe.

Dates	Title
4–5 February 2008	Druggability Portal workshop
10–11 March 2008	Structural biology tools and services provided by the EBI
19–20 May 2008	ChEBI users' workshop
19–20 June 2008	Disease ontologies and information: its linkage to genes, gene ontologies and phenotypic information
30 September – 1 October 2008	Target CV workshop
20–21 October 2008	Integration of genomic information related to crop diseases and pests
7 November 2008	Druggability Portal (for chemists)
20–21 November 2008	New sequencing technologies and their applications in medicine

Table 1. *Industry Programme workshops in 2008.*

Systems and Networking

INTRODUCTION

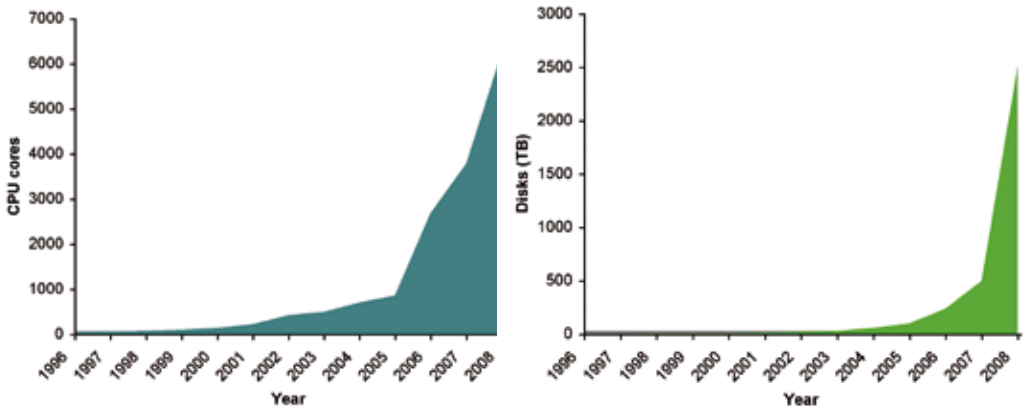
The Systems and Networking team manages the EMBL-EBI IT infrastructure. This includes compute and database servers, storage, desktop systems and networking, as well as managing our campus connection. An important task is supporting EMBL-EBI users in their daily activities. The team works closely with all project groups maintaining and planning their specific infrastructures. The IT environment consists of more than 6,000 CPU cores (Figure 1) and 2.5 petabytes of disk storage (Figure 2). Looking ahead we expect these figures to double again in 2009.

STORAGE

Total storage has more than doubled again this year (Figure 2). As well as expanding our storage we are continuously trying to improve the underlying architecture. This includes data replication and backups. We are now replicating almost all of our disk systems. We have installed a new tape backup system and are in the process of testing it before deployment.

NEW MACHINES

We have continued to add servers to our PC farms. We have begun favouring blade servers because we are filling up our allocated data centre space at an alarming rate (Figure 3). To tackle this issue we have started utilising the densest blades on the market so that we can increase the density of servers. In spite of this, we will soon need to have additional data centre space in order to satisfy the ever growing storage and computational needs.



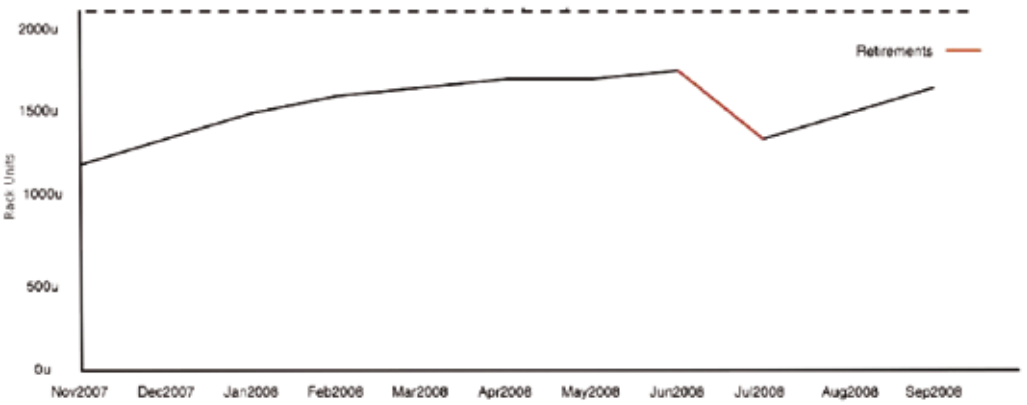


Figure 1 (top left). *CPU cores;* ***Figure 2 (top right).*** *Disk space;* ***Figure 3 (bottom).*** *Data centre space.*

Petteri Jokinen

MSc CS 1990, Helsinki University.
At EMBL-EBI since 1996.

Team Members

Server and Networking team
Jonathan Barker
Kimmo Kallio
Gavin Kelman
Manuela Menchi
Pravin Patel
Radoslaw Ryckowski

Desktop team
William Barber
Karen Briggs
Andy Cafferkey
Kieren Johnson*
John Livingstone

Software Engineer
Ville Silventoinen

Technical Administrator
Carolina Bejar

** Indicates part of the year only*

NETWORKING

Our primary internet connection is via Cambridge and we have an active backup link to London. We have ordered a new 10GB/s connection to London with a faster site router for the new link. The existing Cambridge link (1GB/s) will become a backup link once the new connection is finalised. During this year we will also undertake major work in our internal network enabling us to increase the speed of both internal and external access.

DATABASE SERVER INFRASTRUCTURE

We have configured a centralised scalable MySQL server farm that can be used by all of the EBI teams. We have introduced Linux-based Oracle servers (both clustered and standalone) that can be deployed automatically using our centralised system management software.

NEW PROJECTS

EBI has obtained some exiting new projects. Examples of these are the Trace Archive and BioFocus drug discovery data. These projects have needed a significant amount of systems resources. The Trace Archive project in particular will be pushing the limits of any storage architecture and is thus a major systems challenge in the coming years.

ALPHASERVER RETIREMENT

We have retired almost all of our Alphaservers. They and their ancestors deserve special thanks for serving EBI users so well during the past fourteen years. Sadly they are no longer manufactured and it was time to move on to Linux (mostly) and Solaris (at a lesser scale).

MACINTOSH SUPPORT

For the first time in EBI history we are now officially supporting Mac OSX machines. This is possible because we have recruited a Mac System Administrator, who will be able to start building the infrastructure that is needed.

SUMMARY

This year we have been busier than ever before. There has been a large quantity of equipment to install and maintain as well as an ever increasing number of people to support. We are trying very hard to deploy the best solutions. We continue to put a lot of emphasis on automation. This will result in a consistent and maintainable system that can be rebuilt from scratch with minimum human intervention. We are doing our best to deploy the densest possible solutions. In spite of that we anticipate that we will need considerably more data centre space allocated to us in the future.

Section 2

Services in 2008

External Services Team Scientific Report

Rodrigo Lopez

Vet. Med Degree 1984, Oslo Vet. Hoyskole and NASAS Cand. Scient. Molecular Toxicology and Informatics 1987, University of Oslo. At EMBL-EBI since 1995.

Team Members

Software Engineers
Mickael Goujon
Alberto Labarga*
Robert Langlois*
Hamish McWilliam
Teresa Mivar*
Eric Nzuobontane*
Franck Valentin

Web Developer
Asif Kibria
Thomas Laurent
Gulam Patel
Stephen Robinson
Brendan Vaughan

Web Systems Administrator
Jenny Martin

Visitors
Weizong Li
Menaka Narayanasami

** Indicates part of the year only*

INTRODUCTION

During the last six years the External Services (ES) team has been in charge of the development, management and maintenance of web and FTP resources at EMBL-EBI. These are in high demand as more and newer types of data are archived and distributed from the EBI's databases. The team is also in charge of global search and retrieval services that include the EB-eye search engine and SRS (SRS especially for the European Nucleotide Archive – ENA). It is also responsible for core sequence search and analysis services such as BLAST, FASTA, InterProScan, as well as several mainstream multiple sequence alignment (MSA) services such as ClustalW2, MAFFT, T-Coffee, etc. Today all these services have SOAP/REST web services interfaces, which allows programmatic and systematic use. In parallel to the running of these services, the team is responsible for maintenance, troubleshooting and the provision of guidelines and user support for EBI developers as well as general users. In this context, various members of the team have been heavily involved in training and conference activities, participation at local and international workshops and individual user engagement.

WEB PORTALS

The team is responsible for the main EBI web portal as well as several Wellcome Trust, BBSRC and EU-funded projects websites (amongst others), including: 1000 Genomes Project, 2can, BioCatalogue, BioSapiens, European Genotype Archive (EGA), ELIXIR, EMBRACE, ENFIN, FELICS, 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 training and support websites (2can) in collaboration with the Outreach and Training team and the general EBI staff.

Running web activities involves being able to quickly react to user requirements and maintain a dynamic and highly efficient presence. During 2007–2008, the web administrators have been busy implementing social networking tools that adhere to Web 2.0 standards. These include support for user-generated content, blogs and shared contents (such as calendars and bookings). The security of these systems is paramount and the team has the responsibility of overseeing activity and reacting to the latest threats. They are constantly upgrading systems and maintaining very high standards to ensure we can protect users against web-based attacks, such as phishing and denial-of-service attacks.

The World Wide Web is driven by so-called social technologies (ST) today. These are tools which have been designed as modules or plugins that allow a high degree of customisation between the different portals and encourage users to share content. The team has developed a Content Management System (CMS) which is compatible with current STs and is compliant with the latest W3C recommendations. Use of this CMS translates into higher degrees of freedom between users and reduced human overheads when it comes to maintenance and administration.

Monitoring and supporting use of EBI services

Monitoring and reporting the activity of the 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 powerful RDBMS (MySQL) as the back-end, a complex web and FTP logs analysis infrastructure, and an easy-to-use front-end for browsing and reviewing reports. To run this system efficiently and keep track of services and changes, the group has also developed a services database. This allows us to have an extensive inventory of services maintained by the various groups at the EBI. The information recorded ranges from URL entry points, the physical location of application servers as well as details about who is responsible and source of funding. There are more than 100 services running on a farm comprising more than 20 machines. Furthermore, this database records the whole history of a given service and is updated daily with the data produced by AWSTATS. This system allows for custom queries to be made regarding any service at any specific point in time. The database is also used to maintain and create the required configuration files needed to generate the AWSTATS reports. All the EBI web statistics included in this report were generated using this system. Tables 1, 2 and 3 show web and FTP activity during the period July 2007–June 2008. It is worth mentioning here that these tables represent organic traffic and do not include traffic generated by robots or programmatic web services. They also do not include traffic going to www.ensembl.org.

Publications

2007
Larkin, M.A., *et al*. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948

2008
Chica, C., *et al*. (2008). A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, 9, Article 229

Cochrane, G., *et al*. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 36, D5-D12

Robinson, J., *et al*. (2008). The IMGT/HLA database. *Nucleic Acids Res*., 37 (Database issue), D1013-D1017 Epub ahead of print

Month	Visitors	Visits	Pages	Hits	Bandwidth (GB)
Jul-07	181,004	422,063	18,453,129	53,369,081	447.99
Aug-07	166,760	395,829	13,227,267	40,712,708	386.05
Sep-07	198,017	442,792	18,152,655	47,670,418	546.64
Oct-07	251,825	553,895	21,479,107	59,007,347	567.16
Nov-07	253,310	549,458	25,530,239	62,810,863	555.87
Dec-07	197,710	419,379	17,202,927	45,309,902	463.20
Jan-08	216,957	491,423	18,744,934	54,908,005	604.04
Feb-08	230,748	500,632	15,976,198	54,088,076	590.54
Mar-08	208,240	431,219	17,393,188	50,947,273	639.61
Apr-08	270,532	584,133	26,260,786	68,228,110	816.47
May-08	286,067	588,642	35,195,639	77,791,467	937.10
Jun-08	307,783	602,652	19,107,990	57,234,706	709.96
Total	**2,768,953**	**5,982,117**	**246,724,059**	**672,077,956**	**7,264.63**

Table 1. *Viewed traffic on the main EBI web portal during the period July 2007–June 2008. Figures do not include robots, which comprise on average 15% of all traffic, or traffic from www.ensembl.org.*

Month	Visitors	Visits	Pages	Hits	Bandwidth (GB)
Jul-07	13,632	23,037	317,652	688,830	22.23
Aug-07	35,936	68,477	254,809	2,083,038	44.02
Sep-07	37,315	81,944	893,625	2,728,247	79.57
Oct-07	51,092	107,767	1,071,187	3,588,712	57.18
Nov-07	52,516	103,754	375,924	3,046,276	50.22
Dec-07	41,971	81,115	284,000	2,163,192	33.55
Jan-08	58,137	114,397	415,409	2,910,387	50.68
Feb-08	54,709	107,515	413,269	3,021,659	77.38
Mar-08	51,010	96,389	454,745	2,716,900	187.76
Apr-08	64,158	124,136	786,606	4,339,986	74.10
May-08	74,508	131,252	1,565,066	6,421,843	91.70
Jun-08	98,063	164,000	1,177,927	5,464,950	99.49
Total	**633,047**	**1,203,783**	**8,010,219**	**39,174,020**	**867.88**

Table 2. *Viewed traffic from www4.ebi.ac.uk portal (all .org or .info traffic not logged under www.ebi.ac.uk) during the period July 2007–June 2008. Figures do not include robots.*

Month	Visitors	Visits	Hits	Bandwidth (GB)
Jul-07	6,363	15,623	2,275,019	4,788.30
Aug-07	5,922	16,970	3,398,491	5,682.50
Sep-07	6,387	16,018	2,663,309	6,772.20
Oct-07	7,222	16,598	2,648,572	4,662.35
Nov-07	8,747	18,804	3,078,933	5,178.24
Dec-07	6,705	15,982	3,375,026	8,551.94
Jan-08	7,185	16,385	2,364,426	5,465.23
Feb-08	8,085	17,973	1,973,573	5,929.21
Mar-08	8,390	18,398	2,614,906	8,305.64
Apr-08	8,461	18,263	2,901,781	5,862.89
May-08	9,400	20,585	2,792,208	6,289.46
Jun-08	9,910	20,374	2,941,898	8,494.85
Total	**92,777**	**211,973**	**33,028,142**	**75,982.81**

Table 3. *FTP traffic during the period July 2007–June 2008. Figures do not include ensembl.org.*

As explained earlier, the present 'web farm' comprises some 20 individual machines arranged in pools. Each pool has at least two nodes for redundancy and fail-over. The management of these pools currently uses a pair of load balancers. Activities undertaken by the group in this area include providing support, advice and troubleshooting expertise to the EBI developer community. 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 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 done by building on experience with earlier and existing production systems and by incorporating greater flexibility when software upgrades to Tomcat, Java and database driver infrastructure 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.

There is a long list of service frameworks that are currently supported by the team (see below). During 2007–2008, more have been added as part of the supported infrastructure and include BioMarts, Ensembl and the new UniProt unified website.

- Dasty2: an AJAX-based web client for visualising protein sequence feature information using DAS, developed by the proteomics services group;
- UniProt unified website, UniProt services, UniProt remoting API;
- ELM: a web service for the conservation scoring of predicted Eucaryotic Linear Motifs, developed jointly with Toby Gibson's group at EMBL;
- Iprservice (InterProScan): a web service for the automatic annotation of protein sequences developed jointly with the UniProt team.

New services deployment (infrastructure, testing, etc.):

- Pride BioMart: new BioMart installation to support PRIDE data;
- Uniprot DAS: new Java DAS server based on Kraken;
- ASTD: new web application for the ASTD project;
- Webin: support in the migration to the new web infrastructure;
- HUGO gene nomenclature web application was moved to the EBI from UCL (www.genenames.org).

New SOAP web services developed by other groups but managed by the External Services team:

- IntAct web services;
- Protein Identifier Cross-Reference Service (PICR);
- CiteXplore web services;
- MIRIAM/SBO web services.

WEB APPLICATIONS AND WEB SERVICES

There is constant revision of application interfaces and algorithm code. The group strives to maintain contact with various authors and regularly engages with users to provide information about potential bugs and fixes. We are honoured to count amongst our collaborators Professor Bill Pearson from the University of Virginia (FASTA), Professor Desmond Higgins from University College Dublin (ClustalW2), Professor Geoff Barton from the University of Dundee (SCANPS and JALVIEW), Dr Cedric Notredam from CRG in Barcelona (T-Coffee), Ewan Birney from the EBI (WISE2), staff at the NCBI (BLAST), Dr Warren Gish from the University of Washington (WUBLAST) and many others not mentioned here.

Some of the latest developments include a much enhanced version of the FASTA services. The latest features available include query sequence and library sequence annotations as well as the ability to search using a subset of a library (the latter feature is to be deployed shortly). We have also implemented a very rigorous protein sequence search, PSI-SEARCH, which combines the Smith-

Waterman implementation in the FASTA package known as SSEARCH with NCBI's BLASTPGP. This tool is ideally suited to identify and explore distant relationships. The PSI-SEARCH interface gives the user greater control of the iterative search process. One can run multiple rounds of PSI-SEARCH and select hits in each iteration to build the position-specific scoring matrix (PSSM). The first round of PSI-SEARCH is a standard SSEARCH run. In the second and further iterations, the program builds a PSSM and a checkpoint file from a multiple alignment of the sequences by using NCBI's BLASTPGP and searches the database by using SSEARCH with the checkpoint file.

Another improvement that took place during the reporting period is the launch of ClustalW2. This is not an entirely new version of the program, but one that improves significantly over earlier versions, both in terms of accuracy (with iterative alignment) and ease of maintenance as the code base has been completely refactored.

The EBI's search engine, the EB-eye, has been subjected to a number of modifications and improvements. EB-eye has seen its utilisation double during the reporting period and as a consequence of this, maintaining the speed of indexing and general performance has been the main focus of work. New versions of the underlying Lucene core and major code optimisations now allow us to rebuild the system from a catastrophic failure in less than twelve hours. The main goal is to ensure that the system is up to date and in synch with the underlying databases within one hour following failure.

SRS continues to play an important role, especially since it is the main service currently employed by the European Nucleotide Archive. The SRS server at the EBI will be migrated to SRS8.x as soon as the authors resolve some backward compatibility issues and address some performance problems. The present server receives more than three million queries per month and hosts more than 300 libraries. It continues to represent the fail-over service for many of the core databases held at the EBI.

The adoption of web services and XML technologies, which started during 2002, translates to better and more robust services today. Users now have systematic access to our analytical tools. Our web services present a uniform API for a wide variety of bioinformatic applications that count more than 250 individual algorithms. Access to these represents slightly over 900,000 compute jobs per month (during the reporting period and over one million at the time of writing). Table 4 shows the monthly breakdown of use of these services using browsers, email and the use of the web services API.

In the context of 'finding stuff', the group is involved in the development of browsable catalogues that will enable users to find web services, tools and databases. These projects are run in collaboration with members of the EMBRACE project (the EMBRACE service registry) and the staff at the University of Manchester (BioCatalogue). The projects are financed by the EU and the BBSRC respectively. The aims of these catalogues are to produce a much-needed 'yellow pages' for life sciences resources and their outcome will be linked with mainstream search engines such as EB-eye and Google.

Month	Web	Email	API	Total
Jul-07	427,579	7,367	292,084	727,030
Aug-07	390,423	7,649	193,288	591,360
Sep-07	377,539	8,168	403,836	789,543
Oct-07	533,417	7,674	239,438	780,529
Nov-07	504,842	7,975	604,660	1,117,477
Dec-07	344,679	6,014	326,801	677,494
Jan-08	436,823	7,215	345,745	789,783
Feb-08	471,026	7,795	329,058	807,879
Mar-08	506,174	7,802	580,379	1,094,355
Apr-08	496,051	7,684	618,769	1,122,504
May-08	468,861	7,487	732,426	1,208,774
Jun-08	450,500	6,676	666,555	1,123,731
Total	**5,407,914**	**89,506**	**5,333,039**	**10,830,459**

Table 4. *Access to EBI web services from July 2007–June 2008. Web access signifies browser-based navigation and submission of jobs. Email represents results delivered to the user using e-mail. API represents access using the EBI web services API.*

Work in relation to sequence patents is continuing as part of the team's collaboration with the EPO. The latest developments in this context include non-redundant sets of data, level 1 and level 2 equivalencies, and additional patent sequence databases from both Korea and China. This work leads to a much clearer understanding of when and where a patent was first filed, but importantly, what is actually covered by a patent or by its application. This work is of particular importance to patent examiners currently using EBI services, which include the EPO (based in The Hague, Netherlands and Munich, Germany), Spain, UK, USPTO, KIPO and JPO.

COURSES AND CONFERENCES

Various members of the group have been engaged in training activities during 2007–2008, both at the EBI and elsewhere. These are listed below:

22 September 2008	ECCB tutorial: Interoperability of bioinformatics software and databases (Rodrigo Lopez)
12 September 2008	FELICS Roadshow: Munich (Hamish McWilliam)
8–11 September 2008	EBI hands-on training: Programmatic access in Perl: web services & workflows (Rodrigo Lopez and Hamish McWilliam)
5 September 2008	FELICS Roadshow: The Hague (Hamish McWilliam)
28 August 2008	EMBRACE: Web Services in Systems Biology (Hamish McWilliam)
30 July – 1 August 2008	III Congreso Colombiano de Biotecnología (Rodrigo Lopez)
28–31 July 2008	EBI hands-on training: Programmatic access of Proteomics Resources (Hamish McWilliam)
18–23 July 2008	ISMB 2008 (Rodrigo Lopez, Thomas Laurent, Franck Valentin and Hamish McWilliam)
18 June 2008	EBI/Sanger: Sequence Similarity Search for Curators
9–11 June 2008	EBI hands-on training: Patterns, Similarities and Differences in Biological Data (Hamish McWilliam)
19–22 May 2008	CCGrid 2008 (Rodrigo Lopez
18–20 May 2008	EMBRACE: Advanced Protein Domain Analysis (Hamish McWilliam)
12–16 May 2008	BioSapiens: 8th BioSapiens School
22–24 February 2008	E-MeP: Advanced Training Workshop in Bioinformatics of Membrane Proteins (Hamish McWilliam)
6–8 February 2008	EMBRACE: Workshop on Client Side Scripting for Web Services (Franck Valentin, Teresa Mijar and Rodrigo Lopez)

The Activities of the PANDA Group

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 EMBL Nucleotide Sequence Database, the Ensembl genome browser, the UniProt protein 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 Birney and Apweiler groups have complementary research components. Substantial training and outreach efforts are also part of the PANDA group's activities

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 37. The activities of the Vertebrate Genomics, Ensembl Genomes, Proteomics Services, InterPro, and the Chemoinformatics and Metabolism teams are described in separate reports by their team leaders Paul Flicek, Paul Kersey, Henning Hermjakob, Sarah Hunter and Christoph Steinbeck, respectively.

The main achievements of the PANDA group in 2008 have been:

- handling an ever-growing amount of nucleotide and protein data;
- releasing the first draft of the complete human proteome in UniProtKB/Swiss-Prot;
- launching the European Genotype Archive (EGA) as a central and permanent repository for genetic data;
- opening the next-generation sequencing data archive for submissions. Submissions of data from next-generation sequencing platforms, such as those of 454, Illumina and ABI SOLiD, are now accepted at the EBI as a new service under the European Nucleotide Archive (ENA) that will provide utility for those generating and using data for applications such as genome assembly, resequencing for polymorphism analysis, gene expression, epigenomics and others;
- preparing to provide open access to large-scale drug discovery data. The Wellcome Trust has awarded a grant to the EBI 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 publicly listed company Galapagos NV to the public domain. This led to the formation of a chemogenomics team within PANDA under the leadership of John Overington.

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 2008 we have seen a striking growth in two areas – the use of these new machines for surveying natural variation in populations, in particular the human population (see the report from Paul Flicek, page 59), and the more routine determination of genotypes from large disease cohorts, leading to associations between genetics and disease (also presented in Paul Flicek's report). We expect to see the impact of these technologies become more prevelant across all domains of life, and have anticipated this growth in the forthcoming years by founding Ensembl Genomes, a high-quality, community-led genomic information resource for non-vertebrate species (presented in Paul Kersey's report, page 65). The shift in technology and the repositioning of genomic information as a key organisation 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 below.

The PANDA Nucleotides group has three main sections with Ewan Birney providing strategic oversight across all branches. The sections are: Vertebrate Genomics, which includes Ensembl (under the leadership of Paul Flicek, see page 59); Ensembl Genomes (under the leadership of Paul Kersey, see page 65) and the European Nucleotide Archive, coordinated by Guy Cochrane, which is presented in

Rolf Apweiler

PhD 1994, University of Heidelberg.
At EMBL since 1994.

Ewan Birney

PhD 2000, Sanger Institute.
At EMBL–EBI since 2000.

Team Members

Joint Team Leader PANDA Group (Proteins)
Rolf Apweiler

Joint Team Leader PANDA Group (Nucleotides)
Ewan Birney

Team Leaders
Paul Flicek
Henning Hermjakob
Sarah Hunter
Paul Kersey
John Overington*
Christoph Steinbeck

Group Coordinators
Elspeth Bruford
Guy Cochrane
Lennart Martens
Maria-Jesus Martin
Claire O'Donovan
Glenn Proctor
Esther Schmidt

Project Coordinators
Fiona Cunningham*
Paula de Matos
Emily Dimmer
Xosé M. Fernandez
Javier Herrero-Sanchez
Pascal Kahlem
Daniel Lawson
Rasko Leinonen
David Lonsdale
Michele Magrane
Manuela Pruess
Robert Vaughan

Senior Scientific Database Curators
Paul Browne

Nadeem Faruque
John Stephen Garavelli
Jennifer McDowall
Sandra Orchard

Scientific Database Curators
Ruth Akhtar
Yasmin Alam-Faruque
Sumit Bhattacharraya*
Wei Mun Chan
Louise Daugherty
Catherine Derow*
Ruth Eberhardt
Marcus Ennis
Rebecca Foulger
Phani Garapati
Richard Gibson
Susan Gordon*
Christopher Hunter
Rachael Huntley
Julius Jacobsen
Jyoti Khadake
Bijay Jassal
Kati Laiho
Duncan Legge
Gaurab Mukherjee
Petra ten Hoopen
Ruth Seal*
Inma Spiteri*
David Thorneycroft
Matt Wright

Bioinformaticians
Yuan Chen
David Croft
Martin Hammond
Gemma Hoad
Craig McAnulla
Damian Smedley

Senior Software Engineers
Richard Cote
Alexander Fedotov
Alan Horne
Phil Jones
Samuel Kerrien
John Maslen
Samuel Patient
Antony Quinn
Mark Rijnbeek*
Peter Sterk
Manjula Thimma*

Software Engineers
Premanand Achuthan
Rafael Alcantara Martin
Ricardo Antunes*
Bruno Aranda
Daniel Barrell
Benoit Ballester
Kathryn Beal
Benoit Bely*
David Binns
Lawrence Bower
Mario Caccamo*
Laura Clarke*
Manuel Corpas*
Ujjwal Das
Bernard de Bono
Fehmi Demiralp
Stephen Fitzgerald
Renato Golin
Leo Gordon*
Stefan Graef
Matthias Haimel
Janna Hastings
Jonathan Hinton*
Zamin Iqbal*

this section. In addition, the HUGO Gene Nomenclature Committee (HGNC), a smaller group coordinated by Elspeth Bruford, is part of PANDA Nucleotides and presented here. 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 upon. This dataset is coordinated worldwide by virtue of the International Nucleotide Sequence Database Collaboration (INSDC), forming a single, worldwide 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.

The rest of this section details the two areas (European Nucleotide Archive and HGNC) which directly report to Ewan Birney whereas the two other areas (Vertebrate Genomics, including Ensembl, and Ensembl Genomes) are described in sections by their own team leaders. Finally, Ewan Birney's research group is presented on page 56.

EMBL NUCLEOTIDE ARCHIVE

Ruth Akhtar, Sumit Bhattacharyya, Lawrence Bower, Paul Browne, Guy Cochrane, Fehmi Demiralp, Nadeem Faruque, Richard Gibson, Gemma Hoad, Christopher Hunter, Mikyung Jang, Szilveszter Juhos, Rasko Leinonen, Quan Lin, Rodrigo Lopez, Dariusz Lorenc, Hamish McWilliam, Gaurab Mukherjee, Menaka Narayanasami, Sheila Plaister, Rajesh Radhakrishnan, Stephen Robinson, Siamak Sobhany, Petra ten Hoopen, Robert Vaughan, Vadim Zalunin, Weimin Zhu

Steven Leonard, James Bonfield, Tim Hubbard (Wellcome Trust Sanger Centre)

The European Nucleotide Archive 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. The ENA comprises the EMBL Nucleotide Sequence Database, the Ensembl Trace Archive for capillary sequence traces and the newly launched repository for sequencing data from ultra-high throughput sequencing (UHTS) machines. Databases of the ENA achieve comprehensive coverage through partnership with the 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, has been underway for over a quarter of a century and now serves as a model for data sharing in the life sciences.

Adjusting our overall view of nucleotide sequence archiving, we have abstracted somewhat from the underlying legacy infrastructure, such that sequencing information is classed as:

- '**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 read, assembly and annotation information is data relating to the provenance and treatment of biological samples used for sequencing. In contrast to the other principal PANDA databases, all core information in the ENA is provided and updated solely by submitters. Where possible, data in ENA-Annotation, ENA-Assembly and ENA-Reads are connected in a single integrated system.

In 2008, we have witnessed continued exponential growth in ENA-Annotation; at the time of writing, ENA contains 143 million ENA–Annotation records, covering 233 billion bases (see Figure 1). However, the most striking data growth has been in ENA-Reads, where we have recorded 1.8 billion bases of capillary reads with 10TB of next-generation sequence data. In all, 400,000 different

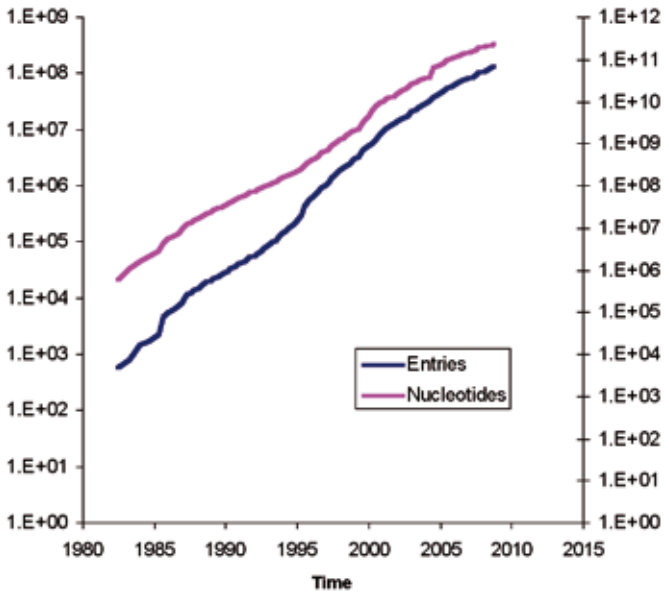


Figure 1. *Data growth for assembled/annotated sequences in ENA-Annotation.*

taxonomic nodes are connected to sequence, over 200,000 published papers are explicitly cross-referenced in ENA records and ENA maintains 99 million cross-references to objects in external resources. In 2008, notable high-volume datasets include:

- the raw sequencing data from the genomes of 83 individuals from the ongoing human 1000 Genomes Project (approximately 460GB; ERA000013-ERA000026);
- all of the underlying data for an extensive genome variation and evolution study including 17 *Salmonella* Typhi isolates (WGS accessions starting CAAQ-CAAZ and ENA-Reads accession ERA000001);
- two newly sequenced genomes from a trio of isolates, including an important drug-resistant nosocomial isolate of *Acinetobacter baumannii (ENA projects 13001 and 28921).*

Submissions and data access at ENA

The ENA provides a variety of submission services, sensitive to the scale of data and bioinformatics expertise available to submitters; a central submission portal is available at http://www.ebi.ac.uk/embl/Submission/webin.html. Data are presented through a variety of means, including search (sequence similarity or text) and FTP (see http://www.ebi.ac.uk/embl/Access/index.html).

Ultra-high throughput sequencing data

UHTS technologies yield unprecedented data volumes, comparatively short sequence reads and low per-read cost. These characteristics expand nucleotide sequencing to a broad range of applications, going far beyond its conventional use in the determination of genome and transcriptome sequences for the purposes of assembly and subsequent functional annotation. These new uses include expression analysis, resequencing for polymorphism discovery, epigenomics and gene discovery. As a response to this development, the ENA has established a new petabyte-scale archive for ultra-high throughput sequencing data under the support of the Wellcome Trust, driven by the 1000 Genomes human resequencing project (see Paul Flicek's report, page 59).

Our design principle for the ENA has involved a radical re-think of the model for storing raw sequence data. The archive uses a hybrid file-database system that maximises data compression and ease of access to metadata (sample, experiment and run details). We have established a shared metadata object and read accessioning system with the Short Read Archive at NCBI and expect to attain comprehensive coverage by regular exchange of data within the coming months. We have deployed submission pipelines for UHTS data for one-off submissions from smaller research groups and for sustained data flow from the large sequencing centres. At the time of writing, all ENA-Reads data can be accessed through the FTP site (ftp://ftp.era.ebi.ac.uk/). We expect to be able to soon replace this service with programmatic access and a more flexible search and browse tool.

Mikyung Jang
Andrew Jenkinson
Nathan Johnson
Szilveszter Juhos*
Andreas Kahari
Arek Kasprzyk*
Damian Keefe
Stephen Keenan*
Arnaud Kerhornou
Rhoda Kinsella*
Michael Kleen
Gautier Koscielny
Stefan Kuhn*
Eugene Kulesha
Vasudev Kumanduri*
Davang Lakhani
Ilkka Lappalainen*
Vincent le Texier*
Quan Lin
Ian Longden
Dariusz Lorenc*
Michael Lush
Jie Luo
Karyn Megy
Patricia Monteiro*
Mylrajan Muthusami*
John F. O´Rourke
Rajesh Radhakrishnan*
Florian Reisinger
Daniel Rios
Emilio Salazar Donate*
Andrey Sitnov*
Guy Slater
Siamak Sobhany
Dan Staines
Gilleain Torrance*
Albert Vilella
Juan A. Vizcaino
Steven Wilder*
Phil Wilkinson
Andy Yates
Vadim Zalunin*

Helpdesk Officers
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*
Markus Fritz*
Michael Hoffman*
Garth Ilsley
Alison Meynert
Michael Mueller
Dace Ruklisa
Daniel Zerbino

Students
Irina Armean*
Dominik Grimm*
Andreas Hoelzlwimmer*
Wolfgang Kluge*
Michael Menden*
Nadin Neuhauser*
Omar Pera Mira*
Jonathan Ramseier*
Avazeh Tashakkori Ghanbarian*

Visitors
Kirill Degtyarenko
Ian Dunham
Alex Mitchell
Augusto Rendon
Will Spooner
Eleanor Stanley
Matthieu Visser

** Indicates part of the year only*

Publications

Apweiler

2007

Asara, J.M., *et al*. (2007). Interpreting sequences from mastodon and T. rex [5]. *Science*, 317, 1324-1325

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

Eisenacher, M., *et al*. (2007). Proteomics data collection - The 1st ProDaC workshop. *Proteomics*, 7, 3034-3037

Field, D., *et al*. (2007). eGenomics: Cataloguing our complete genome collection III. *Comp. Funct. Genomics*, 2007, 6, 363-368

Flikka, K., *et al*. (2007). Implementation and application of a versatile clustering tool for tandem mass spectrometry data. *Proteomics*, 7, 3245-3258

Frigerio, G., *et al*. (2007). Two human ARFGAPs associated with COP-I-coated vesicles. *Traffic*, 8, 1644-1655

Holland, P.W.H., *et al*. (2007). Classification and nomenclature of all human homeobox genes. *BMC Biol*., 5, Article 47

Further key developments in 2008

Integration with genomes: organisation of nucleotide information around sequenced reference genomes is a founding principle of PANDA Nucleotides. It is notable that a high proportion of data in ENA-Annotation (50% of entries, 70% of nucleotides) are derived from organisms that have completely sequenced reference genomes available in Ensembl and Ensembl Genomes/Genome Reviews. We have established a mapping procedure that links entries in ENA-Annotation and Ensembl objects. The first deployment has provided cross-references from archival transcripts that have been used as supporting evidence for Ensembl gene building.

Feature table changes: our continued development of the INSDC Feature Table Definitions (http://www.ebi.ac.uk/embl/WebFeat/index.html) has focused on the removal of information that can be unambiguously calculated (e.g. we have removed /cons_splice) and on simplification (e.g. we have merged satellite, repeat_unit and repeat_region).

Standards and formats activities: we have been involved in the development of the Minimum Information about a Genome Submission (MiGS) standard (Field *et al.*, 2008), the emerging Minimum Information about a Sequence-based Experiment (MINSEQE) and in the development of the short read metadata XML format.

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.

In the past year HGNC publications have included an update paper in the database issue of *Nucleic Acids Research* describing improvements to the HGNC database in the last two years, and six publications on the nomenclature of specific gene families. As well as attending eight international conferences, HGNC staff also took part in the first ever three-way meeting between the HGNC, the Mouse Genomic Nomenclature Committee, and the Rat Genomic and Nomenclature Committee (RGNC), at the Rat Genome Database, Medical Council of Wisconsin, Milwaukee, in June 2008. Following the HGNC's relocation to the EBI in September 2007, the original team of Elspeth Bruford (Group Coordinator), Matthew Wright (Gene Nomenclature Advisor) and Michael Lush (Bioinformatician) was augmented in May 2008 by two new gene nomenclature advisors, Susan Gordon and Ruth Seal. In 2008 we also constituted a new 20-person International Advisory Committee for the HGNC, including the introduction of an executive IAC comprising five key individuals who will meet biennially to discuss future plans for the HGNC (http://www.genenames.org/IAC.html).

Further developments have included:

- curating and making public the locus type (gene with protein product, gene with no protein product, pseudogene, phenotype etc.) for every entry in our database;
- updating the HGNC Comparison of Orthology Predictions search tool, HCOP, to enable users to compare orthologues predicted for human genes in seven other genomes (chimp, mouse, rat, dog, chicken, zebrafish and fruitfly) and providing a total of eleven independent orthology datasets for comparison;
- developing a BioMart tool which will provide users with standalone data mining of the HGNC dataset and will be easily linked to other BioMart instances, including Ensembl and Reactome.

Considerable efforts have been made to increase links from our database to external resources, which have included the Wellcome Trust Sanger Institute's COSMIC (Catalogue Of Somatic Mutations In Cancer) database; the Orphanet portal for rare diseases; the new WikiGenes website which allows author attributed editing; the Consensus CDS (CCDS) project, a collaborative effort to identify a core set of human protein-coding regions that are consistently annotated and of high quality; as well as adding locus-specific database links to over 600 gene entries. Gene naming has also been focused on genes identified by the CCDS project and to date we have named over 98% of the approximate total of 17,000 genes in the current CCDS set. The total of approved gene symbols presently stands at 26,136 (as of 15 October 2008), an increase of over 1,500 in the past year. Some of this increase is also due to continued efforts to name non-protein coding RNA genes, including instituting a new ncRNA nomenclature for non-protein coding RNAs of unknown function and assigning names to all the transfer RNAs in the current genome build. We have also been strengthening our collaboration with UniProt to ensure that links between our genes and their curated protein entries are accurate and updated. A jamboree in early June involving the HGNC, UniProt curators from the Swiss Institute of Bioinformatics (SIB) and EBI, and from Havana and Ensembl, enabled us to examine problematic cases of gene names, putative proteins for withdrawal, and possible candidates for naming and renaming. This has already led to strengthened links between all the groups involved; we continue to feed back on our data to UniProt curators, as well as prioritising the naming of protein-encoding genes that have been annotated in UniProtKB/Swiss-Prot.

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 79). 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 presented in Henning Hermjakob's report, page 73.

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. Finally, Rolf Apweiler's research group is presented on page 55.

THE UNIVERSAL PROTEIN RESOURCE

Yasmin Alam-Faruque, Ricardo Antunes, Daniel Barrell, 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, Jie Luo, Quan Lin, Michele Magrane, Maria Jesus Martin, Patricia Monteiro, Mylrajan Muthusamy, Claire O'Donovan, John F. O'Rourke, Sandra Orchard, Samuel Patient, Emilio Salazar Donate, Andrey Sitnov, Eleanor Stanley

The Universal Protein Resource (UniProt; http://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, 2007). 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.

Jones, A.R., *et al*. (2007). The Functional Genomics Experiment model (FuGE): An extensible framework for standards in functional genomics. *Nat. Biotechnol*., 25, 1127-1133

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

Martens, L., *et al*. (2007). Human proteome organization proteomics standards initiative: Data standardization, a view on developments and policy. *Molecular and Cellular Proteomics*, 6, 1666

Orchard, S. (2007). Proteomics: From technology development to biomarker applications: HUPO 6th Annual World Congress. *Expert Rev. Proteomics*, 4, 709-710

Orchard, S., *et al*. (2007). Submit your interaction data the IMEx way: A step by step guide to trouble-free deposition. *Proteomics - Practical Proteomics*, 2, 28-34

Orchard, S., *et al*. (2007). Five years of progress in the standardization of proteomics data 4 th annual spring workshop of the HUPO-proteomics standards initiative - April 23-25, 2007. Ecole Nationale Supérieure (ENS), Lyon, France. *Proteomics*, 7, 3436-3440

The UniProt Consortium, (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res*., 35: D193-D197

2008
Angiuoli, S.V., *et al*. (2008). Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. *OMICS A Journal of Integrative Biology*, 12, 137-141

Apweiler R. & Mueller M. (2008). Annotating the human proteome: From establishing a parts list to a tool for target identification. In *Cancer Proteomics: From Bench to Bedside*. Daoud S.S. (ed), 211-235, Humana Press Inc., NJ, USA

Braconi Quintaje, S. & Orchard, S. (2008). Completion of the annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: One small step in manual annotation, one giant step for full comprehension of genomes. *Mol. Cell. Proteomics*, 7, 1409-1419

Chiang, T., *et al*. (2008). Rintact: Enabling computational analysis of molecular interaction data from the IntAct repository. *Bioinformatics*, 24, 1100-1101

Cochrane, G., *et al*. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 36, D5-D12

Cole, C.G., *et al*. (2008). Finishing the finished human chromosome 22 sequence. *Genome Biol*., 9, Article R78

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

Degtyarenko, K., *et al*. (2008). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res*., 36, D344-D350

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 (UniProtKB), which represents the centrepiece of UniProt's activities, 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 14.2 (October 2008) consists of 6,932,724 entries (398,181 UniProtKB/Swiss-Prot entries and 6,932,724 UniProtKB/TrEMBL entries). A specimen UniProt report can be found at http://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 lines (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 (Camon *et al.*, 2004) 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 below). 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 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. With the rapid growth of sequence databases, there is an increasing need for reliable functional characterisation and annotation of newly predicted proteins.

For automatic annotation, various systems of standardised transfer of annotation from well-characterised proteins in UniProtKB/Swiss-Prot to non-annotated UniProtKB/TrEMBL entries have been implemented. One system, RuleBase, uses a semi-automatic approach, while Spearmint 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 Spearmint. InterPro (Mulder *et al.*, 2007) 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. The PIRSF (Natale *et al.,* 2004) and HAMAP (Gattiker *et al.,* 2003) rule systems will shortly be integrated into this pipeline to further improve the annotation of UniProtKB/TrEMBL entries.

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 102 external databases. One collaboration of interest is between UniProt and the National Center for Biotechnology Information (NCBI) to provide bi-directional cross-references between Entrez Gene/RefSeq and UniProtKB within each tri-weekly cycle.

A document listing all databases cross-referenced in UniProtKB is available (http://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 (VARSPLIC), 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 (or genes if the translation from different genes in the genome leads to indistin-

Dimmer, E.C., *et al*. (2008). The Gene Ontology - Providing a Functional Role in Proteomic Studies. *Proteomics.* Epub ahead of print

Dunn, M.J., *et al*. (2008). EuPA achieves visibility - An activity report on the first three years. *Journal of Proteomics*, 71, 11-18

Eisenacher, M., *et al*. (2008). Proteomics data collection - 2nd ProDaC workshop: 5 October 2007, Seoul, Korea. *Proteomics*, 8, 1326-1330

Field, D., *et al*. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol*., 26, 541-547

Field, D., *et al*. (2008). Meeting report: The fourth Genomic Standards Consortium (GSC) workshop. *OMICS A Journal of Integrative Biology*, 12, 101-108

Field, D., *et al*. (2008). Meeting report: The fifth Genomic Standards Consortium (GSC) workshop. *OMICS A Journal of Integrative Biology*, 12, 109-113

Flicek, P., *et al*. (2008). Ensembl 2008. *Nucleic Acids Res*., 36, D707-D714

The Gene Ontology Consortium, (2008). The Gene Ontology project in 2008. *Nucleic Acids Res*., 36, D440-444

Hamacher, M., *et al*. (2008). The HUPO brain proteome project wish list - Summary of the 9th HUPO BPP workshop: 9-10 January 2008, Barbados. *Proteomics*, 8, 2160-2164

Han, R., *et al*. (2008). An efficient conformational sampling method for homology modeling. *Proteins: Structure, Function and Genetics*, 71, 175-188

Hermjakob, H. (2008). EBI proteomics services. In *Lecture Notes in Computer Science*, 207

Jenkinson, A.M., *et al*. (2008). Integrating biological data - The Distributed Annotation System. *BMC Bioinformatics*, 9, Article 53

Jones, P. & Cote, R. (2008). The PRIDE Proteomics Identifications Database: Data Submission, Query, and Dataset Comparison. *Methods Mol. Biol*., 484, 287-303

Jones, P., *et al*. (2008). PRIDE: new developments and new datasets. *Nucleic Acids Res*., 36, D878-883

Jungblut, P.R., *et al*. (2008). The speciation of the proteome. *Chemistry Central Journal*, 2, Article 16

Klie, S., *et al*. (2008). Analyzing large-scale proteomics projects with latent semantic indexing. *Journal of Proteome Research*, 7, 182-191

Lovering, R.C., *et al*. (2008). Cardiovascular GO annotation initiative year 1 report: Why cardiovascular GO? *Proteomics*, 8, 1950-1953

Martens, L., *et al*. (2008). Using the Proteomics Identifications Database (PRIDE). *Current protocols in bioinformatics,* Andreas D. Baxevanis *et al*. (editoral board), Chapter 13, 8

Martens, L., *et al*. (2008). Data standards and controlled vocabularies for proteomics. *Methods Mol. Biol*., 484, 279-286

Mathivanan, S., *et al*. (2008). Human Proteinpedia enables sharing of human protein data [4]. *Nat. Biotechnol*., 26, 164-167

Michaut, M., *et al*. (2008). InteroPORC: Automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24, 1625-1631

Mons, B., *et al*. (2008). Calling on a million minds for community annotation in WikiProteins. *Genome Biol*., 9, Article R89

guishable proteins). 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. So far, isoform identifiers have been introduced for splice isoforms. Splice isoforms may differ considerably from one another, with potentially less than 50% sequence similarity between isoforms. The freely available tool VARSPLIC 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. A new development in sequence production, namely, the availability of metagenomic data, has necessitated the creation of a separate database, 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 UniProtKB Sequence/Annotation Version database

Although UniProtKB entries are subject to changes affecting both the sequence and the annotation, only the most recent versions are currently preserved in the database. In response to user demand, we have created the UniProtKB Sequence/Annotation Version database (UniSave; www.ebi.ac.uk/uniprot/unisave), which is a comprehensive archive of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions (Leinonen *et al*., 2006). All updated and new UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entries are loaded into UniSave as part of the public three weekly UniProtKB releases. Unlike UniProtKB, which contains only the latest Swiss-Prot and TrEMBL entry versions, UniSave provides access to previous versions of these entries. This allows retrieval of 'historic' data, which is particularly important in the context of patent claims. We expect that the data in UniSave will double every ten to twelve months. UniSave also forms an integral part of the new unified UniProt website.

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 14.2 (October 2008) contained 17,500,509 unique sequences from 53,933,054 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, 16,865,496 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 http://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 (UniRef), UniRef100, UniRef90 and UniRef50, from UniProtKB 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 facilitates sequence merging in the UniProtKB (based on UniRef100) and allows faster sequence similarity searches (by using UniRef90 and UniRef50).

UniRef100 is based on all UniProtKB records and selected UniProt Archive records. The production of UniRef100 begins with the clustering of records by sequence identity irrespective of species. 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 http://www.uniprot.org/uniref/UniRef90_P57727.

GO ANNOTATION (GOA)

Daniel Barrell, David Binns, Emily Dimmer, Rachael Huntley

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 26,000 terms, distributed over three ontologies that describe the molecular functions, biological processes and locations of action of a gene product in a generic cell. The Gene Ontology Annotation (GOA) database (Camon *et al*., 2004; http://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 and the International Protein Index (IPI) 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 34 million GO annotations to over 4.5 million UniProtKB entries, covering more than 175,000 taxonomic groups. This represents a yearly increase of over 11.8 million GO annotations (a 53% increase) and a 25% increase in taxonomic coverage.

Manual GO annotations created by UniProtKB curators continue to be supplemented with the latest data from 20 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. This activity has continued with a new integration of experimentally-evidenced annotations for human protein subcellular localisation from the Human Protein Atlas project in September 2008. In addition, the Reactome and GOA groups have worked

Montecchi-Palazzi, L., *et al*. (2008). The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol*., 26, 864-866

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

Mueller, M., *et al*. (2008). Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. *Proteomics*, 8, 1138-1148

Mulder, N.J. & Apweiler, R. (2008). The InterPro database and tools for protein domain analysis. *Curr. Protoc. Bioinformatics*, Chapter 2, Unit 2.7

Mulder, N.J., *et al*. (2008). In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol. Biotechnol*., 38, 165-17

O'Donovan, C., *et al* (2008). The Universal Protein Resource. A Growing Online Protein Database. *BIOforum Europe* 12: 32-33

Olohan, L.A., *et al*. (2008). Detection of anoxia-responsive genes in cultured cells of the rainbow trout Oncorhynchus mykiss (Walbaum), using an optimized, genome-wide oligoarray. *Journal of Fish Biology*, 72, 2170-2186

Orchard, S. & Hermjakob, H. (2008). The HUPO proteomics standards initiative - Easing communication and minimizing data loss in a changing world. *Brief. Bioinform*., 9, 166-173

Orchard, S., *et al*. (2008). 6th HUPO annual world congress - Proteomics standards initiative workshop: 6-10 October 2007, Seoul, Korea. *Proteomics*, 8, 1331-1333

Patient, S., *et al*. (2008). UniProtJAPI: A remote API for accessing UniProt data. *Bioinformatics*, 24, 1321-1322

Ping, P., *et al*. (2008). Wiley-VCH and HUPO: A global effort to advance proteomic science. *Proteomics*, 8, 4-6

Quintaje, S.B. & Orchard, S. (2008). The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: One small step in manual annotation, one giant leap for full comprehension of genomes. *Molecular and Cellular Proteomics*, 7, 1409-1419

Reisinger, F., *et al*. (2008). ENFIN - An integrative structure for systems biology. In *Lecture Notes in Computer Science*, 132-143

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

Siepen, J.A., *et al*. (2008). ISPIDER Central: an integrated database web-server for proteomics. *Nucleic Acids Res*., 36: W485-490

Smits, G., *et al*. (2008). Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. *Nat. Genet*., 40, 971-976

Tanaka, T., *et al*. (2008). The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res*., 36, D1028-D1033

Taylor, C.F., *et al*. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nat. Biotechnol*., 26, 889-896

Tharakan, R., *et al*. (2008). OMSSAGUI: An open-source user interface component to configure and run the OMSSA search engine. *Proteomics*, 8, 2376-2378

The UniProt Consortium, (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Res*., 36, D190-195

closely together over the last year to enable Reactome to release over 3,830 high-quality manual GO annotations to 60 different species.

The GOA group is also the primary supplier of electronic GO annotations to the GO Consortium. In UniProtKB there are currently 168,308 species (4,257,090 proteins) for which electronic annotation pipelines, provided by GOA, are the only source of GO annotation. The group is responsible for providing five GO mapping files (Swiss-Prot keyword to GO, UniProtKB Subcellular Location to GO, Enzyme Commission numbers to GO, InterPro to GO and HAMAP to GO) which 'translate' external vocabularies to equivalent GO terms to create annotations for many species. Such mapping files are continually updated so that currently over 14,450 external terms are mapped to GO, providing more than 33 million annotations from entries in UniProtKB. In addition, the electronic annotation collaboration with Ensembl Compara has continued successfully; now providing over 147,858 annotations (a 300% increase over the year) for 30 species by transferring high-quality manual annotations to 1:1 orthologues in closely related species.

A grant from the BBSRC Tools and Resources Fund to redevelop GOA's QuickGO browser ended in 2008. This funding enabled QuickGO to be comprehensively redeveloped, allowing users to query using a range of different keywords or identifier types to find either comprehensive, detailed information on GO terms or sets of GO annotation data, which they can filter to their specific needs and download in a range of formats. Support for the creation, modification and annotation of GO slims is now also provided by this tool.

GOA curators continue to be key members of the GO 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. In 2008 GOA were successful in securing funding from Kidney Research UK, who have agreed to fund one dedicated curator for three years to improve the functional dataset available for mammalian gene products implicated in kidney development and disease. This will provide a valuable community resource for renal researchers and will complement the cardiovascular annotation effort. It is intended that this project will start in January 2009. In 2009 GOA will also be working closely with Swiss-Prot curators from the Swiss Institute of Bioinformatics, who will be providing manual GO annotations to GOA.

RESID

John S. Garavelli

The RESID Database of Protein Modifications (Garavelli 2004; http://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 formulas and masses, enzymatic activities that generate the modifications, keywords, literature citations, cross-references to the Gene Ontology (GO), ChEBI, PSI-MOD, PDB and MSD, structure diagrams and molecular models. Quarterly Release 55 (September 2008) contained 453 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 gathered for the RESID Database 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 has been used as a contributing component of the Proteomics Standards Initiative ontology of protein modifications (PSI-MOD).

Figure 2. *Tag cloud indicating the worldwide distribution of PANDA workshops.*

PANDA OUTREACH AND TRAINING

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

OUTREACH

Xosé M. Fernández, Bert Overduin, Michael Schuster, Giulietta Spudich

The PANDA group has delivered a comprehensive number of workshops worldwide, facilitating the dissemination of our resources. Moreover it has played a main role in many training activities organised at the EBI (e.g. 'A two-day dip into EBI's data resources', BioSapiens European School in Bioinformatics). In addition to PANDA's dedicated outreach representatives, this has also involved several PANDA developers presenting their tools.

The group's outreach activities have also focused on introducing EBI resources to locations emerging in the bioinformatics field, such as China (where we organised five workshops in Shanghai and Beijing), Mexico and India. At the HUGO 13th Human Genome Meeting in Hyderabad, in September 2008, we presented EBI resources as well as participating in a satellite workshop featuring Ensembl, UniProt and InterPro. We have also been active in Africa with workshops in Egypt, Kenya and South Africa. Figure 2 shows a tag cloud of cities hosting Ensembl training activities in the last year.

- Ensembl (VectorBase, Integr8, Genome Reviews) 71 courses
- Proteomics (UniProt, InterPro, IntAct, PRIDE) 28 courses
- Sequence databases (EMBL-Bank, EGA, GOA) 8 courses
- Pathways (Reactome, ChEBI, IntEnz) 11 courses
- EBI (Roadshows and Hinxton-based workshops) 24 courses

Our training brochures and tutorials have been updated. Similarly, the new eLearning platform from the EBI features tutorials for PANDA resources (see the Outreach and Training team report on page 27 for more details on the eLearning project). As a harbinger of future training activities within SLING (Serving Life-science Information for the Next Generation), the UniProt group has begun to share training materials between trainers at SIB and EBI in order to avoid duplication efforts.

Ensembl website redesign

Ensembl has undertaken a major effort to improve the usability of the website. Changes were driven by user feedback, with extensive user testing before and during the redesign process. Face-to-face group discussions with users, and one-to-one sessions allowing us to observe users performing different tasks, highlighted a range of website features to address. Alternative layouts were proposed, test-1, test-2 and a hybrid between both, test-3 (see Figure 3 overleaf).

Following the original selection, a panel of over 300 users was consulted about different options (naming of tracks, labelling, layout options etc.) and their feedback was essential for the new design which will be implemented in release 51 of Ensembl (Figure 4).

Zerbino, D.R. & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, 18, 821-829

Zhang, J., *et al*. (2008). Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria. *Proteomics*, 8, 1564-1575

Zhang, J., *et al*. (2008). Altered proteome biology of cardiac mitochondria under stress conditions. *J Proteome Res*, 7, 2204-2214

Other publications

Camon, E., *et al*. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., 32, D262-D266

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

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

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

Leinonen, R., *et al*. (2006). UniSave: the UniProtKB Sequence/Annotation Version database. *Bioinformatics*, 22,1284-1285

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

Mulder, N.J., *et al*. (2007). New developments in the InterPro database. *Nucleic Acids Res*., 35, D224-228

Wu, C.H., *et al*. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res*., 32, D112-114

Birney

2007

Birney, E. (2007). Double Dutch for duplications. *Nat. Genet*., 39, 1303-1304

Birney, E. (2007). Evolutionary genomics: Come fly with us. *Nature*, 450, 184-185

Del Bene, F., *et al*. (2007). In vivo validation of a computationally predicted conserved Ath5 target gene set. *PLoS Genet*., 3, 1661-1671

Flicek, P. (2007). Gene prediction: compare and CONTRAST. *Genome Biol*., 8, 233

Holland, P.W.H., *et al*. (2007). Classification and nomenclature of all human homeobox genes. *BMC Biol*., 5, 47

Kahlem, P. & Birney, E. (2007). ENFIN A Network to enhance Integrative Systems Biology. *Ann. NY Acad. Sci.*, 1115, 23-31

Nielsen, F., *et al*. (2007). Optimising oligonucleotide array design for ChIP-on-chip. *BMC Bioinformatics*, 8, i195-i204

Prlic, A., *et al*. (2007). Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, 8, 333

Spudich, G., Fernández-Suárez, X.M. & Birney, E. (2007). Genome browsing with Ensembl: A practical overview. *Briefings in Functional Genomics and Proteomics*, 6, 202-219

1. Based on your feedback, we now have three different designs. Have a click around with these sites, navigate to a gene/transcript of choice. Which do you prefer now?

	Response Percent
test-1	15.6%
test-2	4.4%
test-3	57.8%
current ensembl	22.2%

Figure 3. *Feedback from users driving the design of the new Ensembl interface.*

New publications

HGNC has launched a newsletter with the latest release, which was circulated in advance of the Human Genome Meeting in Hyderabad, India.

VectorBase has continued to publish its quarterly newsletter highlighting new available datasets as a way to communicate with the community. Issue 5 is about to be released.

Two tutorial papers were published in the last year: 'Genome Browsing with Ensembl: A practical Overview. (Spudich, Fernández-Suárez & Birney, 2007) and 'Advanced genomic data mining' (Fernández-Suárez & Birney, 2008).

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 2008, the group hosted ten trainees and five visitors, working on a broad variety of projects. The trainees were:

- Irina Armean (FH Hagenberg, Austria), IntAct confidence score, Proteomics Services Team;
- Dominik Grimm (Weihenstephan University, Germany), IntAct InterMine, Proteomics Services Team;

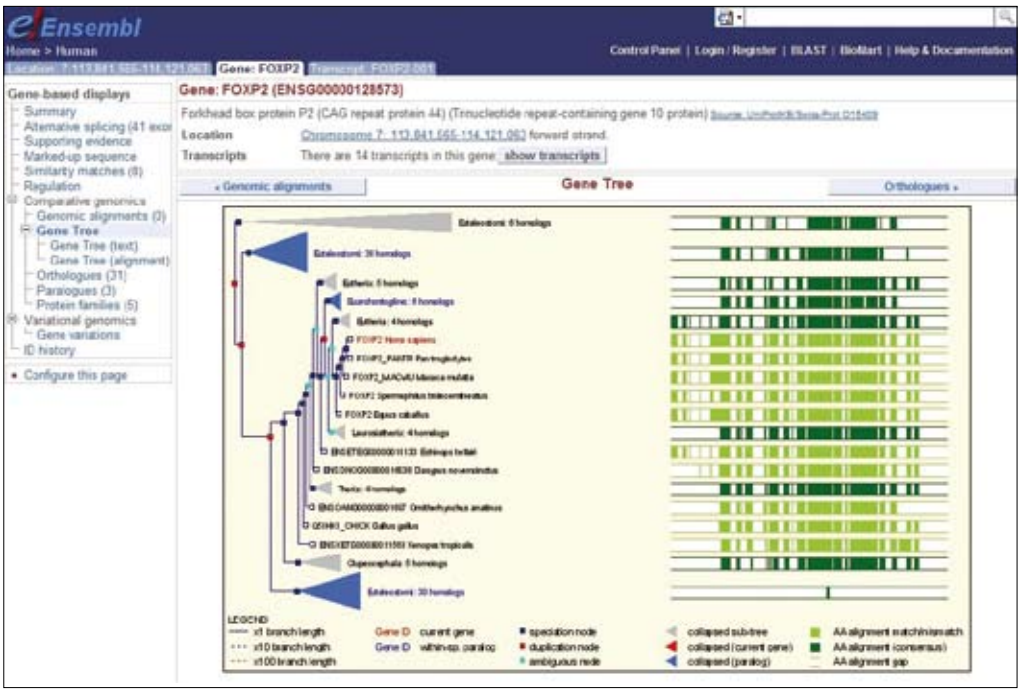


Figure 4. *Screenshot of the new Ensembl interface combining tabs and left-hand menus.*

- Andreas Hoelzlwimmer (FH Hagenberg, Austria), Reactome pathways import, Proteomics Services Team;
- Wolfgang Kluge (Weihenstephan University, Germany), EnVision set-oriented data analysis, Proteomics Services Team;
- Michael Menden (Weihenstephan University, Germany), IntAct BioMart, Proteomics Services Team;
- Nadine Neuhauser (Weihenstephan University, Germany), Interaction graph visualisation, Proteomics Services Team;
- Kieran O'Neill (National Bioinformatics Network, Cape Town, South Africa), OntoDAS, Proteomics Services Team;
- Omar Pera Mira (Polytechnical University of Valencia, Spain), myDAS, Proteomics Services Team;
- Jonathan Rameseder (FH Hagenberg, Austria), PRIDE, Proteomics Services Team;
- Avazeh Tashakkori Ghanbarian (University of Leeds, UK), Interaction visualisation, 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

Three PhD students (Joe Foster, Garth Ilsley and Michael Mueller) and one postdoc (Kai Ye) are currently working under the supervision of Rolf Apweiler. Another ongoing research activity is the automatic annotation of proteins.

Improved exploration of large biological datasets

Garth Ilsley

Most web-based searches of large biological datasets are based on either metadata or a hard-coded statistical function that takes a gene or set of genes as the input. It is not clear whether this limitation is inherent, or whether significantly better searching methods are possible. For example, one might wish to specify a network of interactions with dynamic behaviour and retrieve datasets according to their relevance. Currently, this problem is being explored in the context of *Drosophila* development, in particular with high resolution *in situ* hybridisation data from 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. The question of what inference the data can support is being studied with well-established statistical methods, whereas the appropriate model formalism for searching is being investigated within the computer science field of model checking.

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.

Stark, A., *et al*. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450, 219-232

Sterk, P., *et al*. (2007). The EMBL Nucleotide Sequence and Genome Reviews Databases. *Methods Mol. Biol*., 406, 1-22

Stranger, B.E., *et al*. (2007). Population genomics of human gene expression. *Nat. Genet*., 39, 1217-1224

2008

Aitman, T.J., *et al*. (2008). Progress and prospects in rat genetics: A community view. *Nat. Genet*., 40, 516-522

Birney, E. (2008). Levers and fulcrums: Progress in cis-regulatory motif models. *Nat. Methods*, 5, 297-298

Bruford, E., *et al*. (2008). The HGNC database in 2008: A resource for the human genome. *Nucleic Acids Res*., 36, D445-D448

Cochrane, G., *et al*. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 36, D5-D12

Cole, C.G., *et al*. (2008). Finishing the finished human chromosome 22 sequence. *Genome Biol*., 9, Article R78

Down, T.A., *et al*. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol*., 26, 779-785

Flicek, P., *et al*. (2008). Ensembl 2008. *Nucleic Acids Res*., 36, D707-D714

Han, R., *et al*. (2008). An efficient conformational sampling method for homology modeling. *Proteins: Structure, Function and Genetics*, 71, 175-188

Jenkinson, A.M., *et al*. (2008). Integrating biological data - The Distributed Annotation System. *BMC Bioinformatics*, 9, Article S3

Johnson, D.S., *et al*. (2008). Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res*., 18, 393-403

Lappalainen, I., *et al*. (2008). Genome wide analysis of pathogenic SH2 domain mutations. *Proteins: Structure, Function and Genetics*, 72, 779-792

Margulies, E.H. & Birney, E. (2008). Approaches to comparative sequence analysis: Towards a functional view of vertebrate genomes. *Nat. Rev. Genet*., 9, 303-313

O'Reilly, P.F., *et al*. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res*., 18, 1304-1313

Quintana, F.J., *et al*. (2008). Control of Treg and TH17 cell differentiation by the aryl hydrocarbon receptor. *Nature*, 453, 65-71

Reisinger, F., *et al*. (2008). ENFIN - An integrative structure for systems biology. In *Lecture Notes in Computer Science*, 132-143

Ruan, J., *et al*. (2008). TreeFam: 2008 Update. *Nucleic Acids Res*., 36, D735-D740

Saar, K., *et al*. (2008). SNP and haplotype mapping for genetic analysis in the rat. *Nat. Genet*., 40, 560-566

Smits, G., *et al*. (2008). Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. *Nat. Genet*., 40, 971-976

Topalis, P., *et al*. (2008). How can ontologies help vector biology? *Trends Parasitol*., 24, 249-252

An efficient, versatile and scalable pattern growth approach for string matching in biological sequences

Kai Ye

Traditionally, string matching algorithms require transformation of sequence databases into either hash tables or suffix trees as the first step in facilitating efficient querying of individual sequences. However, as more complete genomes are accumulated, large-scale sequence comparisons across species as well as among different sequence clusters within the same species are required. Since the simple query-oriented algorithms no longer fulfil research requests efficiently, we are exploring our novel pattern growth approach in genome-wide sequence comparisons.

The pattern growth approach has been proved to be linear to the size of the input, scalable to genome-wide research and versatile in finding different pattern types. We are applying this algorithm to the following research topics:

- mining frequent patterns in unaligned protein sequences;
- mapping large amounts of peptides to the genome;
- mapping short reads to the genome;
- identification of indels breakpoints of any length, in genomes using short reads;
- scanning genome sequences for hairpin rich regions.

Automatic annotation of UniProtKB/TrEMBL

Ricardo Antunes, Michael Kleen, Maria Jesus Martin, Claire O'Donovan, John O'Rourke, Samuel Patient

UniProtKB/Swiss-Prot is under strict quality control and kept as consistent, complete and up to date as possible by an experienced team of biologists. The information it contains 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.

In the last year, an intensive effort has been made to redevelop and streamline the existing automatic annotation pipelines for the EBI's automatic annotation approaches of Spearmint and RuleBase. This would allow faster annotation generation, appropriate statistical cross-validation and closer integration with InterPro cross-references. Data is currently exported to UniProtKB/TrEMBL at > 99% confidence value and above a 0.95 score for Spearmint and >97% confidence value and above a 0.80 score for RuleBase. All other predictions are available for review by a newly developed internal curator tool (CAAT). This tool greatly increases the ability for review of the predicted annotation and enables both the development of new rules and the further development/refinement of existing rules. In the near future, CAAT should facilitate the provision of a predicted annotation system for external users in conjunction with the InterPro team at the EBI. Development is also ongoing for a central UniProt repository of rules which will include rules from the HAMAP and PIRSF rule systems, from SIB and PIR respectively. These are currently part of the curation pipeline for UniProtKB/Swiss-Prot but will be applied to UniProtKB/TrEMBL in the near future facilitating a further increase in coverage. 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. This will be further extended to UniProtKB/Swiss-Prot as the HAMAP and PIRSF rules are integrated.

BIRNEY RESEARCH

Ewan Birney's research group focuses on algorithmic methods for genome analysis. Four PhD students are currently working under the supervision of Ewan Birney; Markus Hsi-Yang Fritz, Alison Meynert, Dace Ruklisa and Daniel Zerbino. In addition, one joint EIPOD postdoc works in the group, Mikhail Spivakov. Over 2007/2008 we have worked on the following projects.

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 been 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.

Genetic epistasis models

Dace Ruklisa

We have used protein–protein and pathway information to constrain genetic epistasis models using classical genetic experiments, such as recombinant inbred line data. Using these models we have more statistical power and can detect genetic interactions in these datasets. We have extended this model to genome-wide association studies, requiring some refactoring of the code to handle the data volumes.

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 are also looking at extending this 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.

FUTURE PROJECTS AND GOALS

It is our intention to work on improved integration and synchronisation of all PANDA resources. We will move towards use of Ensembl software and adopt its features and functionalities for all complete genomes. 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 new methodologies to the annotation of UniProtKB/TrEMBL records.

Tsesmetzis, N., *et al*. (2008). *Arabidopsis* reactome: A foundation knowledgebase for plant systems biology. *Plant Cell*, 20, 1426-1436

Twigger, S.N., *et al*. (2008). What everybody should know about the rat genome and its online resources. *Nat. Genet*., 40, 523-527

Vastrik, I. (2008). Installing a local copy of the Reactome Web site and database. *Current protocols in bioinformatics,* Andreas D. Baxevanis *et al*. (editoral board), Chapter 9

Warren, W.C., *et al*. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453, 175-183

Zerbino, D.R. & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*., 18, 821-829

Other EMBL publications

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

Vertebrate Genomics

Paul Flicek

DSc 2004, Washington University.
At EMBL-EBI since 2005.
Honorary Faculty Member, Wellcome Trust Sanger Institute since 2008.
Team Leader at EMBL-EBI since 2008.

Team Members

Project Leaders
Javier Herrero (Ensembl Compara)
Damian Smedley (Mouse Informatics)
Nathan Johnson (Ensembl Production Software)
Fiona Cunningham* (Ensembl Variation)

Scientific Programmers
Mario Caccamo
Laura Clarke*
Stefan Gräf
Jonathan Hinton*
Zamin Iqbal*
Damian Keefe
Vasudev Kumanduri*
Ilkka Lappalainen
Steven Wilder*

Ensembl Developers
Kathryn Beal
Benoit Ballester
Yuan Chen
Stephen Fitzgerald
Leo Gordon*
Daniel Rios*
Albert Vilella

Bioinformaticians
Phil Wilkinson

Team Secretary
Shelley Goddard

** Indicates part of the year only*

INTRODUCTION

The Vertebrate Genomics team is a combined service and research group that creates and manages data resources focusing on genome annotation and human variation. The major service projects of the Vertebrate Genomics team are Ensembl, the European Genotype Archive, and the data coordination centre for the 1000 Genomes Project. In support of these projects, we are developing the specialised, large-scale bioinformatics infrastructure required for each analysis.

The team's research is on computational genome annotation with a particular focus on the integration of diverse data types such as extensive comparative sequencing, DNA–protein interactions, epigenetic modifications, and the DNA sequence itself.

ENSEMBL

Benoit Ballester, Kathryn Beal, Yuan Chen, Fiona Cunningham, Stephen Fitzgerald, Leo Gordon, Stefan Gräf, Javier Herrero, Nathan Johnson, Damian Keefe, Daniel Rios, Albert Vilella, Steven Wilder

The Ensembl project (Flicek *et al.*, 2008) is a comprehensive genome information system featuring an integrated set of tools for genome annotation, data mining and visualisation of chordate genomes. Selected model organism and disease vector genomes are also provided by Ensembl and will be extensively supported in the future by the Ensembl Genomes project. Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute. At the EBI, Ensembl includes members of the Vertebrate Genomics team and components of the PANDA Nucleotides group (see page 43).

Ensembl is one of the world leaders in genome annotation, which is the process of finding functional elements in large genomes using computational and manual approaches. All Ensembl data is freely provided without restriction, and Ensembl is one of the fundamental database resources used to address questions in medical research and molecular biology. This year saw the publication of a number of high-profile projects with Ensembl participation including the platypus genome (Warren *et al.,* 2008), the *Drosophila* 12 genomes project (*Drosophila* 12 Genomes Consortium, 2007), and a rat haplotype map (Saar *et al.*, 2008).

The primary entry point to Ensembl is the website http://www.ensembl.org. As of August 2008, there are 39 fully-supported genomes in Ensembl including human, mouse, chicken, five species of fish, a nematode, and several other mammalian, chordate and insect species. A further five organisms have preliminary support. This report will contain an overview of the entire Ensembl project with a focus on functional genomics, variation, and comparative genomics components of the project within the Vertebrate Genomics team.

At the core of the project are the Ensembl gene sets and the comprehensive annotation pipeline that produces them. We work to develop methods to improve the quality of this annotation and adapt to the range of assemblies and available supporting data. The number of species available in Ensembl continues to grow rapidly, and over the past year full evidence-based gene builds have been performed on several species including human, mouse, cow, chimp, Tetraodon, Takifugu, guinea pig, orangutan and horse. In addition we have carried out gene projection gene builds on nine low coverage mammalian genomes produced as part of a NHGRI-sponsored project to sequence multiple mammalian genomes for comparative annotation (http://www.genome.gov/25521745). We have also regularly updated ncRNA gene predictions on many species.

The Ensembl website facilitates the display of most Ensembl data and significant new resources in functional genomics and variation were incorporated this year. The Ensembl web team has spent much of the previous twelve months developing a new web interface, incorporating extensive user feedback, which uses modern browser and server technologies. To this end, the re-developed web code will consist of smaller, faster-loading user pages; contain standards-compliant HTML, Javascript and CSS; introduce streamlined use of AJAX to include content; incorporate shared memory caching; and provide tweaked server settings to improve browser performance. The new Ensembl website was released in late 2008.

The Ensembl analysis pipeline and the Ensembl website are supported by an extensive software system. The Ensembl software team has grown significantly in size and scope over the past year and now supports the infrastructure needs of the Ensembl project and the Ensembl Genomes project. The

Publications

2007

Drosophila 12 Genomes Consortium, (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450, 203-218

Flicek, P. (2007). Gene prediction: compare and CONTRAST. *Genome Biol*., 8, 233

Stein, C., *et al*. (2007). Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biol*., 8, Article R251

Stranger, B.E., *et al*. (2007). Population genomics of human gene expression. *Nat. Genet*., 39, 1217-1224

2008

Down, T.A., *et al*. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol*., 26, 779-785

Flicek, P., *et al*. (2008). Ensembl 2008. *Nucleic Acids Res*., 36 (Database issue): D707-D714

Johnson, D.S., *et al*. (2008). Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res*., 18, 393-403

Quintana, F.J., *et al*. (2008). Control of Treg and TH17 cell differentiation by the aryl hydrocarbon receptor. *Nature*, 453, 65-71

Ruan, J., *et al*. (2008). TreeFam: 2008 Update. *Nucleic Acids Res*., 36, D735-740

Saar, K., *et al*. (2008). SNP and haplotype mapping for genetic analysis in the rat. *Nat. Genet*., 40, 560-566

Warren, W.C., *et al*. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453, 175-183

latter project has been under development for several months and expects to be formally launched in late 2008 or early 2009 (See Paul Kersey's report on page 65 for more information). The software team now supports the core Ensembl database infrastructure and APIs as well as providing support for annotation hosted on non-Ensembl servers and visualised with Ensembl using the Distributed Annotation System (DAS). The software team has also taken responsibility for the production of the Ensembl BioMart that is part of each Ensembl release. Primary software development over the past year has focused on several optimisations aimed at improving the speed of the API and expanding the number of species that can be stored in a single core database.

Ensembl training is conducted around the world by dedicated PANDA outreach and training staff including courses in Asia, North America and Europe. Furthermore, eLearning courses have been deployed as part of the EBI's user training programme, which supplement existing video tutorials designed to improve the user experience and maximise the utility of Ensembl. In conjunction with the Ensembl website redesign, the PANDA outreach team has undertaken a comprehensive update of the Ensembl help documentation.

Ensembl Comparative Genomics

The comparative genomics team focuses on the evolution of the genome both at the protein and at the genomic level. Their work includes developing methods to annotate orthologues and paralogues, and methods to construct pairwise and whole genome multiple alignments.

Gene tree and orthology/paralogy prediction: the core comparative genomics pipeline infrastructure has remained stable. We start by aligning all the Ensembl genes in a pairwise fashion to cluster them. Proteins within the same cluster are aligned and the TreeBeST method is used to both infer the phylogeny and call speciation and duplication events. During this year, we incorporated the SLR algorithm from Nick Goldman's group at the EBI to calculate site-wise dN/dS values and to identify sites under positive and negative evolution. We have also implemented a basic way to perform incremental builds for pairs of species that have not changed since the previous release to partly mitigate the need for increased computational resources.

Pairwise alignments: we continue to provide pairwise alignments for a selected set of species. We use the BlastZ-net approach for pairs of closely related species, namely among amniotes, between Medaka and stickleback, and between *Ciona savignyi* and *Ciona intestinalis*. The BlastZ-net alignments are the result of post-processing the raw alignments to obtain best-in-genome sets of collinear alignments. For pairs of more distant species, we use BLAT (the BLAST-like alignment tool) in translated mode. Starting from release 50 (July 2008), we have also provided best-in-genome alignments using the same post-processing approach as for BlastZ alignments.

Whole genome multiple alignments: our strategy is based on the detection and alignment of collinear segments between the genomes (see Figure 1). The initial procedure uses Mercator for the detection of orthologous collinear regions and MLagan for aligning them. Over the course of the previous 18 months, we have completely re-engineered this procedure based on software programs developed within the team and in close collaboration with a now-graduated EMBL PhD student. We first replaced MLagan by the Pecan alignment program and started using GERP to calculate conservation scores and detect constrained elements. More recently, we replaced Mercator with Enredo, a graph-based method able to find both orthologous and paralogous collinear regions. Our final step uses Ortheus to infer ancestral sequences from the alignments. Ortheus uses a branch transducer model, a type of Hidden Markov Model to call deletion and insertion events, providing a realistic model under which it can infer the ancestral sequence. The March 2008 Ensembl release (version 49) saw the first set of EPO (Enredo-Pecan-Ortheus) alignments on a set of seven mammals. For the following July release (version 50), we extended this set of alignments to horse and orangutan and added a set of four-way primate only EPO alignments.

The whole genome multiple alignment methodology has been extended to include the low coverage genomes described above. These genome assemblies are far less complete and therefore create too many breakpoints in the Enredo graph. To solve this, we built the Enredo graph using only the high coverage genomes and then mapped the low coverage genomes on the collinear regions by using pairwise alignments to the human genome.

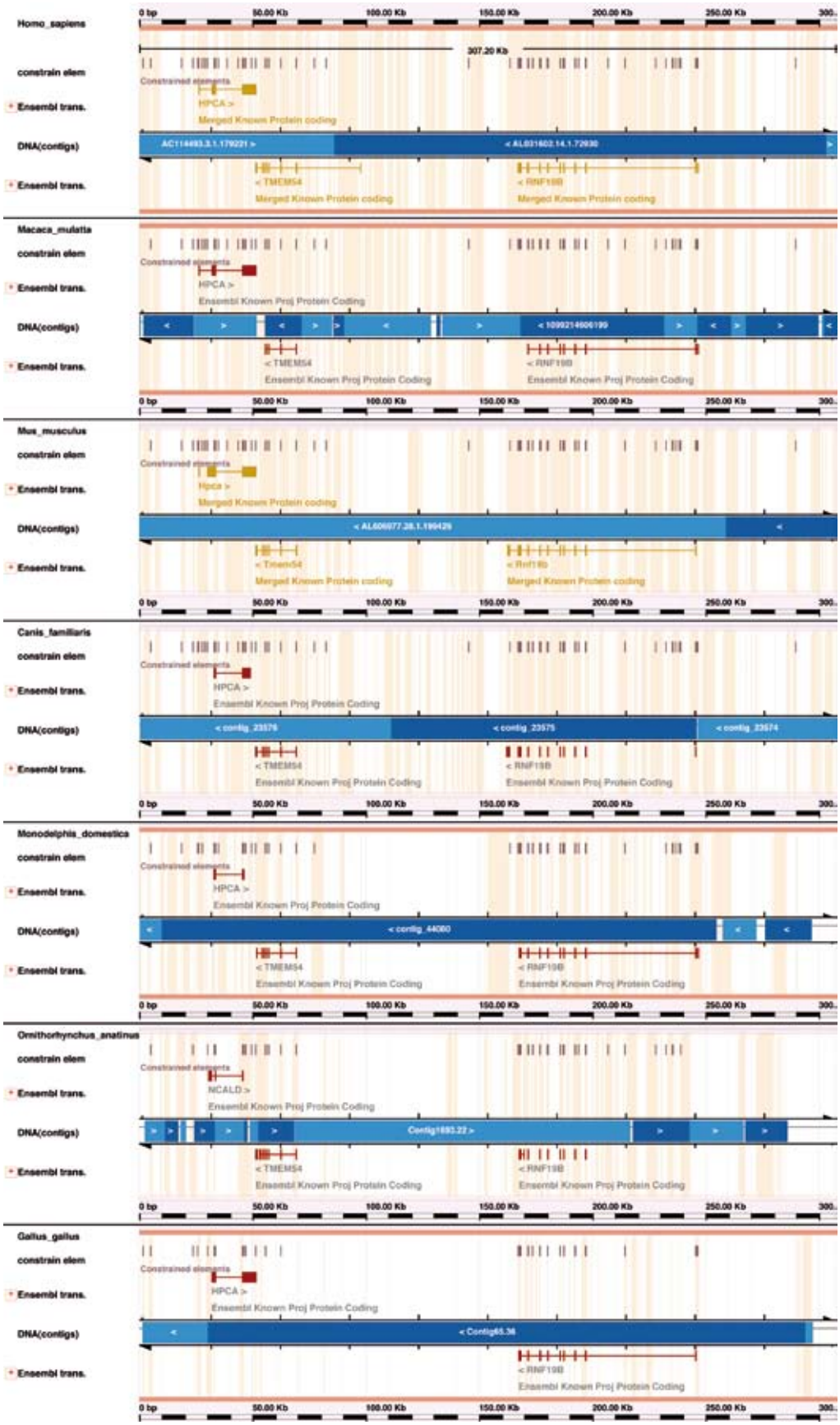


Figure 1. *An Ensembl multiple alignment showing the decreasing amount of alignable sequence to the human genome based on a region of human chromosome 1. Species shown include (from the top): human, macaque, mouse, dog, opossum, platypus and chicken.*

Ensembl functional genomics

Throughout the year, we have continued active development of the Ensembl regulatory build. The regulatory build provides an automatic, evidence-based functional annotation of the genome. The primary inputs for the regulatory build are maps of open chromatin created by assays of DNase I hypersensitivity and covalent modifications of histone protein tails assayed by chromatin immunoprecipitation (ChIP). The first build was released in 2007 in coordination with the ENCODE Pilot Project publication. Since the first release, we have created three additional versions of the regulatory build each adding more data and a more sophisticated analysis of the chromatin conformation and modification data.

The other major effort of the functional genomics group is the development of a ChIP-seq analysis pipeline including a custom algorithm for the analysis of ChIP-seq data. This development has been driven by our participation in a number of collaborative research projects aimed towards understanding protein–DNA binding interactions and how these interactions affect genome function. Another major collaborative effort mapped DNA methylation across a number of tissue types and resulted in the creation of a genome-wide DNA methylation resource (Down *et al.*, 2008) that has now been incorporated into Ensembl.

Ensembl variation

Ensembl variation data is derived from two major sources and represents a significant growth area for the project. Data imported from dbSNP comprise the core variation data, while computationally discovered variations from resequencing data are growing rapidly in parallel with next-generation sequencing technology. In the last year, we have incorporated the data from three successive builds of dbSNP (127, 128 and 129) and have been working more closely with dbSNP. We incorporated resequencing-based SNPs from platypus and orangutan as well as from the resequenced human genomes of James Watson and Craig Venter. Our platypus SNPs were submitted to dbSNP and make up the largest set of SNPs available for that species. Our orangutan SNPs will be submitted in conjunction with the publication of that genome.

Within the variation database, we have increased support for copy number variation data and annotation of individual SNPs including variations associated with common disease and identified in genome-wide association studies and variations associated with expression QTLs identified by Stranger *et al.*, 2007.

EUROPEAN GENOTYPE ARCHIVE

Mario Caccamo, Jonathan Hinton, Vasudev Kumanduri, Ilkka Lappalainen

The European Genotype Archive (EGA) database provides a permanent archive for all types of personally identifiable genetic data including genotypes, genome sequence and associated phenotype data. The EGA contains both data collected from individuals whose consent agreements limit data release to specific research uses or *bona fide* researchers, and data approved for full public release.

The EGA has a distributed access-granting policy. We are partners with the data-generating organisations and access decisions will be made by the appropriate data access-granting organisation (DAO) and not by the EGA. The DAO will normally be the same organisation that approved and monitored the initial study protocol or a designate of this approving organisation. In a typical case, the EGA website will direct users to a project homepage where the user can apply for access that is then managed by the DAO.

Accepted data types include manufacturer-specific raw data formats and analysis results such as genotype calls from the original study authors. The EGA also accepts and distributes any phenotype data associated with the samples. The EGA will accept all data that includes a DAO-approved access plan. It is not required that the data be available for full public release, either at the time of submission or at any time in the future. Prepublication data is accepted and is visible to approved users (e.g. consortium members or manuscript reviewers) only after logging into the EGA. The EGA will exchange summary-level data (consistent with DAO approval) with other similar archives worldwide.

This year we included major datasets from the Wellcome Trust Case Control Consortium (see Figure 2).

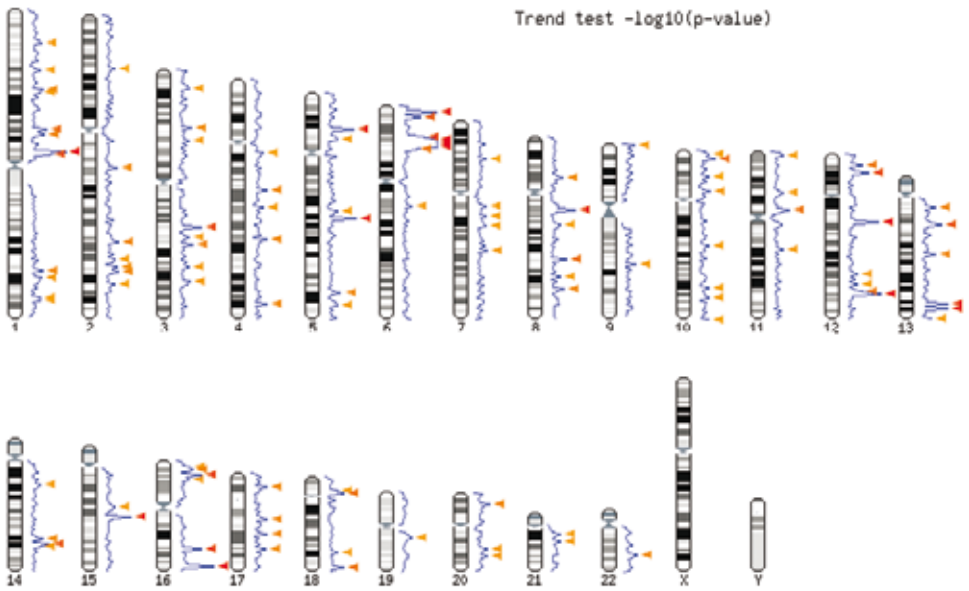


Figure 2. An example GenomeView from the European Genotype Archive showing genomic regions that are significantly associated with type I diabetes.

Implementation

The EGA consists of an optimised database schema for supported data types and accepts data submissions in standard formats such as MAGE-TAB and PLINK. We are implementing a SIMS/AIMS (http://simbioms.sourceforge.net/aims.html) automated submission system compatible with several existing standard data formats.

Security

Data and physical security: all of the EGA data is stored in a MySQL database and an associated file system, to which access has been limited by user group to include only members of the EGA project. All sensitive data are stored and distributed in encrypted form. Furthermore, the EGA disks are attached to a single machine (and not accessible to any other machine on the EBI network). Each dataset will be designated as a specific group to which approved users will be added. The EGA stores individual identifiers and their genotypic data in separate databases, and the connection between these data is made using an abstract identifier.

Web security: security protocols have been implemented by the EBI web team and include a phpBB login system with https for downloading controlled-access data. This system has been successfully applied to other EBI-hosted EU projects containing controlled-access data.

THE 1000 GENOMES PROJECT

Laura Clarke, Zamin Iqbal

The 1000 Genomes Project aims to create a comprehensive and public catalogue of common human genetic variation in three populations by using next-generation sequencing technology. The partners in the 1000 Genomes Project include nine sequence production groups in Europe, North America and Asia, and multiple analysis groups around the world. Due to the broad consent agreements that the underlying DNA samples were collected under, the 1000 Genomes Project is able to release all DNA sequence into the public domain immediately after generation. During 2008, the project conducted three pilot projects to assess the feasibly of creating a deep and accurate catalogue and develop the necessary tools to manage and analyse the data. The pilot projects included the sequencing of 180 individuals to 2x coverage; sequencing two trios consisting of a child and both parents to 20x coverage; and targeted sequencing of 1,000 genes in 1,000 individuals.

In collaboration with the NCBI, the Vertebrate Genomics team is one half of the 1000 Genomes Project Data Coordination Centre (DCC) and has co-leadership of the project's data flow group. Over the course of the year the project produced approximately 2Tb of sequence (equivalent to 8.5

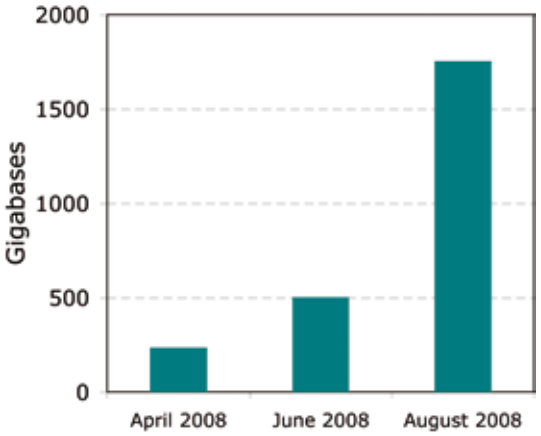


Figure 3. *Amount of sequence data submitted by the 1000 Genomes Project production centres from April 2008 until the end of August 2008.*

times the number of nucleotides in the EMBL-Bank sequence archive) at a rate approaching 30Gb per day (see Figure 3). This data is collected by the DCC and exchanged between the NCBI and EBI groups before being made available to the 1000 Genome Project analysis group and interested researchers worldwide. Significant development on infrastructure for data tracking and low-level analysis has been the most important task within the DCC this year.

In addition, the Vertebrate Genomics team coordinates the 1000 Genomes Project website (http://www.1000genomes.org) in conjunction with the EBI's External Services team.

FUTURE PROJECTS AND GOALS

A major new feature of the Ensembl project is the comprehensive redesign of the web interface. Throughout the project, the impact of next-generation sequencing methods is having a profound impact. For example, we have been investigating ways to use short read transcriptome data in our automatic annotation to support the substantial amounts of data we expect in the future. Initial experience with orangutan and zebrafish data will be generalised for transcript model generation. The availability of an increasing number of genome sequences is challenging the comparative genomics aspects of the project both in terms of scale and complexity. Building on the success of the EPO pipeline, we plan multiple alignments in the teleost fish lineage, which are complicated by an additional whole genome duplication. ENCODE and the 1000 Genomes Project will respectively provide significant new data into the functional genomics and variation resources. Future developments for the EGA include a suite of customised data mining tools, an analysis pipeline infrastructure supporting uniform analysis of the data in the archive, and the development (in collaboration with international partners) of standards for the exchange of genotype data including whole genome sequences.

The Ensembl Genomes Team

Paul Kersey

Ph.D. 2005, University of Edinburgh.
At EMBL since 1999.

Team Members

Ensembl Metazoa Coordinator
Daniel Lawson

Senior Software Engineer
Alan Horne
Gautier Koscielny
Peter Sterk

Software Engineer
Matthias Haimel
Arnaud Kerhornou
Devang Lakhani
Vincent le Texier*
Rajesh Radhakrishnan*
Daniel Staines
Andy Yates

Scientific Programmer
Karyn Megy

VectorBase Bioinformatician
Martin Hammond

** Indicates part of year only*

INTRODUCTION

The Ensembl Genomes team operates within the larger PANDA (Protein and Nucleotide Data) group. The PANDA group was created in June 2007 by merging the former Ensembl (Birney) and Sequence Database (Apweiler) groups. The Ensembl Genomes team was created in May 2008 as a separate entity within PANDA, led by Paul Kersey.

The activities of the Ensembl Genomes team are focused on the representation of genome and genome-related data from non-vertebrate species. It thus complements the activities of the Vertebrate Genomics group, led by Paul Flicek. As is also the case for vertebrate species, a huge influx of data is expected for non-vertebrate species in the near future, driven by the use of ultra-high throughput sequencing technologies. The group's core strategy for handling this data is based on the belief that the demands of these species can be met through the re-use and extension of the Ensembl genome annotation, analysis and visualisation platform that has been used successfully for vertebrate genomes since 2000. This strategy will become apparent with the public launch of five new sites (Ensembl Metazoa, Ensembl Protists, Ensembl Plants, Ensembl Fungi and Ensembl Bacteria) over the course of late 2008 and early 2009.

In association with the launch of Ensembl Genomes, the team is making a major effort to deepen our links with sections of the scientific community working on particular species. While Ensembl provides a powerful toolkit, close links to the community are essential to ensure quality control, scientific relevance and, ultimately, financial viability. Our vision is that Ensembl Genomes will evolve as a project with two faces; to the wider scientific public, the sites will provide an integrative portal to genomes and related data; while simultaneously providing an infrastructure and flexible toolset to empower particular communities to manage genome annotation. Our collaborations with communities may well involve the establishment of independent resources, tailored to individual communities' needs; but easily integrated within the wider Ensembl Genomes site through the use of shared underlying technology.

The team is already involved in one such project: VectorBase, a database of the genomes of metazoan vectors of human pathogens. Additionally, the group is involved in maintaining four databases that precede the foundation of PANDA: Integr8, Genome Reviews, IPI and ASTD. These projects will be wound down after the launch of Ensembl Genomes and their residual functionality incorporated into 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 data access for users, and is a major beneficial consequence of the creation of PANDA.

One of the team's other tasks is to provide data relating to complete proteomes to the UniProt database. Substantial training and outreach efforts are also part of the group's activities. Certain external service aspects of the group's activities are described in the report by Rodrigo Lopez, team leader of the EBI External Services. Development of the Ensembl web code for the purposes of Ensembl Genomes takes place within the EMBL Nucleotides section of PANDA led directly by Ewan Birney.

The main achievements of the Ensembl Genomes group in 2008 have been:

- preparing for the launch of Ensembl Genomes as a public service from late 2008;
- increasing the number of complete cellular genomes available in the Integr8 database to 812 by early October 2008 (an increase of 192 species from the previous year);
- working with the NIAID Microbial Sequencing Centers to generate annotations for two genomes: *C. quinquefasciatus* and *I. scapularis*, as part of the VectorBase project;
- incorporating comprehensive annotation for non-coding RNA genes for all genomes in Genome Reviews.

ENSEMBL GENOMES

Alan Horne, Matthias Haimel, Martin Hammond, Arnaud Kerhornou, Paul Kersey, Gautier Koscielny, Devang Lakhani, Daniel Lawson, Karyn Megy, Daniel Staines, Andrew Yates

The genome is a central concept at the heart of biology. Since the first complete genome was sequenced in the mid-1990s, over 800 more have been sequenced, annotated, and submitted to the

public databases. New ultra-high throughput sequencing technologies are now beginning to generate complete genome sequence at an accelerating rate, both to gap-fill portions of the taxonomy where no genome sequence has yet been deciphered (for example, the GEBA project http://www.jgi.doe.gov/programs/GEBA/, which aims to sequence 6,000 bacteria from taxonomically distinct clades), and to generate data for variation in populations of species of special interest (for example, the 1000 Genomes Project in human http://www.1000genomes.org and the 1001 Genomes Project in *Arabidopsis* http://1001genomes.org). In addition, modern sequencing technologies are increasingly being used to generate data for gene regulation and expression on a genome-wide scale.

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 Ewan Birney on page 43). The main focus of the Ensembl Genomes team is now to leverage the use of these solutions for non-vertebrate species. New sites for invertebrate metazoa, protists and bacteria will be launched in late 2008, to be followed by the subsequent launch of sites for plants and fungi in the first half of 2009.

The initial release of Ensembl Bacteria (illustrated in Figure 1) will focus on six bacterial and one archaeal clade (*Pyrococcus*), with multiple genomes represented in each clade for the purposes of comparative analysis. The bacterial clades represented include *Escherichia*, *Shigella, Bacillus* (home to the model organisms *E. coli* and *B. subtilis*) and four important clades of pathogenic bacteria (*Mycobacterium*, *Neisseria*, *Staphylococcus* and *Streptococcus*).

Ensembl Metazoa will focus on non-chordate metazoan species (this is complementary to the vertebrate focus of Ensembl) and will initially focus on three main groupings: worms, flies and arthropod vectors of human pathogens. We have initiated key collaborations with responsible groups to integrate the genome and annotations from WormBase (www.wormbase.org), FlyBase (www.flybase.org) and VectorBase (www.vectorbase.org). The set of genomes includes the model organisms *Caenorhabditis elegans* and *Drosophila melanogaster* as well the vectors such as *Anopheles gambiae* and pathogens such as *Brugia malayi*. Identifying new collaborations to bring in taxonomic pioneer species such as the sea urchin and planarian flatworm will be a major focus of the coming year.

Ensembl Protists will focus on the Apicomplexa species of human pathogens which includes the causative agents of malaria (*Plasmodium falciparum* and other *Plasmodia*). A key collaborator is the EuPathDB group (http://eupathdb.org/eupathdb/) which hosts 16 eukaryotic pathogen genomes.

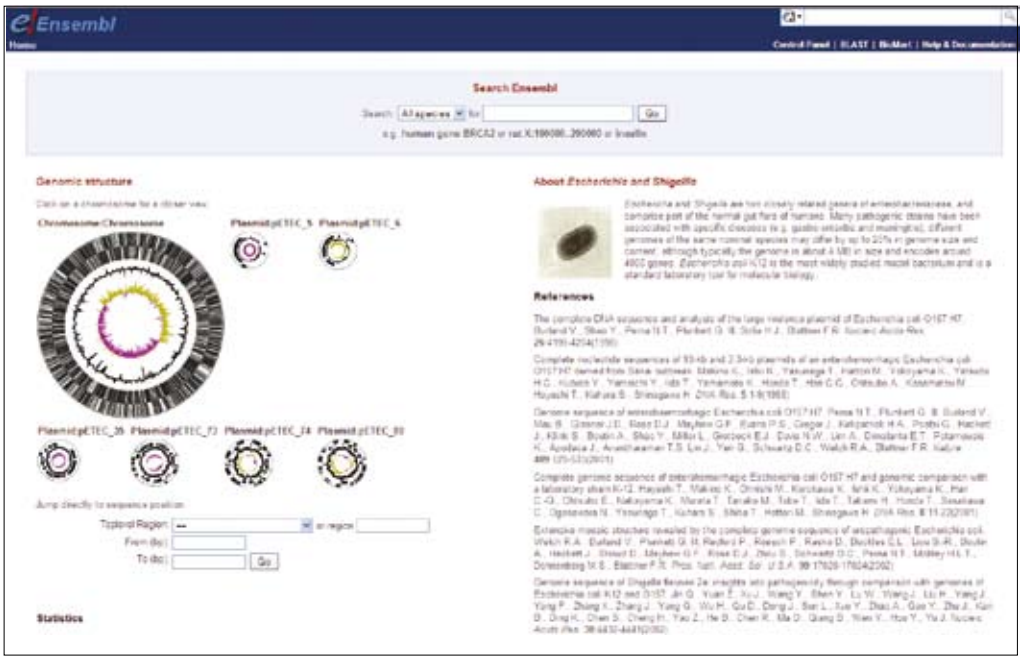


Figure 1. *Ensembl Genomes. The homepage for the* Escherichia/Shigella *clase within the new Ensembl Bacteria site; due for public release at the end of 2008.*

Publications

2007

Sterk, P., *et al*. (2007). The EMBL Nucleotide Sequence and Genome Reviews Databases. *Methods Mol. Biol*., 406, 1-22

2008

Field, D., *et al*. (2008a). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol*., 26, 541-547

Field, D., *et al*. (2008b). Meeting report: The fourth Genomic Standards Consortium (GSC) workshop. *OMICS A Journal of Integrative Biology*, 12, 101-108

Field, D., *et al*. (2008c). Meeting report: The fifth Genomic Standards Consortium (GSC) workshop. *OMICS A Journal of Integrative Biology*, 12, 109-113

Megy, K., *et al*. (2008). Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. *Infection, Genetics and Evolution*, Article in press

Mulder, N.J., *et al*. (2008). In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol. Biotechnol.*, 38, 165-177

Taylor, C.F., *et al*. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nat. Biotechnol*., 26, 889-896

Topalis, P., *et al*. (2008). How can ontologies help vector biology? *Trends Parasitol*., 24, 249-252

Other EMBL publications

Hubbard, T., *et al*. (2007). Ensembl 2007. *Nucleic Acids Res*., 35, D610-D617

Fagnani, M., *et al*., (2007). Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol*., 8, R108

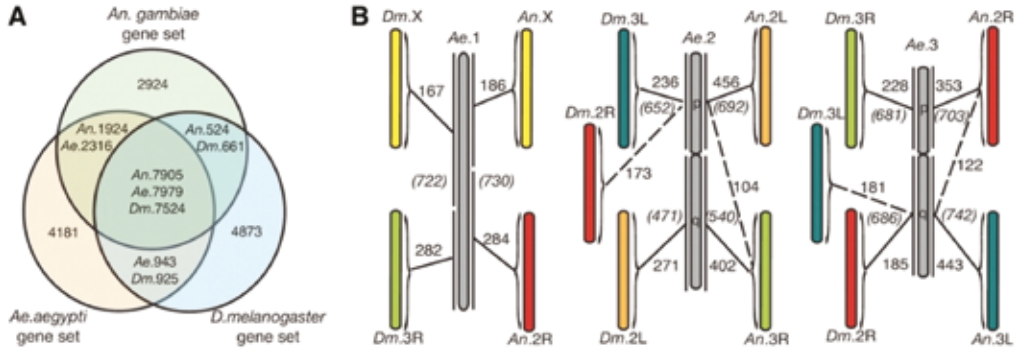


Figure 2. *Orthology and chromosomal synteny among* Aedes aegypti, Anopheles gambiae *and* Drosophila melanogaster*. (A) Each circle represents a gene set for* Ae. aegypti *(*Ae*),* An. gambiae *(*An*), and* D. melanogaster *(*Dm*). (B) Chromosomal synteny for the* Ae. aegypti *chromosomes (represented in grey) and* An. gambiae *and* D. melanogaster *(coloured chromosomes). Solid and dashed lines link each* Ae. aegypti *chromosome to its primary and secondary syntenic chromosome, respectively. The number of* Ae *orthologs to* An *and* Dm *chromosome arms are indicated with the total number of orthologs on the* Ae *chromosome arm to* Ae *or* Dm *is shown in italics and parentheses.*

Features of the initial release of Ensembl Genomes will include the availability of BioMarts (query optimised databases for efficient data mining) and the application of comparative analysis pipelines initially developed by the Compara team for vertebrate genomes to bacteria and metazoa. Compara will be applied selectively and at different levels: for example, DNA-based comparisons will be run on closely related bacterial strains; while protein-based analysis will be run within bacterial clades, across the bacterial and metazoan space, and across a selection of species from throughout the taxonomy.

As well as preparing for the launch in a technical sense, we have also been making contact with selected important groups in the *Arabidopsis*, monocot, plant pathogen, human pathogen, and model organism communities. The first practical fruit of these contacts, a project to capture *Arabidopsis* variation data within an Ensembl-based infrastructure is due to begin in late 2008.

VECTORBASE

Martin Hammond, Daniel Lawson, Karyn Megy

VectorBase is an NIAID Bioinformatics Resource Center (http://www.brc-central.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*), arborial 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 2008 VectorBase has worked with the NIAID Microbial Sequencing Centers to generate annotations for two genomes: *C. quinquefasciatus* and *I. scapularis*. The availability of the *C. quinquefasciatus* genome annotation facilitates comparison of the three main families of mosquitoes (Anopheline, Aedine and Culicine) with the model Dipteran (*Drosophila melanogaster*) to address questions about hematophagy and vectorial competence. The *I. scapularis* genome is the first chelicerate (a major subphylum of arthropods including spiders, mites and ticks) to be sequenced. A comparison of the genomes of some of these species is shown in Figure 2.

VectorBase related data is available from the VectorBase website (http://www.vectorbase.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.

Gaidatzis, D., *et al*. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics, 8*, 69

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

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

INTEGR8

Alan Horne, Matthias Haimel, Arnaud Kerhornou, Paul Kersey, Devang Lakhani, Rajesh Radhkrishnan, Andrew Yates

Integr8 (Kersey *et al.*, 2005; http://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. There are essentially two classes of information available in Integr8: information about the species as a whole, and information about individual genes found within these species.

For each species, a short descriptive overview is provided, as well as a list of recently published literature. A detailed statistical analysis of the genome and proteome of each species is also presented using a combination of tabular and graphical displays. Key tools used in this analysis include InterPro, CluSTr and GO, which each provide ways of classifying the proteins that collectively define a proteome. A Genome Annotation Score is also calculated, which attempts to measure how completely each genome is characterised and allows improvements in annotation to be tracked over time. Many resources are also available for download, including EMBL-Bank and Genome Reviews files, non-redundant DNA, transcript, and protein sets, information about orthologues, and GO/InterPro annotation files. The number of cellular organisms present in the database increased by 192 between October 2007 and October 2008 to its current value of 811; the increase is illustrated in Figure 3. A clear majority (almost 700) of these genomes are bacterial; the other half are split, roughly equally, between eukaryotic and archaeal genomes. Since November 2007, bacteriophage genomes can also be accessed through the site.

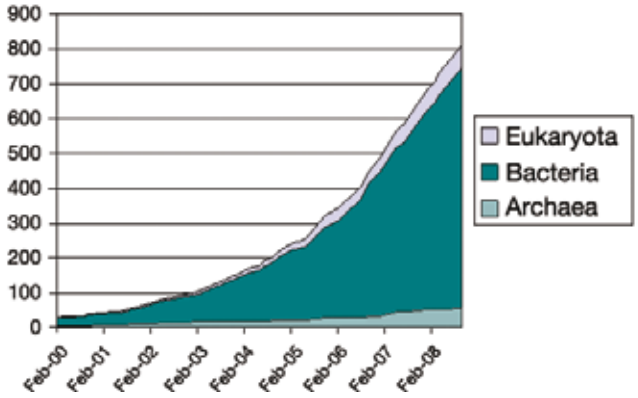


Figure 3. *Complete genomes in Integr8. The figure shows the increase in the number of complete genomes in Integr8 (and its predecessor projects) since 2000. This year might be the last year of relatively moderate growth, as new ultra-high throughput sequencing technologies start to generate previously unheard-of quantities of data.*

Information about individual genes can be accessed through the 'Integr8or' component of the Integr8 portal. Access to Integr8or is provided through a graphical search interface that allows users to direct their search to a specified portion of the taxonomic tree. Once a gene has been selected, Integr8or provides a clear view of the products (transcripts and proteins) of the gene and the annotations (including cross-referencing entries in external resources) associated with each. A schematic representation of the location of each gene in the context of its neighbours is also provided. Additionally, a list of orthologues and paralogues is provided for each gene; alignments of protein sequence and the potentially syntenic genomic regions centred on homologous genes can be viewed; and a graphical view, showing the distribution of orthologous families over the taxonomy, has been added recently (see Figure 4).

Another component of Integr8, 'Inquisitor', provides an expert system for protein sequence analysis, applying a number of different tools to identify each sequence in Integr8, or, if the sequence is novel, to classify it as fully as possible. As when searching for genes, sequence searches can be restricted to a given taxonomic scope.

In addition to accessing the Integr8 web portal and FTP site, web services are also available, providing programmatic access. This year, a new Perl package for remote access to Integr8 was added to complement an existing package available for the Java programming language.

In order to concentrate resources behind a limited number of public interfaces, the Integr8 brand will be retired in 2009, following the launch of Ensembl Genomes. Data that currently appears in Integr8 will be made available through the Ensembl Genomes and UniProt interfaces.

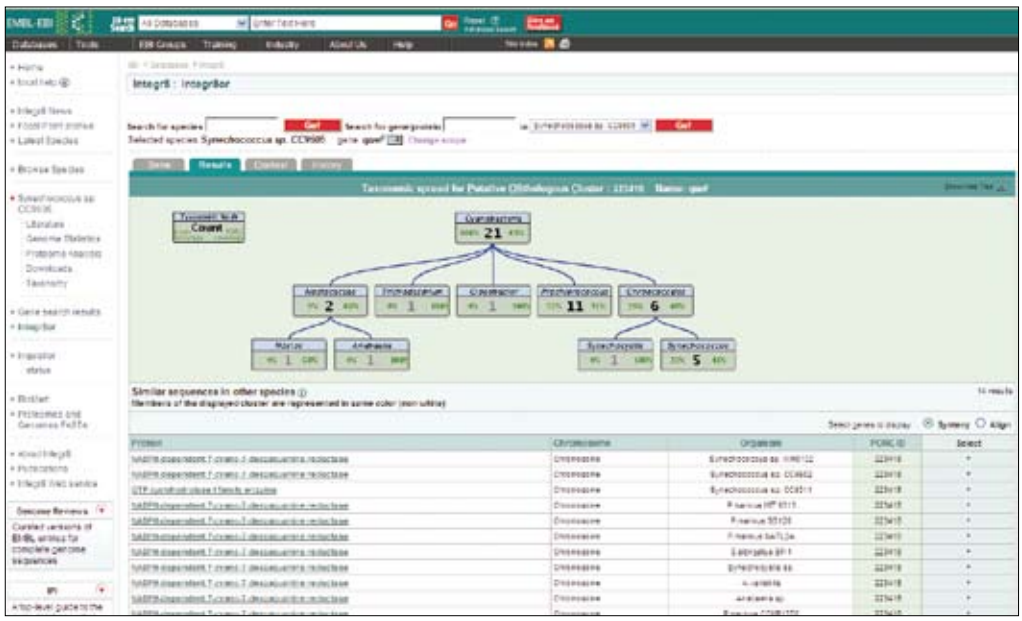


Figure 4. Integr8 orthologue viewer. Putative orthologues are calculated between the proteomes of all species in Integr8. The taxonomic spread of each orthologoues family is visible in this viewer. Individual orthologues can be selected for alignment or synteny display.

GENOME REVIEWS

Arnaud Kerhornou, Paul Kersey, Peter Sterk

Genome Reviews (Kersey *et al.*, 2005; http://www.ebi.ac.uk/GenomeReviews) has been developed as a database of complete genome sequences, focusing on bacteria, bacteriophage and selected lower eukaryotes. Data is derived from submissions to the public nucleotide sequence archives and improved by the automatic import of manual annotation from curated resources such as UniProtKB/Swiss-Prot, Regulon DB; and up-to-date, sequence-derived functional annotation (e.g. annotation from GOA). Additionally, discrepancies in DNA and protein-level annotation are identified through sequence analysis and labelled; and certain gene types not universally annotated in archived submissions (e.g. tRNA, rRNA genes) are added using standard software pipelines. Since February 2008, annotation of predicted non-coding RNA genes has been added for all genomes where this was not part of the archive submission.

Another interesting development has been the assignment of stable, transcript IDs for all genes in Genome Reviews, and the release of datasets representing each (verified or predicted) transcript to provide a mapping between identified genes and their protein products. Annotated gene, transcript and protein sequence can be accessed through the Integr8 website on a per-gene basis; or downloaded in bulk from the Genome Reviews FTP site.

Genome Reviews are distributed in EMBL-like flat files. In addition, cellular and bacteriophage genomes in Genome Reviews have been available in an Ensembl-style browser since 2005 and 2007 respectively. In order to concentrate resources behind a limited number of public interfaces, the Genome Reviews brand will be retired in 2009, following the launch of Ensembl Genomes. Data that currently appears in Genome Reviews will be made available through the Ensembl Genomes interface.

IPI

Matthias Haimel, Devang Lakhani

IPI (Kersey *et al.*, 2004; the International Protein Index, http://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 IPI releases. IPI is released at three weekly intervals.

The rationale for the creation of IPI was the presence of large numbers of divergent gene predictions in different databases describing higher eukaryotic proteomes; and the problem that a nominally non-redundant set of protein sequences may reflect the historical accumulation of sequence variants (and errors) from across a population. The IPI algorithm, which uses a combination of sequence analysis and primary annotation to merge different predictions, creates non-redundant datasets for each species defining redundancy according to a tighter criterion than sequence identity, so that partial sequences, and variant alleles, are not independently represented in the non-redundant set.

In 2009, the EBI's services will be rationalised by the direct incorporation of data from Ensembl and Ensembl Genomes into the UniProt Knowledgebase, ensuring consistent data is available in the primary EBI services delivering genome and protein annotation. When this happens, IPI will be discontinued as a separate service.

ALTERNATIVE SPLICING AND TRANSCRIPT DIVERSITY (ASTD)

Gautier Koscienly, Vincent le Textier

Transcript expression in eukaryotes is subject to variation at three main biological stages: transcription initiation, splicing and polyadenylation. In mammals, most genes undergo some kind of alternative transcription. Current data for human indicates that at least 81% of genes are subject to alternative transcription initiation, 69% to alternative splicing and 60% to alternative polyadenylation. Abnormal expression of alternative transcripts has been linked to multiple diseases, and especially to cancer. The sheer number and wide biological impact of alternative transcripts has created a high demand for tools enabling the identification, classification, functional annotation and expression profiling of alternative transcription in major model genomes.

ASTD (http://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. The database 1) covers all three aspects of alternative transcription; 2) includes three model vertebrate species and enables extension to new species; 3) offers a powerful interface for expression pattern-based queries; 4) is fully integrated with other genomic data and genome browsing capabilities; and 5) is extensible with respect to functional features and regulatory motifs.

Alternative transcripts are derived from the mapping of transcribed sequences (EST/cDNA sequences from EMBL) to the complete human, mouse and rat genomes using an extension of the computational pipeline developed for the ASD and ATD databases, which are now superseded by ASTD. For the human genome, ASTD identifies splicing variants, transcription initiation variants and polyadenylation variants in 68%, 68% and 62% of the gene set, respectively, consistent with current estimates for transcription variation.

The scientific community can access ASTD through a variety of browsing and query tools, including expression state-based queries for the identification of tissue-specific isoforms.

We have placed a significant emphasis on the experimental validation of predicted transcripts using RT-PCR. Together with participating labs from the ATD Consortium (INSERM ERM206 in Marseille, France and University Hospital of Heidelberg, Germany), we have completed the experimental validation of over 500 different polyadenylation and splicing events in human and mouse that were not previously described in the literature (Koscielny *et al.*, submitted).

For polyadenylation sites, our validation efforts have focused on events conserved between human and mouse that produce alternative 3' isoforms with size variations of 3kb or more. Out of 86 such events, 84 have been individually confirmed using a specially devised RT-PCR strategy. For splicing variants, we have focused on the identification of cancer-specific events. From a list of ASTD-predicted tissue-specific splice variants, we have confirmed 73 isoforms with specific expression in human cancer cell lines versus normal tissues. The validation details for the model of alternative splicing for one such gene are shown in Figure 5.

Many analyses rely on the alignments of EST/cDNA sequences to genomes sequences and several resources like ASTD have been built in the same way. However the *de novo* identification of alternative splicing events that have been missed by EST and cDNA sequencing requires more sensitive and efficient strategies. In recent years, high-throughput technologies like custom splicing junction microarrays have been designed to detect changes in splicing on a genome-wide scale and address questions such as the identification of alternatively spliced exons and the cellular context within

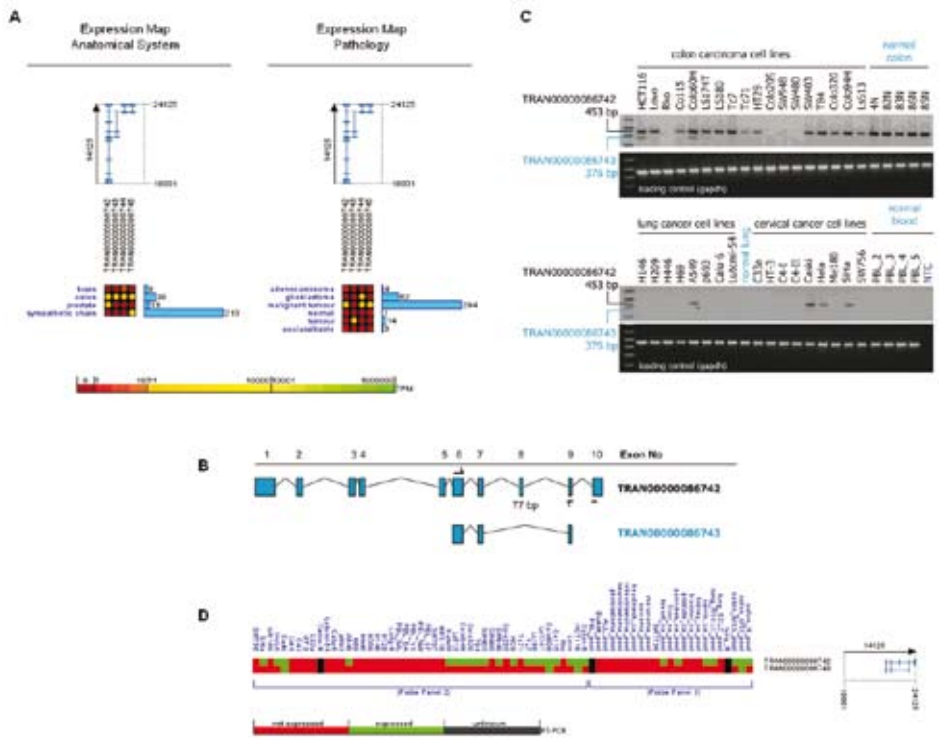


Figure 5. *Validation detail for a model of alternative splicing. The figure shows validation detail for the transmembrane Mucin 12 gene (UniProtKB/TrEMBL accession Q9UKN1). The digital expression map obtained from the ASTD database (A) shows that while this gene is mainly expressed in the colon, transcript TRAN00000086743 is predicted to be colon tumour-specific. The latter transcript (B) lacks exon 8 relative to transcript TRAN00000086742 which should be present in colorectal normal and cancer cells. RT-PCR experiments confirmed these predictions. The transcript including exon 8 was mainly detectable in colon cell lines and normal colon tissues while the skipped exon transcript was present in colorectal cancer cell lines and absent in all five tested normal colon samples (C). PCR validation data are incorporated in the ASTD database in the form of graphical overviews showing positive/negative PCR results for each tested splice variant and condition (D).*

which such events take place. To date, we have been collaborating with Genome Canada to incorporate data from a quantitative profiling study on alternative splicing microarray, combining set of exons and splicing junction probes (Fagnani *et al.*, 2007). ASTD gives access to this data analysis for 3,707 mouse cassette alternative splicing events and the estimates for percentage exon exclusion levels for these events in 27 mouse tissues. This information is also available as a DAS track on the Ensembl website.

Another project has been undetaken in collaboration with the National Institute of Allergy and Infectious Diseases (NIAID) at the NIH. The project aims to identify novel transcripts using an alternative splicing microarray-based method. In this work, the microarray analysis relies on the existing ASTD alternative transcripts dataset to mine and assess true positive alternative splicing events. This strategy increases the detection of *de novo* candidates confirmed in a second round by RT-PCR.

We also play an active role in EURASNET (http://www.eurasnet.info), the Alternative Splicing Network of Excellence which brings together leading research groups and young investigators from eleven European countries. The network aims to investigate and understand the mechanisms of alternative splicing. Our role is to maintain the ASTD service for the scientific community and to incorporate valuable data from other members of the network. For example, we have recently worked with the Biozentrum at the University of Basel to map miRNA functional target sites on the 3' UTR of the ASTD transcripts (Gaidatzis *et al.*, 2007). These miRNA targets are made available as a DAS track on the ASTD website.

New sequencing technologies have tremendous potential for the detection and monitoring of alternative splicing by offering advantages in terms of increased throughput and sensitivity. Several very recent studies have applied analyses of short cDNA read data to profile alternative splicing in human

and mouse tissues and cell lines. We would like to broaden the scope of the analysis and integration of alternative splicing related information based on emerging or existing technologies. At the same time, we want to rationalise the incorporation of such data into Ensembl and Ensembl Genomes and future developments are being focused in this direction.

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 workshops and contributing to the development of the emerging standards.

The fifth GSC workshop was held from 12–14 December 2007 at the EBI. The key outcome of the workshop was the finalisation of a stable version of the Minimal Information about a Genome Sequence (MIGS) specification (v2.0) for publication in *Nature Biotechnology* (Field *et al.*, 2008a). The proceedings of the fourth and fifth GSC meetings (Field *et al.*, 2008b; Field *et al.*, 2008c) were published in a special issue of *OMICS* produced as a direct result of the fifth workshop. The sixth GSC workshop was held from 13–17 October 2008 at the EBI, and focused on carrying forward the core GSC projects described in the June 2008 special issue of *OMICS*.

The GSC is also involved in the Minimum Information for Biological and Biomedical Investigations (MIBBI) project, which provides a resource for those exploring the range of extant minimum information checklists and fosters coordinated development of such checklists (Taylor *et al.*, 2008).

FUTURE PROJECTS AND GOALS

2009 will see the launch of the Microme Project, a 13 member European partnership to develop a database and downstream services for metabolic pathways in bacteria. Based on the successful Reactome software infrastructure, Microme is especially timely as pathway-based inference is likely to become increasingly important in the annotation and functional interpretation of bacterial genome sequence. Novel genome sequence has the potential to suggest biotechnological solutions to problems such as energy and food production, waste decontamination and industrial catalysis.

We will continue to work with the UK and international research communities to establish collaborations to exploit the Ensembl infrastructure in new domains and bring fresh content to Ensembl Genomes. Currently, new proposals are being prepared for human and plant pathogen data, while we are seeking to deepen existing relationships with leading resources in the areas of plant and protist pathogens, such as Gramene and EuPath DB. The relationship between hosts, vectors and pathogens is of particular interest as Ensembl Genomes (and Ensembl) expand to include increasing numbers of complete disease systems.

We will also be working to develop the interface between genomic and metagenomic data, an area of increasing scientific interest and data production.

The Proteomics Services Team

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 Cold Spring Harbor Laboratory, New York.

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, BioSapiens, FELICS, ENFIN, ProteomeBinders and Transfog) 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

Phil Jones, Samuel Kerrien, Lennart Martens, Luisa Montecchi-Palazzi, Sandra Orchard

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 Proteomics Standards Initiative (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 June 2008, the mzML format for the representation of mass spectrometry data was released by the PSI (Deutsch, 2008);
- **controlled vocabularies:** while XML schemas 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

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.

Team Members

Coordinator
Lennart Martens
Sandra Orchard
Esther Schmidt

Senior Software Engineer
Richard Côté
Phil Jones
Samuel Kerrien

Postdoctoral Fellow
Manuel Corpas*
David Gloriam*
Juan Antonio Vizcaíno

Scientific Database Curator
Bernard de Bono
Cathy Derow*
Phani Garapati
Bijay Jassal
Jyoti Khadake
Luisa Montecchi-Palazzi*
David Thorneycroft

Software Engineer
Premanad Achuthan
Bruno Aranda
David Croft
Antony Quinn*
Florian Reisinger

Visitor
Matthieu Visser

Marie Curie Fellow
Magali Michaut*

Trainee
Irina Armean*
Michael Menden*
Omar Pera Mira*
Nadin Neuhauser*
Kieran O'Neill*
Jonathan Rameseder*
Wolfgang Kluge*

** Indicates part of the year only.*

vocabulary where necessary, for example, protein interaction detection technologies (Martens, Palazzi & Hermjakob, 2008). The combination of reasonably stable XML schemas 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 Proteomics Services team's 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 addition to the work group-specific deliverables, the Proteomics Services team contributed to a cross-work group for the standardised representation of protein modifications (PSI-MOD; Montecchi-Palazzi *et al.*, 2008) and genomic sequence metadata (MIGS; Field *et al.*, 2008). To address the problem of a proliferation of potentially uncoordinated domain-specific standardisation efforts, we also contribute to the development of a central registry of minimum requirement specifications (MIBBI; Taylor *et al.*, 2008b) and a proposed domain-independent data model for molecular biology (FuGe; Jones *et al.*, 2008).

The major PSI event is the annual PSI spring meeting which was held in Toledo, Spain, in 2008 (Orchard *et al.*, 2008), and will next take place in Turku, Finland, in April 2009. The PSI is an open, collaborative initiative. We invite comments and participation in new and existing work groups. Full project information is available from the PSI website www.psidev.info.

MOLECULAR INTERACTIONS

Premanand Achuthan, Bruno Aranda, Cathy Derow, David Gloriam, Samuel Kerrien, Jyoti Khadake, Magali Michaut, Luisa Montecchi-Palazzi, Sandra Orchard, David Thorneycroft

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.

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. From March 2007 to February 2008, IntAct recorded more binary interactions resulting from direct data deposition by authors than from literature curation for the first time. Overall, IntAct now contains more than 170,000 molecular interactions from direct data depositions, literature curation focusing on specific journals and literature curation focusing on specific topics like cancer, chromatin or *Arabidopsis* (Morsy *et al.*, 2008).

The toolkit around IntAct has been extended by InteroPORC, a system for the homology-based prediction of molecular interactions (Michaut *et al.*, 2008), and an R package for the analysis of molecular interactions in PSI MI format (Chiang *et al.*, 2008).

PROTEIN IDENTIFICATIONS

Richard Côté, Phil Jones, Lennart Martens, Dave Thorneycroft, 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. In 2007–2008, PRIDE has been completely refactored to cope efficiently with the rapid increase in data depositions (Jones *et al.*, 2008), but also to provide an improved user experience. In particular, PRIDE now supports the annotation of fragment ions on identified peptides (Figure 1). Both the PRIDE dataset view and the PRIDE BioMart link to Reactome to allow analysis of proteomics datasets in the context of human pathways. PRIDE data content has more than tripled from 2.6 million spectra in September 2007 to 10.5 million spectra in August 2008.

Publications

2007

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

Eisenacher, M., *et al*. (2007). Proteomics data collection - The 1st ProDaC workshop 26 April 2007: Ecole Normale Supérieur, Lyon, France. *Proteomics*, 7, 3034-3037

Flikka, K., *et al*. (2007). Implementation and application of a versatile clustering tool for tandem mass spectrometry data. *Proteomics*, 7, 3245-3258

Jones, A.R., *et al*. (2007). The Functional Genomics Experiment model (FuGE): An extensible framework for standards in functional genomics. *Nat. Biotechnol*., 25, 1127-1133

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

Martens, L., *et al*. (2007). Human proteome organization proteomics standards initiative: Data standardization, a view on developments and policy. *Molecular and Cellular Proteomics*, 6, 1666

Orchard, S. & Hermjakob, H. (2007). Bioinformatics approaches in proteomics. In *Proteomics* (O'Connor, C. D. and Hames, B.D., eds), 227-243, Scion Press

Orchard, S. (2007). Proteomics: From technology development to biomarker applications: HUPO 6th Annual World Congress, 6-10 October, 2007, Seoul, Korea. *Expert Rev. Proteomics*, 4, 709-710

Orchard, S., *et al*. (2007). Submit your interaction data the IMEx way: A step by step guide to trouble-free deposition. *Proteomics - Practical Proteomics*, 2, 28-34

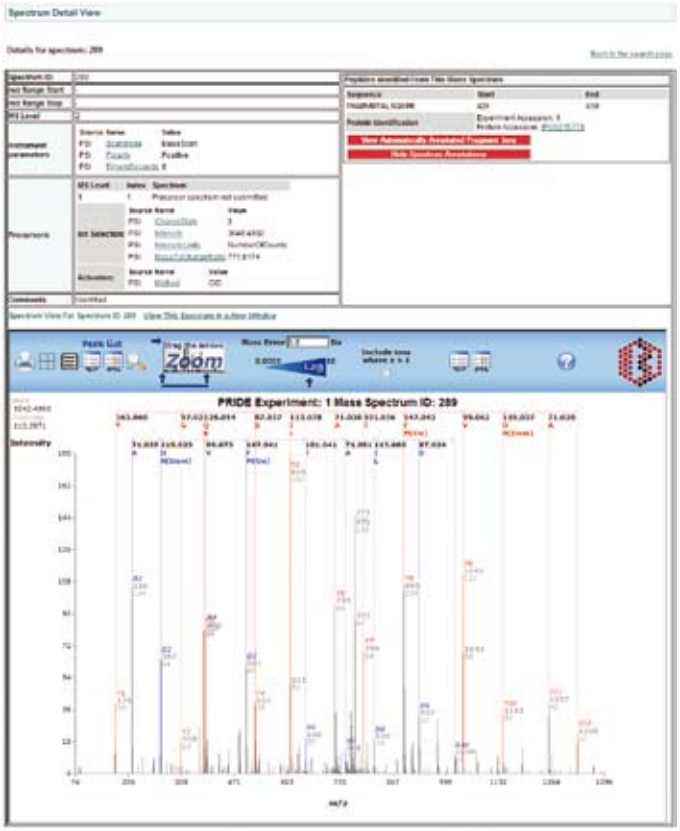


Figure 1. PRIDE detail view for a submitted spectrum, including annotation of submitted fragment ions in the spectrum.

The research use of PRIDE data has been exemplified by two recent publications (Mueller *et al.*, 2008 and Klie *et al.*, 2008). Klie *et al.* illustrate the capacity of different experimental technologies to sample different subsets of the human proteome, while Mueller *et al.* analyse the detection of central nervous system-related gene products in the human brain and cerebrospinal fluid. Both publications build on the large amounts of standardised and annotated experimental data in PRIDE to gain new insights across multiple experimental datasets of different origin.

PATHWAYS

David Croft, Bernard de Bono, Phani Garapati, Bijay Jassal, Esther Schmidt

Reactome (http://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 and CSHL, 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 hormonal signalling. The current release includes 2,975 human proteins, 2,907 reactions and 4,455 literature citations. This represents a coverage of approximately 12.5% of 20,000 curated UniProtKB human proteins, a 2.7-fold increase over the last three years.

A new entity-level pathway viewer and improved search and data mining tools facilitate searching and visualising pathway data and the analysis of user-supplied high-throughput datasets. Reactome has increased its utility to the model organism communities, with improved orthology prediction methods allowing pathway inference for 22 species and through collaborations, to create manually curated Reactome pathway datasets for species including *Arabidopsis* (Tsesmetzis *et al.*, 2008), *Oryza sativa* (rice), *Drosophila* and *Gallus gallus* (chicken).

DATA INTEGRATION

Richard Côté, Rafael Jimenez, Omar Pera Mira, Antony Quinn, 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

Orchard, S., *et al*. (2007). Five years of progress in the standardization of proteomics data 4 th annual spring workshop of the HUPO-proteomics standards initiative - April 23-25, 2007. Ecole Nationale Supérieure (ENS), Lyon, France. *Proteomics*, 7, 3436-3440

2008

Braconi Quintaje, S. & Orchard, S. (2008). Completion of the annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: One small step in manual annotation, one giant step for full comprehension of genomes. *Mol. Cell. Proteomics*, 7, 1409-1419

Chiang, T., *et al*. (2008). Rintact: Enabling computational analysis of molecular interaction data from the IntAct repository. *Bioinformatics*, 24, 1100-1101

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

Field, D., *et al*. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol*., 26, 541-547

Hamacher, M., *et al*. (2008). The HUPO brain proteome project wish list - Summary of the 9th HUPO BPP workshop: 9-10 January 2008, Barbados. *Proteomics*, 8, 2160-2164

Hermjakob, H. (2008). EBI proteomics services. In *Lecture Notes in Computer Science*, 207

Jenkinson, A.M., *et al*. (2008). Integrating biological data - The Distributed Annotation System. *BMC Bioinformatics*, 9, Article S3

Jones, P. & Cote, R. (2008). The PRIDE Proteomics Identifications Database: Data Submission, Query, and Dataset Comparison. *Methods Mol. Biol*., 484, 287-303

ENFIN, Transfog, BioSapiens and FELICS projects, we contribute to the development and application of data integration infrastructure based on DAS and XML technologies.

The Dasty protein DAS client (www.ebi.ac.uk/dasty; Jimenez *et al.*, 2008) provides an integrative view of protein features from currently more than 60 annotation servers. In collaboration with the Thornton group and the BioSapiens and GO consortia, we have developed a standardised feature ontology, facilitating the consistent display and analysis of protein feature annotation from multiple sources.

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 the 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 has been achieved in the domain of protein identifications, with three published MIAPE modules and the release of the mzML format for mass spectrometry data representation. For 2009, we plan to build on these standards and their implementation in PRIDE and IntAct, and initiate a regular exchange of proteomics data with international collaborators in the ProteomExchange and IMEx consortia (imex.sf.net).

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 the Druggability Portal, where we explore drug–target interaction data. We also hope to achieve this in the context of closer integration between IntAct and Reactome, supplementing canonical pathway information with relevant molecular interaction data.

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.

Jones, P., *et al*. (2008). PRIDE: New developments and new datasets. *Nucleic Acids Res*., 36, D878-D883

Klie, S., *et al*. (2008). Analyzing large-scale proteomics projects with latent semantic indexing. *Journal of Proteome Research*, 7, 182-191

Martens, L., *et al*. (2008). Using the Proteomics Identifications Database (PRIDE). *Current protocols in bioinformatics,* Andreas D. Baxevanis *et al*. (editoral board), Chapter 13, 8

Martens, L., Palazzi, L.M. & Hermjakob, H. (2008). Data standards and controlled vocabularies for proteomics. *Methods Mol. Biol*., 484, 279-286

Michaut, M., *et al*. (2008). InteroPORC: Automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24, 1625-1631

Mons, B., *et al*. (2008). Calling on a million minds for community annotation in WikiProteins. *Genome Biol*., 9, Article R89

Montecchi-Palazzi, L., *et al*. (2008). The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol*., 26, 864-866

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

Mueller, M., *et al*. (2008). Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. *Proteomics*, 8, 1138-1148

Orchard, S. & Hermjakob, H. (2008). The HUPO proteomics standards initiative - Easing communication and minimizing data loss in a changing world. *Brief. Bioinform*., 9, 166-173

Publications continued from left

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

Orchard, S., *et al*. (2008). 6th HUPO annual world congress - Proteomics standards initiative workshop: 6-10 October 2007, Seoul, Korea. *Proteomics*, 8, 1331-1333

Quintaje, S.B. & Orchard, S. (2008). The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: One small step in manual annotation, one giant leap for full comprehension of genomes. *Molecular and Cellular Proteomics*, 7, 1409-1419

Reisinger, F., *et al*. (2008). ENFIN - An integrative structure for systems biology. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in *Bioinformatics*), 132-143

Siepen, J.A., *et al*. (2008). ISPIDER Central: an integrated database web-server for proteomics. *Nucleic Acids Res*., 36, W485-490

Taylor, C.F., *et al*. (2008b). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nat. Biotechnol*., 26, 889-896

Tharakan, R., *et al*. (2008). OMSSAGUI: An open-source user interface component to configure and run the OMSSA search engine. *Proteomics*, 8, 2376-2378

Tsesmetzis, N., *et al*. (2008). *Arabidopsis* reactome: A foundation knowledgebase for plant systems biology. *Plant Cell*, 20, 1426-1436

Vastrik, I. (2008). Installing a local copy of the Reactome Web site and database. *Current protocols in bioinformatics,* Andreas D. Baxevanis *et al*. (editoral board), Chapter 9

Publications continued from left

Other EMBL publications
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*. (2008a). 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

Deutsch, E. (2008). mzML: a single, unifying data format for mass spectrometer output. *Proteomics*, 8, 2776-2777

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

The InterPro Team

INTRODUCTION

The InterPro team coordinates the InterPro and CluSTr projects and develops the software used by the Gene Ontology Annotation (GOA) group 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, Gene3D and PANTHER.

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, including table and graphical views and the InterPro Domain Architectures view.

InterPro has a number of important applications, including the automatic annotation of proteins for UniProtKB/TrEMBL and genome annotation projects. InterPro is used by the Ensembl and Integr8 databases 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 complete genomes in the Integr8 database.

Members of the team also develop and maintain the protein2GO tool used by the Gene Ontology Annotation (GOA) team and the QuickGO browser.

INTERPRO

The InterPro project (Hunter *et al.*, 2008; http://www.ebi.ac.uk/interpro) aims to provide an integrated resource for protein families, domains and functional sites. InterPro includes data from ten member databases, and continues to grow along with its members. The resource continues to expand and 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. InterPro currently integrates data from Pfam, Prints, PROSITE, ProDom, SMART, TIGRFAMs, PIRSF, PANTHER and the structure-based resources SUPERFAMILY (based on SCOP superfamilies) and Gene3D (models of CATH superfamilies).

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 two types of relationships in InterPro: 'parent/child', which shows family relationships, and 'contains/found in', which displays domain composition. 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

Sarah Hunter

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

Team Members

Coordinator
David Lonsdale
(Annotation Coordinator)

Senior Scientific Database Curators
Jennifer McDowall

Scientific Database Curators
Louise Daugherty

Bioinformatician
Craig McAnulla

Senior Software Engineers
Anthony Quinn
John Maslen
Manjula Thimma*
Phil Jones*

Software Engineers
David Binns
Ujjwal Das

Team Secretary
Kerry Smith

** Indicates part of the year only*

Publications

2008

Dunn, M.J., *et al*. (2008). EuPA achieves visibility - An activity report on the first three years. *Journal of Proteomics*, 71, 11-18

Hunter, S., *et al*. (2008). InterPro: the integrative protein signature database. *Nucleic Acids Res*., doi: 10.1093/nar/gkn785

Jones, P., *et al*. (2008). PRIDE: New developments and new datasets. *Nucleic Acids Res.*, 36, D878-D883

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

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, new 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 'parent/child' and 'contains/found in' relationships for matched entries. A standalone version of InterProScan is available for download for users who require privacy or bulk searches; users can also submit searches programmatically via web services. The latest version of InterProScan, 4.4, was released in September 2008, and a new version, with new Java-based architecture will be released in 2009.

InterPro database

The number of entries and coverage of proteins by InterPro continues to grow. The latest release of the database (18.0) contains over 16,500 entries. The growth of the database is shown in Figure 1. In its infancy, InterPro covered around 66% of all proteins in UniProtKB, and this has increased to 95.6% for UniProtKB/Swiss-Prot, 78.8% for UniProtKB/TrEMBL, and approximately 80% for UniProtKB (Swiss-Prot and TrEMBL). In the last year InterPro released versions 16.1, 17.0 and 18.0. InterPro data can now be accessed programmatically via web services to retrieve specific information within InterPro entries and protein matches.

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. These are listed below:

- July 2008: Humboldt University, USA. Bioinformatics Roadshow
- July 2008: EBI hands-on training: Programmatic Access of Proteomic Resources
- July 2008: Toronto, Canada. ISMB 2008 (demo)
- June 2008: EBI hands-on training: Patterns, Similarities and Differences in Biological Data

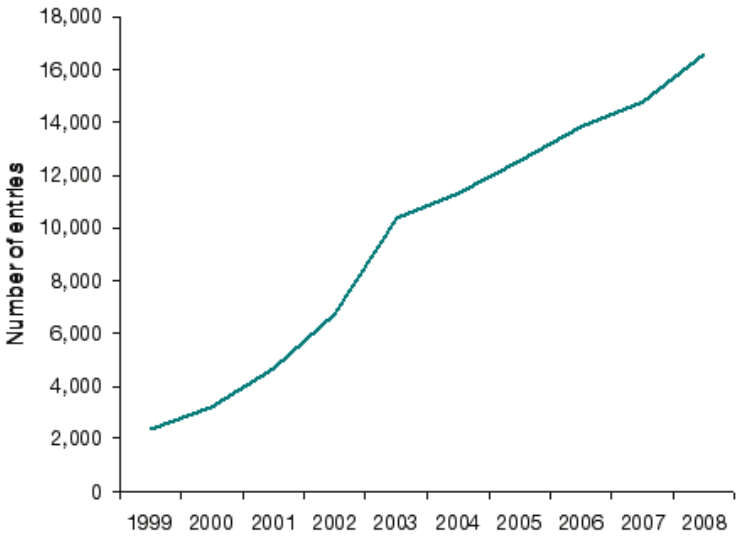


Figure 1. *Chart of the growth of InterPro showing how the number of entries is increasing each year.*

- June 2008: Trieste, Italy. EBI/NCBI Roadshow
- May 2008: EBI hands-on training: Protein Structural Annotation
- May 2008: EBI. 8th BioSapiens European School of Bioinformatics
- May 2008: Geneva/Bern, Switzerland. Bioinformatics Workshops
- April 2008: Poitiers, France. Bioinformatics Roadshow
- March 2008: EBI. Masters Open Day
- March 2008: London, UK. NIMR Introduction to Proteomics
- March 2008: EBI. Bioinformatics for Immunologists
- February 2008: Australia Lorne Conference on Protein Structure and Function (demo)
- February 2008: EBI hands-on training: Proteins to Proteomes
- February 2008: EBI. Membrane Workshop
- January 2008: Glasgow, UK. Scottish Bioinformatics Forum
- January 2008: San Diego, USA. Plant and Animal Genome XVI Workshop (workshop)
- November 2007: Aachen, Germany. Computational Proteomics Workshop (workshop)

CLUSTR

The CluSTr database (Petryszak *et al.*, 2005; http://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:

- Integr8 (http://www.ebi.ac.uk/integr8/): proteome analysis pages with CluSTr statistics; orthologue and paralogue predictions based on similarity data in CluSTr;
- UniProt (http://www.uniprot.org/): mapping UniProtKB proteins to clusters to which they belong;
- CluSTr search facility (http://www.ebi.ac.uk/clustr);
- InterPro (http://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:

- 6,959,413 sequences from UniProt Knowledgebase release 14.2;
- 303,139 sequences from IPI;
- 3.5 billion similarities, with pairwise alignments generated on-the-fly;
- 13,808,133 clusters;

- clustering for 793 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 homology predictions to Integr8 projects and best reciprocal hit orthology predictions for all species in CluSTr.

During 2008, bulk downloads of CluSTr data were made by the following institutions to be used in their own research: Wellcome Trust Sanger Institute, UK; University of Reading, UK; Hebrew University of Jerusalem, Israel; Helwan University, Egypt and University of Toronto, Canada.

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. To be able to cope with the increasing number of new signatures being developed by InterPro consortium members, we are in the process of improving our internal curation tools to aid rapid signature annotation and integration.

Future plans for CluSTr include use of its coverage of unique sequence sets to identify potential conserved protein families for new InterPro signature building. As a consequence, it is likely we will have to improve the web services available to access the data.

Chemoinformatics and Metabolism

Christoph Steinbeck

PhD 1995, Rheinische Friedrich-Wilhelm-Universitaet, 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-Universitaet, Jena, 2003.
Head of Research Group for Molecular Informatics, Cologne University Bioinformatics Center (CUBIC), Cologne, 2002-2007.
Lecturer in Chemoinformatics, University of Tuebingen, 2007.
Team Leader Chemoinformatics and Metabolism at EMBL-EBI since 2008.

Team Members

Coordinators
Paula de Matos*

Software Engineers
Rafael Alcantara
Janna Hastings
Stefan Kuhn*

Senior Software Engineers
Mark Rijnbeek*

Scientific Database Curators
Marcus Ennis
Inma Spiteri*

Postdoctoral Fellow
Gilleain Torrance*

Visitors
Kirill Degtyarenko
Alan McNaught

** Indicates part of the year only*

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 further 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

Marcus Ennis, Kirill Degtyarenko, Paula de Matos, Janna Hastings, Alan McNaught, Inma Spiteri

The database of Chemical Entities of Biological Interest (ChEBI) 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 MSDchem 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 http://obo.sourceforge.net/.

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 http://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; http://www.iupac.org/inchi/) and InChIKey are automatically generated. ChEBI release 49 (September 2008) comprises 15,833 annotated entries containing 14,307 IUPAC names, 41,513 synonyms, 13,741 chemical structures and 15,071 registry numbers. In addition ChEBI has over three million cross-references to other bioinformatics resources.

Significant developments in 2008

Drug annotation: in 2008 ChEBI increased its annotation of pharmaceuticals due to requests from the EBI industry partners. This resulted in new terminology being added in the form of brand names and International Nonproprietary Names (INN). NC-IUPHAR terminology has also been used to annotate relevant entries.

Publications

2007
Kuhn, S., *et al*. (2007). Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. *J. Chem. Inf. Model*, 47, 2015-2034

Willighagen, E.L., *et al*. (2007). Userscripts for the life sciences. *BMC Bioinformatics*, 8, 487

2008
Blinov, K.A., *et al*. (2008). Performance validation of neural network based (13)c NMR prediction using a publicly available data source. *J. Chem. Inf. Model*, 48, 550-555

Other publications
Steinbeck, C. (2001). SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *J. Chem. Inf. Com. Sci.*, 41, 1500-1507

Steinbeck, C., Krause, S. & Kuhn, S. (2003). NMRShiftDB - Constructing a free chemical information system with open-source components. *J. Chem. Inf. Com. Sci.*, 43, 1733-1739

Steinbeck, C. & Kuhn, S. (2004). NMRShiftDB -- compound identification and structure elucidation support through a free community-build web database. *Phytochemistry*, 65, 2711-2717

Automatically generated cross-references: in addition to its manually annotated cross-references to other small molecule databases, ChEBI provides a chemistry-centric view on a number of biologically relevant resources by automatically linking to databases using ChEBI. The BioModels and ArrayExpress databases have been included into the automatic linking of ChEBI entities.

Patent annotation: in collaboration with the European Patent Office, patent cross-references and terminology found in patents have been manually annotated in ChEBI. This allows users to search for chemicals based on patent identifiers. In addition we introduced language qualifiers to aid text mining projects, mining patents in various languages; German, French, Latin and Spanish language qualifiers were introduced.

Chemical structure search: chemical structure searching facilities were introduced into ChEBI allowing users to perform chemical structure queries searching for a substructure, a similar structure or an identical structure. These functions were implemented using the Chemistry Development Kit (CDK) and the MarvinSketch applet developed by ChemAxon. In addition, the functionality to narrow down a search based on datasets was also implemented.

INTENZ – THE RELATIONAL ENZYME DATABASE AND NOMENCLATURE RESOURCE

Rafael Alcantara, 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) provides a complete, freely available database focused on enzyme nomenclature approved by the NC-IUBMB, combined with additional information from the ENZYME database (http://www.expasy.org/enzyme/). Currently, IntEnz contains 4,111 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 the EBI. The IntEnz entries contain cross-references to various resources such as BRENDA, GO, KEGG, MEROPS, PDB, UM-BDD, NIST Thermodynamics of Enzyme-Catalysed Reactions and the Catalytic Site Atlas. IntEnz entries have cross-references to their corresponding proteins via UniProtKB (http://www.uniprot.org/). IntEnz also contains literature references to relevant publications. Biochemical compound names from IntEnz are used in the creation of the dictionary of ChEBI (http://www.ebi.ac.uk/chebi/).

IntEnz is available at http://www.ebi.ac.uk/intenz/ where EC numbers can be searched or browsed via the EC number hierarchy. IntEnz is freely available for download in XML format and flat file format from the EBI FTP site.

Significant developments in 2008

Cofactors annotated with ChEBI entities: previously a vocabulary of cofactors was maintained in IntEnz in a free text format. However this meant that a chemical vocabulary was being maintained unnecessarily as the same terminology could be found in the ChEBI database. In 2008 all cofactors were linked to ChEBI ensuring that only one set of terminology is maintained.

Reaction annotation: reactions in IntEnz have been stored as free text without links to compound information in the reactants. The IntEnz team has attempted to address this problem by creating a new database with reaction annotations linked to the ChEBI database. This new reaction database, named Rhea, will contain chemically balanced reactions with indicated directionality and have all reaction participants linked to ChEBI. More information on this can be found at the Rhea homepage (http://www.ebi.ac.uk/rhea).

CDK-TAVERNA – A WORKFLOW ENGINE FOR CHEMICAL INFORMATICS

Thomas Kuhn, Gesellschaft für naturwissenschaftliche Informatik

The recent release of large open access chemistry databases into the public domain generates a demand for flexible tools to process the data therein and discover new knowledge. To support open drug discovery and open notebook science on top of these data resources, it is desirable for the processing tools to be open source and available for everyone. Here we describe a combination of the workflow engine Taverna and our chemoinformatics library the Chemistry Development Kit

(CDK), resulting in an open source workflow solution to attack chemoinformatics problems. We have implemented more than 160 different workers to handle specific chemoinformatics tasks, allowing researchers to create chemoinformatics workflows for the processing, curation, conversion and analysis of large datasets in a Lego™-like manner (see Table 1 and Figure 1).

CDK-Taverna has the potential to be of use in a number of ongoing projects at EBI, for example to support the curation workflow in ChEMBL, EBI's new resource of drug discovery data funded by the Wellcome Trust.

COMPUTER-ASSISTED STRUCTURE ELUCIDATION AND PREDICTION OF NMR SPECTRA

Stefan Kuhn, Gilleain Torrance

Current efforts in metabolomics, such as the Human Metabolome Project, collect structures of biological metabolites as well as data for their characterisation, such as spectra for identification of substances and measurements of their concentration. Despite these efforts, only a fraction of the existing metabolites and their spectral fingerprints are known. Computer-Assisted Structure Elucidation (CASE) of biological metabolites will be an important tool to overcome this lack of knowledge (Figure 2). Indispensable for CASE are modules to predict spectra for hypothetical structures. In the course of this research, we evaluated different statistical and machine learning methods to perform predictions of proton NMR spectra based on data from our open database NMRShiftDB. Our best result with a mean absolute error of 0.18ppm was achieved for the prediction of proton NMR shifts ranging from 0 to 11~ppm. Random Forest, J48 decision tree and Support Vector Machines achieved similar overall errors. HOSE codes – being a notably simple method – achieved a comparatively good result of 0.17~ppm mean absolute error (see Figure 3).

Table 1. *Classification and count of workers implemented in CDK-Taverna*

Grouped worker	No. workers
File I/O	15
Database I/O	7
Molecular descriptors	42
Atom descriptors	27
Bond descriptors	6
ART2A classificator	10
SimpleKMean/EM clusterer (Weka)	3
SMILE tools	2
Inchi Parser	2
Miscellaneous	50

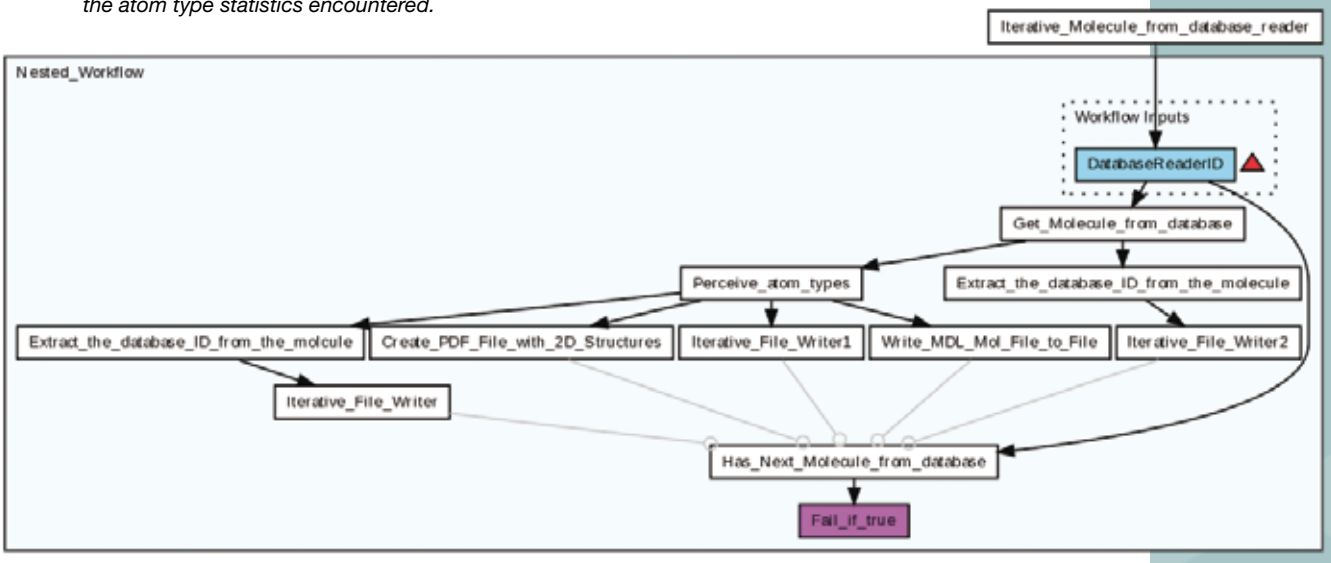


Figure 1. *Curating chemical databases such as the new chemogenomics resource at the EBI may involve plausibility checks by so-called atom typing, where nodes in a molecular graph are classified based on their molecular environment. The workflow shown here iterates over a database of molecules, performs a CDK-based atom typing and produces a report on the atom type statistics encountered.*

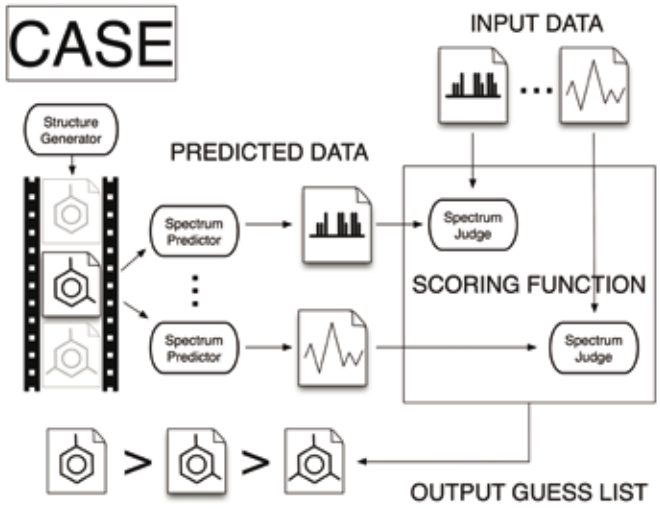


Figure 2. *Computer-Assisted Structure Elucidation uses a structure generation engine to produce chemical spaces based on boundary conditions such as the gross formula of the unknown compound, determined for instance by mass spectrometry. These chemical spaces are then crawled and candidate structures in them inspected for fitness by comparing predicted and measured properties such as NMR spectra. Based on calculated fitness values, a ranking is presented to the user.*

NMR prediction methods applied in the course of this work delivered precise predictions which can serve as a building block for CASE of biological metabolites. Here we aim to create a state-of-the-art set of algorithms for CASE for metabolome investigations. We have previously shown (Steinbeck, 2001, Steinbeck & Kuhn, 2004) that stochastic structure generation schemes (simulated annealing and genetic algorithms) are capable of elucidating the chemical constitution of metabolites of up to 32 heavy (non-hydrogen) atoms based on the 1D and 2D NMR experiments. The success and speed of the structure generation runs depend on the choice of the starting structure(s). Instead of starting with a random population, we will create a large-scale resource of predicted NMR spectra based on all publicly available organic molecules (PubChem, ChemSpider, ZINC, etc.). A starting population will be selected based on a spectrum-similarity search using existing technology developed in our group (NMRShiftDB).

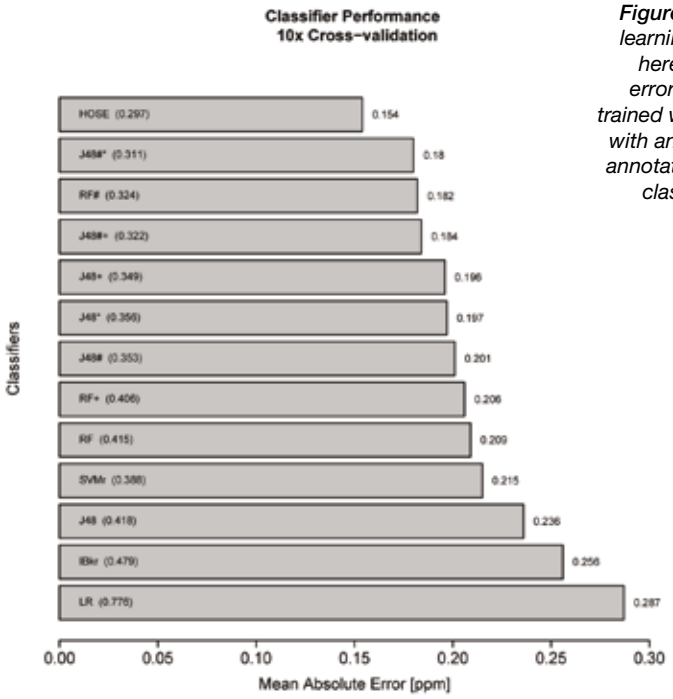


Figure 3. *Mean absolute error of machine learning methods (classifiers) investigated here for proton prediction, the standard error (SE) is given in brackets. Classifiers trained with selected features are annotated with an additional #. Bagged classifiers are annotated with an additional * and boosted classifiers carry an additional + symbol.*

PERFORMANCE VALIDATION OF NEURAL NETWORK-BASED ^{13}C NMR PREDICTION WITH NMRSHIFTDB

In 2003, we published the first open access database of NMR spectra of organic molecules (http://www.nmrshiftdb.org/) to open up resources for spectrum prediction in computer-assisted structure elucidation (Steinbeck, Krause & Kuhn, 2003). In collaboration with partners at ACD/Labs, we have now worked on the validation of the performance of a neural network based ^{13}C NMR prediction algorithm, which was trained based on their database ACD/CNMR DB, using a test set available from NMRShiftDB (Blinov *et al.*, 2008). The validation was performed using a version of NMRShiftDB containing approximately 214,000 chemical shifts as well as for two subsets of the database to compare performance when overlap with the training set is taken into account. The first subset contained approximately 93,000 chemical shifts that were absent from the ACD/CNMR DB (the 'excluded shift set' used for training of the neural network and the ACD/CNMR prediction algorithm) while the second contained approximately 121,000 shifts that were present in the ACD/CNMR DB training set (the 'included shift set'). This work has shown that the mean error between experimental and predicted shifts for the entire NMRShiftDB is 1.59ppm, while the mean deviation for the subset with included shifts is 1.47ppm and 1.74ppm for excluded shifts. Since similar work has been reported online for another algorithm, we compared the results with the errors determined using Robien's CNMR Neural Network Predictor using the entire NMRShiftDB for program validation. This work may serve as a showcase for the importance of open source and open access principles in chemistry.

FUTURE PROJECTS AND GOALS

The recently acquired resource of large-scale drug activity data at the EBI creates exciting new opportunities both on the research and service side (http://www.ebi.ac.uk/Information/News/pdf/Press23July08.pdf). Our team has started to create an open source chemical search engine for the new resource, which will be the first open source chemistry search engine for the widely used Oracle™ Database system. A combination of the new chemogenomics data and the Chemistry Development Kit will allow us to create open structure-activity models and to assist efforts in wet lab screening in areas such as library design.

In January 2009, the Chemoinformatics and Metabolism team, under the auspices of the ELIXIR Standards workpackage, will organise a workshop of leading metabolomics researchers worldwide to define the scope of a new metabolomics resource at the EBI.

On the service side, ensuring a sustainable growth for the ChEBI database will be the focus of our attention. The number of marketed and developed drugs in the world drug index alone currently amounts to more than 80,000 compounds. Assuming only a handful of metabolites are produced by organisms upon application of these drugs, the task ahead takes shape. Not only does this task 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. Last but not least, 2009 will reveal the EBI's solution on how to integrate the chemogenomic data with existing chemical resources at the EBI.

Database Research and Development Group Activities

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.

Team Members

Senior Software Engineer
Rasko Leinonen

Software Engineers
Lawrence Bower
Fehmi Demiralp
Mikyung Jang
Szilveszter Juhas
Quan Lin
Dariusz Lorenc
Siamak Sobhay

Database Administrators and Consultants/ Developers
Mike Donnelly
Giuseppe di Martino
Beat Ramseier
Pieter van Rensburg
Mark Rynbeek

INTRODUCTION

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

As the Database Application team, the team's main functions were to provide software/tool development and maintenance support for the EMBL Nucleotide Sequence Database, and Oracle database support for all the Oracle databases in the PANDA group. The team also coordinated hardware resources with the EBI's Systems and Networking team (p33) to ensure the hardware requirements for the projects within PANDA were met.

The group'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.

ACTIVITY UPDATE IN 2008

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 require upgrading in order to realise the full potential of data-rich resources, and optimise the usage of our human and hardware resources.

This year, our main focus has been to analyse the uniqueness of bioinformatics datasets, and conduct research into possible solutions to large dataset distribution problems, especially over slow networks.

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, 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 major new bottleneck. Also, unclustered changes in files, which are common in bioinformatics databases, can result in full data transfer instead of deltas. Some peer-to-peer models suggested the de-centralisation of the source data by splitting it into smaller subsets accessed from geographically distributed computers. This solution shortens the network latency by parallel downloading, possibly on a faster network than the original sources. Some scientific applications of this model have been implemented within a clearly defined and closed user community. Currently, it is unconceivable to distribute large biological datasets in this fashion, while ensuring data integrity and sufficient coverage. However, this parallel download model has the potential to be part of the more complex integrated solution in the long run. Other models also suggest using the data and storage grid to distribute the data. Since these do not directly address the performance problem over slow networks, and grid technology is still in its early stage, their possible application in large dataset distribution requires further development in the future

In a preliminary analysis of EBI database files from different database releases, some useful characteristics were found:

- an entry in a typical data file uses common atomic data units, 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

Publications

2008
Cochrane, G., *et al*. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 36, D5-D12

database files. Changes, such as deletions, updates and new entry insertions, occur on the level of the entries in a file;

- 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 of data need to be transferred for a database with total size of 600GB;
- when these changed entries were further analysed, a significant percentage were due to updates, not new insertions. The updated entries count 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 result were obtained when analysing Ensembl datasets.

These unique features tell us two things:

- delta compression has a significant role to play in dramatically reducing the data that needs to be transferred;
- delta compression is feasible. To use entry comparision between database releases as the first level of delta-ing, huge computation power and time can be saved by removing those unchanged entries from the fingerprint computing and matching processes. It will also make the later block-size-based computing more accurate and manageable.

Network latency can be further reduced by querying into a RDBMS-managed database in the run-time, which contains pre-computed information essential to finding the deltas, and transporting the data and instructions for the client to reconstruct the new version of the files.

An algorithm using these principles is in the early stages of development.

FUTURE PROJECTS AND GOALS

The development of the new delta compression algorithm will be continued. The goal is to develop prototype software, along with a data model and repository to store the pre-computed metadata for the algorithm, to test the key features of this new distribution system. The later stage of this work will be undertaken in collaboration with the Systems and Networking team and database projects requiring the distribution of large data files, as well as with external partners in countries with slow network connections to the EBI. This will allow us to benchmark the network saving for distributing large datasets using delta compression.

The data distribution problem is not unique to the bioinformatics community. It is also a problem for other domains, such as particle physics and earth science. A workshop is planned for 2009 to share knowledge among scientists across specialisations. Another focus of this workshop is to bring together Chinese bioinformaticians and network backbone administrators to explore the network optimisation possibilities between China and the EBI.

We will continue to seek external funding for the project's long term goal – to have a stable and well-maintained distribution system for the EBI's large databases.

The GO Editorial Office

Midori Harris

GO Editor

Team Members

Curation Coordinator
Jane Lomax

Scientific Database Curators
Jennifer Deegan
Amelia Ireland

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 the 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 parent–child relationships between terms are specified. A child term is a subset of its parent(s). The GO vocabularies have long defined two semantic relationships between parent and child terms: *is_a*

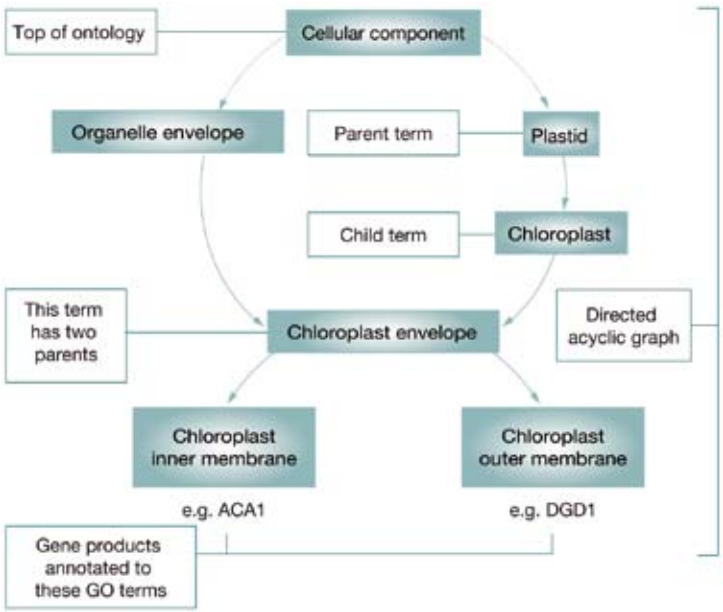


Figure 1. *GO terms are organised in directed acyclic graphs (DAGs) – hierarchical structures in which any 'child' (more specialised term) can have many 'parents' (less specialised terms). For example, the cellular component term chloroplast envelope has two parents, reflecting the fact that it is a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

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 discoideum*)
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

2008

Harris, M.A. (2008). Developing an ontology. *Methods in molecular biology* (Clifton, N.J.), 452, 111-124

Blake, J.A. & Harris, M. (2008). The Gene Ontology (GO) Project: Structured Vocabularies for Molecular Biology and Their Application to Genome and Expression Analysis. *Current Protocols in Bioinformatics*, Chapter 7, Unit 7.2

Harris M.A. & Feltrin E. (2008). Browsing and searching Gene Ontology resources using AmiGO. *EMBnet News*, 14, 17-21

The Gene Ontology Consortium (2008). The Gene Ontology (GO) Project in 2008. *Nucleic Acids Res*., 36 (Database issue), D440-D444

Other publications

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 implementation. *Genome Res*., 11, 1425-1433

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

Biological Process Terms	15018
Molecular Function Terms	8220
Cellular Component Terms	2176
Sequence Ontology Terms	1652

Table 1. *Current status of the GO vocabularies (as of 1 September 2008).*

and *part_of*. The *is_a* relationship means that a term is a subclass of its parent; part of may mean 'physically part of' (as in the cellular component ontology) or 'subprocess of' (as in the biological process ontology). New relationships representing biological regulation have recently been added, as described below. 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;
- 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

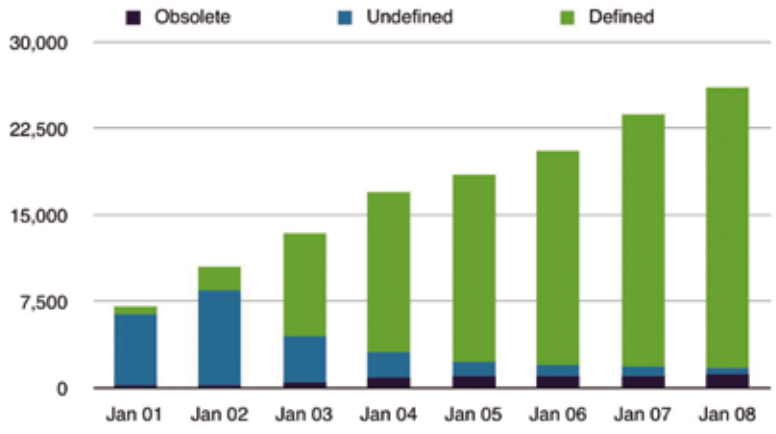


Figure 2. *Annual growth of the GO vocabularies, reflecting the addition of new terms and of definitions for existing terms (the latter mainly in 2003). The graph shows the number of terms in the Molecular Function, Biological Process, and Cellular Component ontologies combined. Obsolete terms are those that have been removed from active use.*

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.

The most significant change introduced in GO in 2008 affects both biological and logical aspects of the ontologies: three new relationship types have been introduced to represent biological regulation, namely *regulates*, *negatively_regulates*, and *positively_regulates*. Whereas regulatory processes have been represented in the past as *part_of* the processes they regulate, the new relationship types better reflect the fact that regulation is not necessarily integral to the regulated processes. Accompanying the new relationship types, new high-level regulation terms have been added to represent processes that regulate the activity of gene products and processes that regulate measurable biological attributes. Regulation of molecular function terms have been aligned with the corresponding terms in the molecular function ontology. In addition to capturing the biological nature of regulation more accurately, the *regulates* relationships will allow GO to make cross-products between regulatory processes and regulated processes, improving computability and supporting more sophisticated tool development.

Biological topics of interest in 2008 include:

- improvements to process and function terms relevant to signal transduction have begun;
- electron transport terms have been revised;
- process terms describing the organisation of subcellular structures are being improved;
- function, process and component terms for transcription 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 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 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 in 2008 on signal transduction, transcription, and other topics will therefore continue in 2009. The new *regulates* relationships will be used to create the first links between the biological process and molecular function ontologies, with additional types of links to follow. Work on recasting many complex process terms as explicit cross-products with orthogonal ontologies such as the ChEBI ontology and the cell ontology will also continue.

The Microarray Informatics Team

Alvis Bramza

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

Team Members

Technical Team Leader
Ugis Sarkans

Coordinators
Misha Kapushesky
Maria Krestyaninova
Helen Parkinson
Susanna-Assunta Sansone

Technical Coordiantor
Philippe Rocca-Serra

Software Developers
Tony Burdett*
Marco Brandizi
Eamonn Maguire
Sudeshna Guha Neogi
Nataliya Sklyar

Scientific Database Curators
Anna Farne
Ele Holloway
Margus Lukk
James Malone
Tim Rayner*
Eleanor Williams
Holly Zheng Bradley

Scientists
Richard Coulson
Johan Rung*
Gabriela Rustici
Chris Taylor

Bioinformaticians
Tomasz Adamusiak*
Natalja Kurbatova*
Pavel Kurnosov*
Chris Taylor

Postdoctoral Fellow
Daniel Schrober

PhD Students
Nils Gehlenborg
Katherine Lawler

Visitors
Juok Cho
Samuel Kaski
Rodrigo Santamaria Vicente
Aravind Venkatesan

INTRODUCTION

The Microarray Informatics team provide two major EBI resources:

- ArrayExpress archive of functional genomics data;
- ArrayExpress Atlas of Gene Expression.

We also work on:

- medical and translational bioinformatics;
- data standards and ontologies for systems biology and medical informatics;
- high-throughput data analysis, models for gene regulation networks and algorithm development for functional genomics and systems biology.

Training provision is also a major activity of the group.

Our team was among the first to use microarray data to study transcription regulation mechanisms on a genomic scale (Brazma *et al.*, 1998). In 1999 we realised the importance of standards in microarray data reporting (Brazma *et al.*, 2000, Brazma *et al.*, 2001) and began work to establish the ArrayExpress database. The database now holds data from ~200,000 assays, including deep sequencing based transcriptomics and epigenomics assays. The Atlas of Gene Expression builds on these data to allow the user to find genes differentially expressed in particular conditions or tissue types across multiple studies and technologies. Our PhD students focus mostly on analysing these data and building models for systems biology (e.g. Rustici *et al.*, 2004, Schlitt & Brazma, 2006). Integration of data across multiple platforms, including genotypes is among the latest activities of the team.

ARRAYEXPRESS

ArrayExpress development team – Ugis Sarkans, Mohammadreza Shojatalab, Niran Abeygunawardena, Hugo Berube, Karim Chine, Miroslaw Dylag, Ibrahim Emam, Nikolay Kolesnikov, Roby Mani, Ekaterina Pilicheva, Anjan Sharma, George Tsun-Po Yang – for more detail on this work see the report by Ugis Sarkans, p103

ArrayExpress Atlas – Misha Kapushesky, Pavel Kurnosov

ArrayExpress production – Helen Parkinson, Tomasz Adamusiak, Tony Burdett, Anna Farne, Ele Holloway, Margus Lukk, James Malone, Tim Rayner, Eleanor Williams, Holly Zheng

Important contributions were also made by other team members (acknowledged in the text)

ArrayExpress developments over the past two years are described in a publication to appear in *Nucleic Acids Research* database issue 2009; here we present a summary concentrating on the last twelve months. The ArrayExpress Repository is one of the three recommended international repositories to archive publication-related functional genomics data. It was launched in 2002, and currently it contains data from over 6,000 experiments (studies) and ~200,000 assays. The ArrayExpress Warehouse of gene expression profiles was added in 2005, and currently it includes data from over 600 experiments and 20,000 samples. There have been three major developments in the last two years, which have mostly come to a fruition over the last twelve months. First, the new ArrayExpress Atlas of Gene Expression provides experimental condition-based queries and an overview of gene expression across multiple experiments. Second, we have established a procedure to import data from the Gene Expression Omnibus (GEO) and almost 4,000 GEO data series have been imported and released via the ArrayExpress interface. Third, ArrayExpress has started to accept data from ultra high-throughput sequencing (UHTS) experiments and the first datasets have been made public.

ArrayExpress Atlas of Gene Expression

The ArrayExpress Atlas of Gene Expression allows the user to query for condition-specific gene expression across multiple datasets. The user can query for a gene or a set of genes by name, synonym, Gene Ontology term or for a biological sample property or condition, e.g. tissue type, disease name, developmental stage, compound name or identifier. Combined queries for both genes and conditions are also possible. For example, the user can query for all 'DNA repair' genes up-regu-

Personal Assistant
Helen Crowe*
Lynn French*

** Indicates part of the year only*

lated in 'cancer'. The Atlas data content is produced using a meta-analytical approach which uses Bioconductor. Linear models are applied to experiments in the ArrayExpress Warehouse to calculate moderated *t*-statistics that simultaneously test the significance of differential expression in each condition versus the mean level of expression across all conditions for each gene. We then adjust the computed *p*-values and aggregate the computed multiple test statistics for interpretation. The high quality of annotation and curation of experiments in the ArrayExpress Warehouse allows the application of this generic method to data from multiple platforms.

Data import from Gene Expression Omnibus

ArrayExpress now integrates gene expression data produced on Affymetrix and Agilent array platforms from GEO. This allows users to view and search GEO and ArrayExpress data from a common interface and access the data in the standard MAGE-TAB format. GEO datasets are imported, the data files are then checked and free text information is text mined using Whatizit (a text mining tool developed by the Rebholz-Schuhmann group at the EBI) and a custom ArrayExpress dictionary derived from the NCI Thesaurus with local extensions for non-human terms. Imported and text-mined datasets are then curated and the annotations mapped to ontology terms. We apply Bioconductor quality metrics and perform scoring for MIAME compliance to decide which datasets are appropriate for integration into the ArrayExpress Data Warehouse/Atlas. Curated GEO data that satisfy both the quality metrics and MIAME compliance and have curated experimental factors are included in the ArrayExpress Data Warehouse and consequently made available via the Atlas.

New high-throughput sequencing and other supported data types

ArrayExpress accepts data generated on all array-based technologies, including gene expression, protein array, ChIP-chip and genotyping. More recently, data from transcriptomic and related applications of UHTS technologies such as Solexa (Illumina), and 454 (Roche) are also accepted. The first sequencing based datasets to be made public are the transcriptomics landscape for *S. pombe* and epigenomics analysis of the human genome.

The ArrayExpress Warehouse now includes gene expression profiles from *in situ* gene expression measurements, as well as other molecular measurement data from metabolomics and protein profiling technologies. Where *in situ* and array-based gene expression data are available for the same gene, these are displayed in the same view and links to the multi-species 4DXpress database of *in situ* gene expression are provided (Haudry *et al.*, 2008).

Other developments

In addition to integrating processed data from multiple array platforms in the ArrayExpress Atlas, we have also performed integration for each individual platform using a re-annotation, data quality assessment and re-normalisation approach (see below). All ArrayExpress data are now available for download in MAGE-TAB format. To aid bioinformaticians and other users interested in large-scale functional genomics analysis, a Bioconductor package called ArrayExpress has been developed in collaboration with the Huber group (EBI). The ArrayExpress Atlas, Repository and Warehouse have web service APIs enabling programmatic queries. ArrayExpress can also be queried along with all EBI core databases via the EBI general query interface 'EB-eye'. The ArrayExpress submission tools, MIAMExpress and Tab2MAGE, are undergoing continuous improvement to facilitate submissions of large experiments, to work with MAGE-TAB files, and to accept UHTS-based transcriptomics data.

To improve the ArrayExpress Atlas queries, we have developed an application ontology called the Experimental Factor Ontology (EFO) (see page 98). The EFO is deployed in the Atlas interface where queries can be expanded via the 'hierarchies'. For example, a query for the condition 'cancer' will also retrieve conditions 'sarcoma', 'carcinoma' and other cancers.

A Java-based framework for integrating the R/Bioconductor statistical package within a distributed computing context has been developed. Biocep is a multiclient-multiserver platform for high performance computing, and includes a fully-fledged workbench application with an integrated numerical spreadsheet, code editor, multi-window graphical display and help browser.

MEDICAL AND TRANSLATIONAL BIOINFORMATICS

The Nutrigenomics, Environmental Genomics and Toxigenomics (NET) Project

Susanna-Assunta Sansone, Philippe Rocca-Serra, Marco Brandizi, Eamonn Maguire, Daniel Schober, Nataliya Sklyar, Chris Taylor

The NET Project (http://www.ebi.ac.uk/net-project) includes a series of collaborative projects with nutrigenomics, environmental genomics and toxicogenomics communities. Susanna-Assunta Sansone and Philippe Rocca-Serra coordinate bioinformatics and information management in two European projects: Carcinogenomics and NuGO. A major activity of the group is the development of the Bio Investigation Index (http://www.ebi.ac.uk/bioinvindex), a new sample and experimental metadata database and infrastructure which aims at creating a common storage mechanism for metadata across several resources at the EBI. The first release is scheduled for December 2008. The new infrastructure also includes an emerging format for multi-platform based studies, ISA (Investigation-Study-Assay) TAB (Sansone *et al.*, 2007) for submission/import and download/export, and the ISAcreator, a tool for creating and editing ISA-TAB formatted files. In addition, converters between ISA-TAB and MAGE-TAB are work in progress, to allow ArrayExpress to accept data in both formats. Jointly with the ENGAGE/MolPAGE consortia and the EGA group we have customised the MAGE-TAB and ISA-TAB formats for representing population genomics studies (genotyping, functional genomics, phonotype data and metadata). We have also established strategic collaborations (NERC-NEBC and NCTR-FDA) to maximise data exchange and software interoperability.

Susanna-Assunta Sansone, Phillipe Rocca-Serra and Chris Taylor are also major contributors to standards development activities (see Standards and Ontologies section overleaf) and, having obtained BBSRC funds, have organised a series of workshops designed to advance the coordinated development of synergistic standards.

MUGEN and Gen2Phen

Holly Zheng Bradley, Helen Parkinson

MUGEN (www.mugen-noe.org) is an EU-funded Network of Excellence examining mouse models of immune disease. Our task is data capture and annotation, the development of ontologies for sample description and to integrate detailed mouse phenotypes with available human data. An meta-analysed set of Affymetrix data has been created and the next step will be to perform an orthologue/condition analysis with a human meta-analysed dataset.

Gen2Phen is an FP7 project which aims to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-To-Phenotype (G2P) data, and to link this system into other biomedical knowledge sources via genome browser functionality. Our major role in this project is the standardisation of formats and semantics. We are working with data providers EGA and the ENGAGE project and external locus-specific databases to integrate this data with EBI resources.

MolPAGE and ENGAGE

Maria Krestyaninova, Mikhail Gostev, Natalja Kurbatova, Sudeshna Guha Neogi, Johan Rung, Ugis Sarkans from the Microarray Software Development team

Maria Krestyaninova coordinates bioinformatics and information management in two genetic epidemiology projects – MolPAGE (www.molpge.org) and ENGAGE (www.euengage.org). The objective of the Molecular Phenotyping to Accelerate Genomic Epidemiology Consortium (MolPAGE) project is to find biomarkers that can identify individuals likely to suffer from diabetes and vascular disease in advance of the emergence of physical symptoms. The MolPAGE experiment repository now contains data from 22,400 multiomics assays, from which 3,600 have been loaded in the ArrayExpress Warehouse.

ENGAGE (European Network of Genomic and Genetic Epidemiology) aims at translating the wealth of data emerging from large-scale research efforts in genetic and genomic epidemiology into information relevant to future clinical applications. A System for Information Management in BioMedical Studies – SIMBioMS, which was originally developed for MolPAGE, was scaled up for population-wide studies and has been used to manage data for the ENGAGE project from the project's first month. Additionally, a new software system (Sample Availability system – SAIL) has been developed to assist the design of population genomics studies via indexing of sample availability in various cohorts. The

Publications

2007

Fiehn, O., *et al*. (2007). The metabolomics standards initiative (MSI). *Metabolomics*, 3, 175-178

Griffin, J.L., *et al*. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Mammalian/*in vivo* experiments. *Metabolomics*, 3, 179-188

Hancock, J.M., *et al*. (2007). Integration of mouse phenome data resources. *Mamm. Genome* 18, 157-163

Hardy, N.W. & Taylor, C.F. (2007). A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics*, 3, 243-248

Jones, A.R., *et al*. (2007). The Functional Genomics Experiment model (FuGE): An extensible framework for standards in functional genomics. *Nat. Biotechnol*., 25, 1127-1133

Morrison, N., *et al*. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Environmental context. *Metabolomics*, 3, 203-210

Sansone, S.A., *et al*. (2007). Metabolomics standards initiative: Ontology working group work in progress. *Metabolomics*, 3, 249-256

Schlitt, T. & Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8, Article 59

Smith, B., *et al*. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol*., 25, 1251-1255

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

Chalmers, I.W., *et al*. (2008). Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the Schistosoma mansoni venom allergen-like (SmVAL) gene family. *BMC Genomics*, 9, Article 89

Corpas, M., *et al*. (2008). Ten simple rules for organizing a scientific meeting. *PLoS Comput. Biol.*, 4, e1000080

Deutsch, E.W., *et al*. (2008). Minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE). *Nat. Biotechnol*., 26, 305-312

Field, D., *et al*. (2008). Meeting report: The fifth Genomic Standards Consortium (GSC) workshop. *OMICS A Journal of Integrative Biology*, 12, 109-113

Field, D., *et al*. (2008). Foreword to the special issue on the fifth Genomic Standards Consortium workshop. *OMICS A Journal of Integrative Biology*, 12, 99

Field, D., *et al*. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol*., 26, 541-547

Garrity, G.M., *et al*. (2008). Toward a standards-compliant genomic and metagenomic publication record. *OMICS A Journal of Integrative Biology*, 12, 157-160

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

Giegerich, R., *et al*. (2008). The BREW workshop series: A stimulating experience in PhD education. *Brief. Bioinform*., 9, 250-253

system has currently been released in beta version and is already being used in ENGAGE. In addition to being used in these two projects, a distribution version of the SIMBioMS software is available and six biomedical research laboratories are currently testing it for their local use.

We have also established a federated data management system based on SIMBioMS. This involves three different European nodes supporting over 50 user accounts from different groups across Europe and we are maintaining a stable flow of data from population multiomics studies from across Europe. This system is used in data submission to the core EBI repositories ArrayExpress and the European Genotyping Archive (EGA). ENGAGE submissions are the first non Wellcome Trust Case Control Consortium data submissions to the EGA.

Marine Genomics and Tara-Oceans

Maria Krestyaninova

Since May 2008, Maria Krestyaninova has been leading data management planning in the preparatory stage of a biodiversity project, Tara-Oceans, led by Eric Karsenti. Having studied the existing marine genomics and oceanographic data resources and carried out the preliminary analysis of the Tara-Oceans informatics requirements, we proposed a concept for the informatics infrastructure to collect and manage the data produced by expedition and by marine genomics centres worldwide.

STANDARDS AND ONTOLOGIES

In order to systematically annotate the data in ArrayExpress we have developed an application ontology called the Experimental Factor Ontology (EFO). Currently it contains over 1,000 terms including diseases, multi-species anatomy, compounds and cell type terms. EFO maps to several non-orthogonal ontologies, such as those for human anatomy, the disease ontology, the cell type ontology and the NCI Thesaurus. The EFO allows users to tune the ontology based on analysis of user queries and provision of annotation at an appropriate level of granularity for the database content. The EFO is available from the EBI Ontology Lookup Server – OLS (Côté *et al*., 2006) and is available in OBO and OWL formats (http://www.ebi.ac.uk/microarray-srv/efo/).

Susanna-Assunta Sansone and Phillipe Rocca-Serra have organised the first Open Biomedical Ontology (OBO) Foundry workshop at the EBI in July 2009, which will bring together over 30 biomedical ontology groups; followed by a hands-on workshop for the Ontology for Biomedical Investigation, both funded by the BBSRC. Chris Taylor and Susanna-Assunta Sansone organised the first Minimum Information about Biological and Biomedical Investigation (MIBBI) workshop, held at the EBI in April 2008, bringing together over 20 groups that defined minimal reporting requirements. The MIBBI paper was published in *Nature Biotechnology* in August 2008 (Taylor *et al*., 2008b). Lately, a grant award has been received from the BBSRC to continue supporting Chris Taylor on the MIBBI project.

Phillipe Rocca-Serra and Susanna-Assunta Sansone led the development of the ISA-TAB format (http://isatab.sf.net), a MAGE-TAB format generalisation which can be used for all technology platforms, and moreover for describing investigations where different platforms are used (e.g. transcriptomics and proteomics). The second ISA-TAB workshop was held at the EBI in June 2008; resulting in the preparation and publication of the ISA-TAB v1.0 specification. The third ISA-TAB hands-on workshop will be held at the EBI in December 2008.

In collaboration with the Huber group at the EBI we have developed a proposal for microarray data quality standards as a part of our participation in the EC-funded project EMERALD. This has resulted in the ArrayExpress Bioconductor package for processing data in MAGE-TAB format.

Several team members actively participate in various community activities to promote standards and data sharing. Helen Parkinson, Susanna-Assunta Sansone and Alvis Brazma are members of the Microarray and Gene Expression Data (MGED) Society Board of Directors (Helen Parkinson is the Board Secretary), while Maria Krestyaninova is a co-leader of the Information Curation and Information Technology group in the Canadian-funded Public Population Project in Genomics (P3G). We actively participated in organising the tenth annual MGED society meeting, MGED 11, Riva Del Garda, Italy in September 2008.

DATA ANALYSIS AND ALGORITHM DEVELOPMENT

Misha Kapushesky, Richard Coulson, Nils Gehlenborg, Katherine Lawler, Margus Lukk, Helen Parkinson, Gabriela Rustici, Holly Zeng Bradley

Over the past twelve months we have been working on ArrayExpress data meta-analysis, the use of hybrid systems for describing and analysing gene networks, and on using mathematical modelling for describing RNA degradation in *S. pombe*. Four manuscripts are in preparation, which we plan to submit to journals during November 2008 – February 2009.

Analysis of human gene expression in 5,372 samples representing 363 different biological conditions

Margus Lukk, Misha Kapushesky, Janne Nikkila (University of Helsinki), Helen Parkinson, Esko Ukkonen (University of Helsinki), Alvis Brazma

We have constructed a systematically annotated and consistently normalised human gene expression data matrix of 5,372 samples and ~18,500 genes using publicly available, high-quality raw data from a single high-density oligonucleotide array platform. The sample descriptions provided with the original datasets were converted into controlled vocabulary terms concisely describing the cell or tissue types, developmental stages and disease states, dividing all samples into 363 different biological groups. Information about the 163 laboratories which performed the assays was also included in the annotation. Analyses show that the biological signal in the data was clearly stronger than the laboratory effects. We analysed the relationships between expression profiles of different biological sample groups (see Figure 1) and determined the genes specifically expressed in particular biological conditions. Six major clusters were identified – the samples derived from blood or blood cells clustered together, as did the solid tumor samples, brain and brain parts, heart and muscle. Almost all

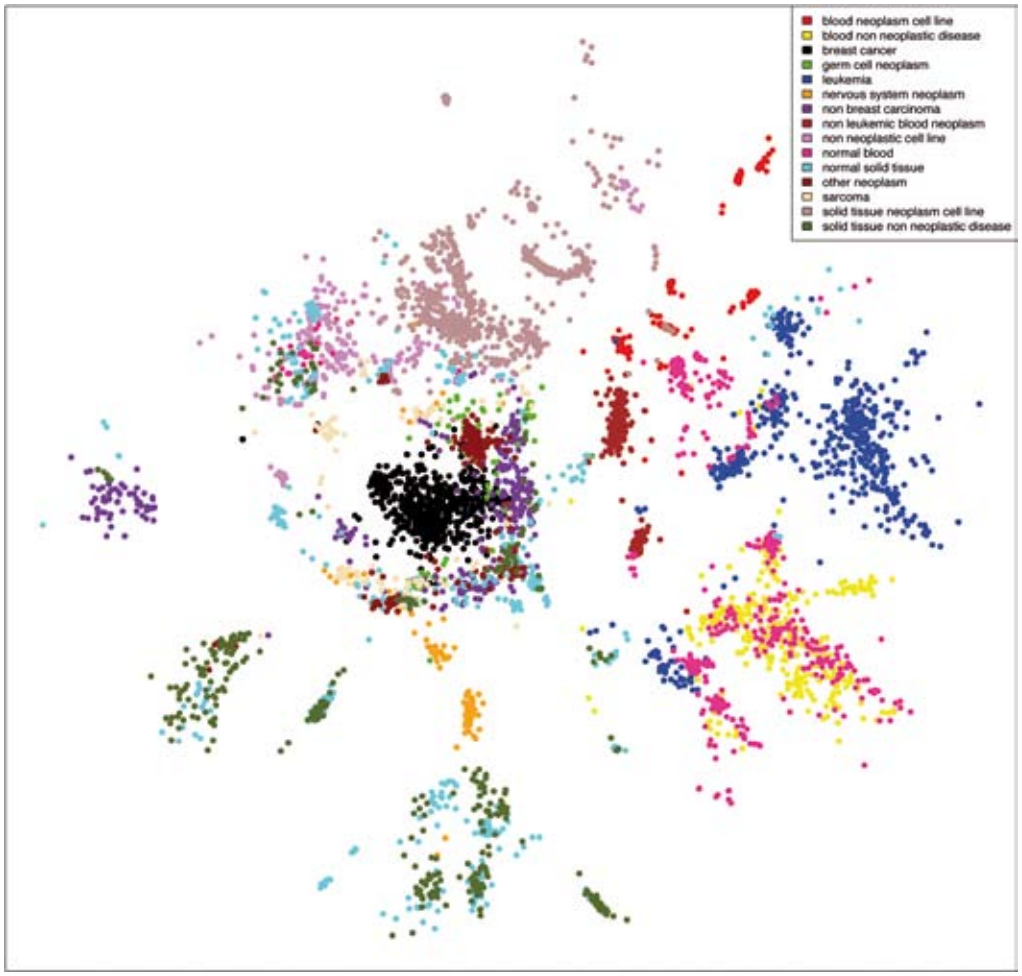


Figure 1. *Visualisation of relationships transcriptomes of ~5,300 human samples categorised in 15 biological classes using Neighbor Retrieval Visualizer (NeRV; Venna & Kaski, 2007) developed by our collaborators in Helsinki University of Technology.*

Haudry, Y., *et al*. (2008). 4DXpress: a database for cross-species expression pattern comparisons. *Nucleic Acids Res*., 36, D847-D853

Jen, C.H., *et al*. (2008). Signature Evaluation Tool (SET): A Java-based tool to evaluate and visualize the sample discrimination abilities of gene expression signatures. *BMC Bioinformatics*, 9, Article 58

Rustici, G., *et al*., (2008). Data storage and analysis in ArrayExpress and Expression Profiler. *Current protocols in bioinformatics*, Chapter 7, Unit 7.13

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

Spasic, I., *et al*. (2008). Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*, 9, S5

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

Taylor, C.F., *et al*. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nat. Biotechnol*., 26, 889-896

Other EMBL publications

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

Brazma, A., *et al*. (2000). One-stop shop for microarray data. *Nature*, 403, 699-700

Brazma, A., *et al*. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29, 365-371

Brazma, A. & Schlitt, T. (2003). *Genome Biol.*, 4, P5

Côté, R.G., *et al*. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 28, 97

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

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

cell line-based samples clustered together, rather than with the cell types they were derived from. A mixture of incompletely differentiated cells and adhesive tissues formed a separate cluster, which was most similar to non-neoplastic cell lines. We studied the distribution of gene expression levels across all 5,372 samples, as well as in particular sample groups and clusters, to determine genes expressed specifically in particular conditions. We studied gene ontology term over-representation in different groups of genes, for instance, genes that are highly variable overall are over-represented with gene ontology terms related to stress response. The least variable genes that are highly expressed in all samples are candidates for 'housekeeping' genes. All data, including raw and processed, and full sample annotation are publicly available from ArrayExpress.

A similar dataset has been prepared and is being analysed for mouse.

Finding condition-specific genes – lessons from integrating and exploring gene expression data from over 20,000 assays on various platforms

Misha Kaphushesky, Helen Parkinson, Wolfgang Huber, Alvis Brazma

Gene regulation pathways have been shown to be active simultaneously in numerous systems in living cells, functioning in different organs and under a diversity of conditions. It is hypothesised, and there exists strong evidence in individual cases, that the expression of some regulatory elements and targets of pathways is restricted to specific tissues or conditions, or is specifically required in certain tissues in developmental processes. We have developed an analytical framework where the wealth of high-throughput gene expression data can be examined cohesively in order to test this hypothesis and related questions.

Using a robust statistical framework (based on the limma package in Bioconductor), we tested the differential expression strength of genes across more than 2,000 conditions studied in over 20,000 assays in nine species (including human, mouse, rat, fruit fly and two yeasts). Furthermore, we developed a means to measure the confidence of specificity of expression of a given gene in a condition, including tissue, disease, cell type and developmental stage. An analysis of the distribution of condition-specific active genes, allowed us to elucidate genes whose expression patterns tend to be determined by few conditions as opposed to those expressed without significant condition association. The obtained global statistics for studying functionally related groups of genes (e.g. pathways) can be used to discover core subsets whose patterns of condition-specific expression are similar.

A simple mathematical model of the transcription regulation network in λ-phage reveals its dynamic properties and leads to experimentally testable hypothesis

Dace Ruklisa, Karlis Cerans (University of Latvia), Juris Viksna (University of Latvia), Thomas Schlitt (Kings College London), Alvis Brazma

We have developed a new mathematical model of the transcription regulation network in λ-phage and explored its properties including its attractor structure and stability to changes in initial conditions. This model is used to generate experimentally testable hypothesis about λ-phage behaviour if certain mutations are introduced. The model uses a simple mathematical formalism called hybrid systems, which includes both continuous variables to model biochemical substance concentrations, and discrete variables to model transcription factor binding site occupation states. A heater with a thermostat is a simple example of a hybrid system. It has two discrete states – on and off, while the temperature is a continuous value, and the state of the system depends on the temperature and the previous state. The presented model is a generalisation of the Finite State Linear Model (FSLM) first introduced in Brazma & Schlitt, 2003 and Schlitt & Brazma, 2006. Each transcription factor binding site has two affinity thresholds (constants) defining the transcription factor concentration at which it binds or detaches from the particular binding site. We were able to show mathematically that when described as a hybrid system, λ-phage has only two attractors corresponding to biologically meaningful states – lysis and lysogeny. More importantly, we show that this attractor structure does not depend on the exact values of the binding site affinity thresholds, nor on the exact shapes of the substance concentration growth or degradation functions, as long as some simple properties of the model are preserved. This allows us to potentially study all possible behaviours of the λ-phage network wiring diagram with different affinity threshold configurations. We found that although many of the possible parameter configurations led to the same attractor structure, it is possible to change these properties. This leads to an experimentally testable hypothesis, where if the relative affinity of two groups of binding sites in the wild type λ-phage are interchanged, then the behaviour of the system is predicted to change in a particular, experimentally testable way.

Stability is defined as the invariance of the system's behaviour to small changes in the initial values of the substance concentrations. In general, hybrid systems are not stable, however we show that the system defined by the λ-phage network is stable. By factoring the continuum of all possible states of the system (combining the continuous values of substances and binding site state configurations) we generated approximately 1,000 different symbolic states. Using an exhaustive computer enumeration, we demonstrated that the state transition diagram has a particular topology, which we refer to as a modular pseudo-DAG. Each module in this diagram, excluding the attractor modules, is characterised by the property that the system can stay in its particular state for only a limited time. The attractor modules are either single nodes or simple loops. Together this leads to the stability of the system. We hypothesise that such a structure is a general property for all or most biomolecular regulation networks.

Evidence of RNA degradation active regulation in *S. pombe*

Katherine Lawler, Samuel Marguerat (Sanger Institute), Jurg Bahler (Sanger Institute), Alvis Brazma

Gene expression is shaped by a balance between transcription rate and degradation of mRNA. In collaborative work with the Fission Yeast Genomics group at the Wellcome Trust Sanger Institute, we have investigated the kinetics of stress response in fission yeast. We are using a time course of two simultaneous microarray measurements – expression microarrays and RNA Pol II ChIP-chip arrays – to identify possible modes of regulation during the stress response. In particular, what contribution does the control of mRNA degradation make to shaping the stress response? We are investigating different modes of response and have identified genes whose mRNA appears to be tightly regulated by a varying transcription rate; mRNA which may be destabilised during the stress response, enabling rapid removal from the cell; and mRNA whose accumulation can be explained by a rapid change in transcription rate or stability soon after exposure to stress. Our results provide evidence that for a substantial number of genes, the RNA degradation rate in *S. Pombe* is regulated actively.

Visualisation of complex datasets

Nils Gehlenborg, Samuel Kaski (Helsinki University of Technology), Alvis Brazma

Visualisation of gene expression in a large number of conditions (over 1,000) is challenging and the traditional methods (heat-maps) do not work well since all conditions do not fit on a computer screen without wrapping. We propose that this can be overcome by using Hilbert Curves, which provides a good approximation to optimise the mean distances between pixels representing similar conditions. This method can be expanded in various ways, for instance gene expression profiles wrapped in Hilbert Curves can be used in combination with the traditional 2F PCA visualisation, to represent the relative distances between genes and their expression profiles in the same plot. The results have been presented in several data visualisation conferences and workshops.

We have established a collaboration with one of the leading labs in multidimentional data visualisation at the Helsinki University of Technology (Samuel Kaski) and we are working on applications of visualisation methods to gene expression (Venna & Kaski, 2007; see Figure 1).

Comparative analysis of human and mouse transcription factor expression

Richard Coulson, Alvis Brazma

Human and mouse protein-coding genes containing DNA binding domains (transcription factors) have been identified. This facilitates the creation of co-expressed clusters by seeding with these transcription factors. A gene becomes a member of the cluster if it displays differential expression under the same experimental conditions as the transcription factor, and if its expression pattern is significantly and highly correlated to that of the transcription factor. Over-represented GO terms and KEGG pathways in the clusters have been determined and compared with the corresponding annotations in the set of genes co-expessed with the orthologous transcription factor. These analyses suggest orthologous transcription factors control the expression of functionally distinct genes.

TRAINING

Gabriella Rustici, Misha Kapushesky, Helen Parkinson, Eleanor Williams

The Microarray Informatics team is actively involved in the EBI's user training programme. In particular we have been major contributors to the EBI's hands-on, in-house training programme and roadshow activities. ArrayExpress and Expression Profiler have been presented at 21 training events including seven Roadshows, five EBI training events and three major international conferences. The

transcriptomics resources are one of the most popular topics and always receive very high evaluation marks in the surveys following each training event.

Our Transcriptomics module (including ArrayExpress, Expression Profiler and Atlas) was one of the first courses in the EBI's new eLearning platform, launched in September 2008. Users can now interactively learn how to use the EBI's transcriptomics resources using a variety of learning materials including PDF documents, videos, quizzes and reflective tasks.

We have also published a chapter in *Current Protocols for Bioinformatics* (Rustici *et al.*, 2008).

FUTURE PROJECTS AND GOALS

There are two main challenges for ArrayExpress in the near future. The first is to develop the Gene Expression Atlas into a robust production database and increase its data volume. This will create a new database at the EBI allowing our users to easily query and display gene expression for any gene under any condition. The second is to build a standard pipeline for accepting the ultra high-throughput sequencing-based transcriptomics data in the ArrayExpress archive, and presenting these data via the Atlas. These data will be closely integrated with other relevant data at the EBI, in particular via the Bio Investigation Index. Moreover, we will continue building a distributed system for integrating functional genomics, genetics and medical bioinformatics data via the System for Information Management in Biomedical Studies (SIMBioMS). We will continue developing new data analysis algorithms and applying these to the integrated data. Integration and meta-analysis of different datasets, including human genome variation and medical data, will continue to be a major focus of our research, with the goal of understanding and modelling basic biological processes or diagnostics.

The Microarray Software Development Team

Ugis Sarkans

PhD in Computer Science, University of Latvia, 1998. Postdoctoral research in University of Wales, Aberystwyth, 2000. At EMBL-EBI since 2000.

Team Members

Software Development Project Leaders
Misha Kapushesky*
Nikolay Kolesnikov
Mohammadreza Shojatalab

Software Engineers
Niran Abeygunawardena
Hugo Berube*
Miroslaw Dylag
Ibrahim Emam
Sudeshna Guha Neogi
Ekaterina Pilicheva
Anjan Sharma
George Tsun-Po Yang*

Database Administrator
Roby Mani

**Indicates part of the year only*

INTRODUCTION

Our team has been developing software for ArrayExpress since 2001 (Sarkans *et al.*, 2005). As of October 2008 ArrayExpress holds data from almost 200,000 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;
- MIAMExpress – a data annotation and submission system;
- Expression Profiler – a web-based data analysis toolset;
- components used internally by the ArrayExpress production team.

In 2008 we continued our efforts to rebuild the ArrayExpress infrastructure, with an emphasis on simplification of data management tasks and software components, tighter integration of data submission and management, and use of automated workflows for data processing.

The team is also working on data management and integration solutions for domains beyond microarray data. We participate in the MolPAGE project, an integrated EU effort that aims to find biomarkers for genetic diseases, and type II diabetes in particular. Our team's main achievements in this project are: the basic principles and architecture of a framework for managing human sample information and output of high-throughput analytical platforms; and a generic data reannotation, integration and warehousing solution.

We also participated in data standardisation efforts, in particular MAGE (Spellman *et al.*, 2002) and MAGE-TAB (Rayner *et al.*, 2006) and regard this work as crucial for the successful development and support of a high-throughput data management infrastructure.

ARRAYEXPRESS – NEW INFRASTRUCTURE

Misha Kapushesky, Nikolay Kolesnikov, Mohammadreza Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw 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). ArrayExpress comprises two databases (the ArrayExpress Repository and the ArrayExpress Data Warehouse) and two related tools (the online data submission/annotation tool, MIAMExpress, and the data analysis tool, Expression Profiler).

One of the current central tasks of the Microarray Software Development team is simplification of ArrayExpress software. Due to the mode of evolution of the current infrastructure, data transfer between the various components of ArrayExpress has gradually demanded greater resources. We are addressing this problem by building pipelines that enable automation of data transfer. However, in the long run it is more efficient to restructure the system to minimise the need for unnecessary data and metadata transformations.

In the field of data exchange standards MAGE-TAB has become the *de facto* standard. It is a simple (compared to MAGE-ML) spreadsheet-based format which has been developed with a view that it should be easy to read and create in the absence of sophisticated tools. We are building the object model and database schema of 'new' ArrayExpress, and this work is heavily influenced by MAGE-TAB, as well as the original ArrayExpress model, MAGE version 1, MIAMExpress, ArrayExpress Warehouse, FuGE and MAGE version 2. This restructuring allows us to unify software development efforts that have previously been split across at least four different database schemas and software components (MIAMExpress, ArrayExpress Repository, ArrayExpress Data Warehouse and Expression Profiler).

We collaborate with a team (part of the Microarray Informatics team) led by Sussana-Assunta Sansone and involved in the development of standards and software for multi-omics data management. One

Publications

2007
Brazma, A. & Sarkans, U. (2007). Gene expression databases. *Encyclopedia of Life Sciences.* John Wiley & Sons Ltd, Chichester.

Jones, A.R., *et al*. (2007). The Functional Genomics Experiment model (FuGE): An extensible framework for standards in functional genomics. *Nat. Biotechnol*., 25, 1127-1133

2008
Haudry, Y., *et al*. (2008). 4DXpress: a database for cross-species expression pattern comparisons. *Nucleic Acids Res*., 36, D847-D853

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

Sarkans, U., *et al*. (2005). The ArrayExpress gene expression database: a software engineering and implementation perspective. *Bioinformatics*, 21, 1495-1501

Spellman, P.T., *et al*. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*., 3, research0046.1-research0046.9

of the components of Susanna's work is a system for integrating biological investigations, employing transcriptomics, proteomics and metabolomics technologies. A generic object model that is expressive enough to represent the 'sample' side of multi-omics experiments (i.e. metadata about studies, samples used, protocols, links to data files etc.) is being adopted for the needs of ArrayExpress, and supplemented by ArrayExpress-specific components.

ARRAYEXPRESS USER INTERFACES

Misha Kapushesky, Nikolay Kolesnikov, Mohammadreza Shojatalab, Niran Abeygunawardena, Hugo Berube, Ibrahim Emam, Anjan Sharma

ArrayExpress Repository

A good separation between the underlying infrastructure and the software driving the user interfaces of ArrayExpress has been achieved. This allows us to pursue evolution of user interfaces quite independently from the evolution of the database back-end.

The first generation of the ArrayExpress interface presented a fairly close mapping of the underlying object model (MAGE-OM) concepts to the web paradigm, providing complete, detailed access to every aspect of the stored data, from all protocol details to sample properties. This was rather overwhelming for the user. For example, it was not possible to obtain a simple overview of what was available in the database, nor did it permit programmatic access to the repository query mechanism by large-scale researchers.

The improvements to the next generation of the interface resolved these two insufficiencies of the older interface: a flexible data browsing graphical user interface was designed and implemented, and programmatic access for querying the database was simplified.

A commercial framework for building Rich Internet Applications (RIAs), called Backbase, was used for this implementation. This did not correspond well to our policy of building open source software components, but enabled us to prototype and build a successful interface in a relatively short timeframe. Also, at the time no suitable open source alternatives existed. Subsequent experience with this framework uncovered other drawbacks, in particular, suboptimal performance of interfaces resulting from the size and nature of the Backbase implementation.

The main achievement in 2008 for the repository interface is invisible for the end user, but it is very important in terms of future development. We have built another generation of the ArrayExpress repository interface, dependant only on open source software components. The framework that replaced Backbase is jQuery. From the end user point of view the differences are minor, but the underlying architecture has been significantly changed, resulting in a more robust software product.

The lightweight XML data representation that bridges data in the ArrayExpress relational database and the visual user interface components has been further improved. This enables additional functionality of repository querying, in particular by experiment type and by all sample and factor annotations.

The browser interface has been extended to provide a simple overview of several internal properties of experiments, including biological sample properties and experimental factors, which before were only accessible through the older detailed interface. Also, more flexible, detailed APIs for data querying and retrieval were developed.

ArrayExpress Data Warehouse

The granularity of information as managed in the ArrayExpress Repository is on the level of experiments. It is possible to find experiments satisfying certain criteria (keywords, experiment types etc.), retrieve data and associated annotation, but users cannot obtain information about the behaviour of individual genes. For this purpose, repository data needs to be analysed by some other means.

With the ArrayExpress Data Warehouse we have gone one step further. Array design annotation in the Data Warehouse is unified and therefore it is possible to retrieve and visualise data about the expression of specific genes in a particular subset of experiments. However, deeper biological conclusions about gene expression are left for the users to make; the Warehouse facilitates only data retrieval and visualisation.

When searching a gene expression database, biologists ideally wish to retrieve information in the form 'gene X is over/underexpressed in condition y'. To be able to distil and present relevant infor-

mation in this fashion, we have built a gene expression atlas (ArrayExpress Atlas; www.ebi.ac.uk/microarray/doc/atlas/), where data from the Data Warehouse is subjected to uniform statistical processing. The Atlas is fronted by a user-friendly interface, facilitating queries similar to the one above, as well as more complex searches. We think that for a user familiar with the ArrayExpress Warehouse functionality, who wants to find out how a defined gene or set of genes is behaving in a particular set of experimental conditions, it is more useful to provide the Atlas results as the first set of query results. This then allows the user to drill down into plots of gene expression, as provided by the current Warehouse interface. Therefore the current Warehouse and the Atlas prototype are scheduled to merge into a single user experience.

The Atlas project has been developed by Misha Kapushesky, and a separate project team has been spun out from the Microarray Software Development team, headed by Misha.

ARRAYEXPRESS MAINTENANCE

Misha Kapushesky, Nikolay Kolesnikov, Mohammadreza Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Ekaterina Pilicheva, Anjan Sharma, George Tsun-Po Yang, 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 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 workflow system is one aspect of our infrastructure that can be carried over to the new environment with minimal changes. This includes automated pipelines for array reannotation and data transfer from the repository into the Warehouse/Atlas.

A diverse set of tools has been developed in the ArrayExpress team for various data acquisition tasks. These tools are different in nature and use different technologies (MySQL, Oracle, Perl and Java etc.). Use of different languages and database platforms is not problematic *per se* since different tasks need different underlying mechanisms and lend themselves to different technologies. However, there is clearly a need to integrate different components in a workflow environment, with the potential to incrementally replace components with new versions as data communication standards evolve and new services are added.

We have adopted several open source libraries from the open source framework 'Open Symphony'. The main two are OSWorkflow and Quartz, which provide workflow functionality and job control facilities, respectively. We have implemented our own web front-end for these libraries in order to satisfy the requirements of the ArrayExpress production team.

On top of this generic workflow engine, several workflows have been built. The most mature of these is an array design reannotation system and this has been working in production mode for more than six months. This system solves the problem of Ensembl fluidity, and meets our requirement to update array design annotation to the most recent version of Ensembl. This process is now carried out for each Ensembl release and all array designs in the Warehouse have been reannotated. Two types of events happen during the reannotation process: 1) associating a particular set of probes on an array design to a different gene; 2) updating gene information, i.e. refreshing the set of cross-references to various databases and ontologies. In addition to consulting Ensembl, this process also involves UniProt lookup, and it is immensely useful to have a workflow system in place that allows various pieces of software to be connected.

Another workflow that will automate mundane tasks is currently being tested. The workflow will score experiments present in the ArrayExpress Repository (both from the point of completeness of sample annotation, as well as the ability to map array design to the genome) and automatically load high-scoring experiments into the Data Warehouse. Due to our two main databases having vastly different schemas and APIs, this is a nontrivial task. Although the upcoming restructuring of ArrayExpress will minimise the need for data transfer, there will still be a need to score experiments and choose the most useful expression level measurements for gene/sample pairs, and this workflow will be reused in that context.

MEDICAL INFORMATICS

Sudeshna Guha Neogi, Ugis Sarkans

In 2008 several projects adopted the system built for the EU project MolPAGE for acquisition and management of sample information, as well as the repository for high-throughput technology data acquisition and management, collectively known as SIMBioMS. While this work (coordinated by Maria Krestyaninova) was mostly done externally to the Microarray Software Development team, we contributed by working on summary data integration, cleaning and presentation through the Warehouse interface.

Software for the MolPAGE Data Warehouse (MoDa) is the same as used for the ArrayExpress Data Warehouse. MoDa has served as a proof of concept that we can use the same annotation pre-processing pipelines, data structures and interface paradigms to serve not only microarray gene expression data, but also data generated by methylation, metabolomics, protein array and tissue array experiments. This could potentially be expanded to any technologies that produce data as a matrix: 'gene or other biomolecular entity versus various biological samples'.

FUTURE PROJECTS AND GOALS

The main goal for 2009 is gradually rolling out the next-generation infrastructure for ArrayExpress. The main tasks of this project include:

- replacing the central databases of ArrayExpress with a single unified database;
- organising a system for synchronising the old and new databases, enabling testing and gradual switchover to the new infrastructure;
- porting the existing workflows to the new system;
- porting ArrayExpress interfaces to work with the new back-end.

There are significant changes planned also from the perspective of external users. There is a clear conceptual separation between ArrayExpress Repository and ArrayExpress Warehouse/Atlas. The former serves the need to provide an archive of scientific record, working in tandem with journal publications, while the latter provides researchers with biologically meaningful information about expression of genes in various conditions – either in a more 'raw' form (as in the Warehouse), or already processed (as in the Atlas). However, there is no need to maintain a distinction between these resources from the web access point of view. We plan to build a unified interface that, in response to user queries, will provide information on all levels of granularity: whole experiments (Repository), sets of expression profiles (Warehouse), and interesting facts about gene expression (Atlas), providing users with a choice for further investigations.

Macromolecular Structure Database Team

INTRODUCTION

The Macromolecular Structure Database (MSD; www.ebi.ac.uk/msd) is one of the five core molecular databases (genomes, nucleotides, proteins, 3D structures and expression data) hosted by the EBI. The MSD holds detailed knowledge of protein structure and function of biological macromolecules. Access to this information is vital for many different users, for example, the identification of potential targets for therapeutic intervention as well as of lead structures for pharmaceutical use. MSD usage averages more than two million requests a month. Through our membership of the Worldwide Protein Data Bank organisation (wwPDB; www.wwpdb.org) we are an equal PDB partner and work closely with the United States (RCSB) and Japanese (PDBj) partners to provide the single international archive for structural data. We uniquely 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. Our aim is to continue to provide integrated data resources that evolve with the needs of structural biologists; 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 PDB archive is growing annually in value, importance and size. The current conservative estimate of the value of the entire PDB archive is in excess of $4.5 billion (just over €3 billion), with the annual research cost involved in elucidating these structures estimated to be $750 million (€530 million). More than 15 different funding agencies worldwide fund the wwPDB centres, generating an estimated total of $9 million (€6.4 million) per year. When these figures are compared, the costs of maintaining the archive represents approximately 1% of the annual research investment. This year saw the PDB make the deposition of scientific evidence for a macromolecular structure mandatory. From 1 February 2008, structure factor amplitudes/intensities (for crystal structures) and restraints (for NMR structures) are now mandatory for PDB deposition. This policy was published at http://www.wwpdb.org/news.html. In accordance with this policy, the PDB ID should be included in publications and authors must agree to release the atomic coordinates and associated experimental data when the article is published.

WORLDWIDE PROTEIN DATA BANK ACTIVITIES

To continue to meet the goals of the PDB as a critical global scientific resource we must evolve with and anticipate the needs of the scientific community. This includes the five- to ten-year vision to enrich the annotation of entries to include biological function as well as all aspects of the PDB services. The wwPDB members have started work for a common deposition and annotation project. Throughout 2008 a series of meetings were held to engage the full staff in assessing current bottlenecks and future needs of PDB users and to propose strategic projects that will insure our ability to meet those needs. The goal of this project is to maximise the efficiency and effectiveness of macromolecular structure data handling, and support the scientific community through the development and adoption of common deposition and annotation processes and tools. These shared processes and tools will enable balancing of deposition load across the wwPDB member sites and allow for shared maintenance and updates to the archive and tools in the future. The next deposition and annotation product will be a unified system, which will be used at all deposition sites. It is designed to provide significant reduction in curation time, assure consistency and will be extensible to accommodate hybrids and new structural determination methods. It will also provide users with portable validation tools and 'table 1' of their experiment. Development is under the control of a business plan.

As a joint project undertaken by the wwPDB partners, we now provide entry-level download statistics collected from all three distribution sites. The service is run under http://www.wwpdb.org/downloadStats.php. As of August 2007, statistics have been available for FTP and http (web) downloads and views for each PDB structure, from all wwPDB sites on a monthly basis.

At the EBI, MSD staff have continued to process deposited entries via the PDB deposition tool, http://www.ebi.ac.uk/msd-srv/autodep4/. Deposition is the process of submitting experimental data, currently from crystallography (X-ray), Nuclear Magnetic Resonance (NMR) or electron microscopy (EM), by scientists to the wwPDB central site that stores these data.

Kim Henrick

Postdoctoral research at the Polytechnic of North London, Imperial College of Science and Technology, SERC Daresbury Laboratory, Cambridge Centre for Protein Engineering and National Institute for Medical Research. At EMBL-EBI since 1996.

Team Members

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

Database Development
Harry Boutselakis
Norman Cobley
Dimitris Dimitropoulos
Robert Hulme
Anne Pajon
Jorge Pineda
Antonio Suarez Uruena
Miri Hirshberg

Data Search and Retrieval
Adel Golovin
Matt Harrison
Eugene Krissinel
Thomas Oldfield

NMR Validation Tools
Chris Penkett
Wim Vranken

Database Administrator
Melford John

Database Project Leader
Sameer Velankar

Publications

2007
Vranken, W. (2007). A global analysis of NMR distance constraints from the PDB. *J. Biomol. NMR,* 39, 303-314

2008
Dutta, S., *et al*. (2008) Data deposition and annotation at the world-wide Protein Data Bank in *Structural Proteomics High-Throughput Methods Series: Methods in Molecular Biology*, 426, 81-101 Kobe B, Guss M, Huber T (eds.)

Golovin, A. & Henrick, K. (2008). MSDmotif: Exploring protein sites and motifs. BMC *Bioinformatics*, 9, Article 312

Haquin, S., *et al*. (2008) Data management in structural genomics: an overview. *Methods Mol Biol.,* 426, 49-79

Henrick, K., *et al*. (2008) Remediation of the protein data bank archive. *Nucleic Acids Res*., 36 (Database issue):D4-26-D433.

Lawson, C.L., *et al*. (2008). Representation of viruses in the remediated PDB archive. *Acta Crystallogr. Sect. D-Biol. Crystallogr.*, 64, 874-882

Markley, J.L., *et al*. (2008). BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR,* 40:153-155

MSD DATABASE ACTIVITIES

The annual MSD Scientific Advisory Board (SAB) was held in February 2008. A recommendation of the SAB was that the group changed its name to become the Protein Data Bank in Europe (PDBe). We are in the process of migrating to this new name and logo (see left).

In the past year, the MSD has consolidated its search and retrieval services. Between January 2008 and February 2008 all the services were ported to new 64-bit machines with a new disk structure organisation and a new monitoring system. Some services were retired including DALI, MSDtarget, Biotech validation and Relibase while the services PQS, CAPRI, OCA and OLDERADO were changed from virtual services to URLs under www.ebi.ac.uk/msd-srv/.

The MSD has continued to expand its search and retrieval services to permit database searches using both generic and expert specific viewpoint systems. The team has redesigned the core MSD relational database as the source of the data has changed. With the completion of the first round of the remediation project, the wwPDB mmCIF files are now the primary source of flat files to be loaded into the database. The system used to follow the wwPDB data model and nomenclature is now simplified, easy to maintain and stable without external dependencies. If problems are encountered in the data during loading, the data held in the primary public data source is corrected. The new database has the following benefits:

- it is open source capable;
- the database is vendor independent;
- de-coupling of the logic from the database allows modularity and portability due to an in-house Java loader framework.

We have made considerable progress for the smooth continuity of the project in the future with new staff taking responsibility of key developmental areas. Good progress on improving performance has been achieved, with the ability to load the complete PDB in less than 14 hours (it was previously 14 days).

A new service, MSDmapping (http://www.ebi.ac.uk/msd-as/MSDMapping/) provides the residue and chain level mapping between the PDB and external databases (both ways): UniProt, InterPro, CATH, SCOP, EC, Pfam, NCBI, PubMed and GO. This formalises access to these data which was initiated through our SIFTS project (http://www.ebi.ac.uk/msd-srv/docs/sifts/).

MSD has also been expanded by the adoption of the OLDERADO service from the University of Leeds (http://www.ebi.ac.uk/msd-srv/olderado/). This service provides information on the core atoms, domains and representative structures from current NMR-derived ensembles of protein structures deposited in the PDB. The combined MSD services (providing access to 20 different views of PDB data) deliver 3.6 million hits per month, transferring about 238GB to approximately 53,000 users. In addition, users download around 1.7 million and 4 million files per month from the MSD and PDB FTP sites respectively.

Grid and e-Science Research and Development

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.

Team Members

Scientists
Alan Bleasby
Tony Burdett*
Syed Haider
Jon Ison
Shaun McGlinchey
Tom Oinn*
Mahmut Uludag

**Indicates part of the year only*

INTRODUCTION

The team's focus is on the integration of bioinformatics tools and data resources. We 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 eScience project) and for various projects (including EMBRACE and ComparaGrid) that integrate access to bioinformatics tools and data content.

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 steerage or high-throughput sequence analysis. The most visible Grid project is currently CERN's Large Hadron Collider 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 existing 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 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.

THE MYGRID PROJECT

Our group has been the biological specialist participant in the UK-funded myGrid project and we are continuing 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. EMBL-EBI's role is through the Taverna workbench and as an application and data service developer and provider which continues through the EMBRACE and EMBOSS projects.

Taverna

MyGrid's flagship application is the Taverna E-Scientist's Workbench, initially developed and now coordinated by Tom Oinn. Taverna currently has over 9,000 downloads, 1,500 installations, and users in Europe and worldwide, including industry and in areas beyond bioinformatics (e.g. 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, SeqHound from the BluePrint/BIND group in Toronto and EMBL-EBI's BioMart databases.

Publications

2008

Belhajjame, K., *et al*. (2008). Metadata management in the taverna workflow system. In *Proceedings CCGRID 2008 - 8th IEEE International Symposium on Cluster Computing and the Grid*, 651-656

Lanzen, A. & Oinn, T. (2008). The Taverna Interaction Service: Enabling manual interaction in workflows. *Bioinformatics*, 24, 1118-1120

Li, P., *et al*. (2008). Performing statistical analyses on quantitative data in Taverna workflows: An example using R and maxdBrowse to identify differentially-expressed genes from microarray data. *BMC Bioinformatics*, 9, Article 334

Li, P., *et al*. (2008). Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics*, 24, 287-289

Internally, Taverna uses Life Science Identifiers (LSIDs) throughout to uniquely identify and retrieve objects in the 'myGrid information model'. Our experience in creating workflows using existing standards such as WSFL (Web Services Flow Language) has been frustrated by the lack of appropriate editing tools and by the level at which business-targeted workflow definitions are stored. Taverna therefore uses our own workflow language, SCUFL (Simple Conceptual Unified Flow Language), which has been very positively received by other members of the myGrid consortium and by the wider research community.

Taverna is made available under the Lesser General Public License through SourceForge at http://taverna.sourceforge.net/.

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.

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.

We have already started an ambitious programme of new developments in database indexing, graphics outputs, pattern searches, automated documentation and help text. Support for the GFF3 (Generic Feature Format version 3) annotation standard was extended to cover protein sequence features in the new EMBOSS 6.0.0 release in July 2008. GFF3 makes extensive use of the Sequence Ontology (SOFA) which has been extended to cover protein features. These extensions are now part of the development version of EMBOSS and will be included in the next release. The release also completes the standardisation of library function naming and interfaces for developers, significant performance improvements, improved documentation in output files, and an extended test suite. The new syntax for the EMBL-Bank and UniProt databases is now the default.

We are keen to encourage contributions from outside developers, and have completed a major refactoring and redocumenting of the source code and the programming interface together with new documentation for system administrators. This will be published as a series of open source books by 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.

We have a new collaboration with SciTegic Inc. to develop Pipeline Pilot components for EMBOSS. This has required us to build and support a native Windows implementation (mEMBOSS) which was released as a beta version with release 4.0.0 in July 2006. The current EMBOSS release (6.0.0, July 2008) has a fully-supported Windows implementation with the Jemboss 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.

EMBOSS is available from http://emboss.sf.net/.

EMBRACE

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

SoapLab (http://soaplab.sf.net/) does not access individual analysis programs 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 (Object Management Group, Life Sciences Analysis Engine) 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.

Martin Senger, the developer of SoapLab, has been visiting the group on behalf of the Generation Challenge Programme (a unique worldwide project focused on crop bioinformatics, trying to improve life and resources of the resource-poor regions all over the world). The main objective of our collaboration is to develop 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 External Services. 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 (6.0.1), 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 developments 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 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

BioMart is a project within the Ensembl group at the EBI and 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 the EBI and by other partners in EMBRACE. Following the recent development and restructure 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 that will bring forward cases from wet-lab biologists, and subsequently add to the usefulness of the mart data. We have conducted several training workshops on BioMart web services during the last year to reach a wide bioinformatics community.

COMPARAGRID

In a project led by the Roslin Institute, we took a more specific approach to a problem similar to EMBRACE. The aim of ComparaGRID is to enable a biologist with, for example, a quantitative trait locus in sheep to track from genetic information in the species of interest, through links to various kinds of physical and genetic map information, to sequence data and experimental validation in one or more of the sequenced model organisms.

We worked with computer scientists in Newcastle and with ontology specialists in Manchester and at Roslin to define the relationships between the various data resources at EMBL-EBI and among the more specialised plant, animal and microbial research communities. The resulting ontological structure was then supported by a set of services that allow biostatistical research groups to analyse sets of data and for ontology researchers to 'reason' over the structured results. The ontology was successfully established, with workshops involving biologists and computer scientists to further develop the current working model.

Data is manipulated using the Web Ontology Language (OWL). As a data publisher in the project we provide access to our core data resources in a compatible form (database OWL) to which we can apply a set of mapping rules to translate the results into the ComparaGrid OWL representation. We are providing services in support of this prototype for validation of the model and further exploratory research. We also provide a Protege plugin to map the ComparaGrid ontology to a relational database schema, define Java classes, and aid in the definition of mapping rules for other data publishers in the project.

FUTURE PROJECTS AND GOALS

The services provided by the group remain largely SOAP-based web services. These have proved 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 will move beyond sequence data and analysis services to cover the remaining areas of the EBI's core databases and to integrate services from our partners using the same standards and interfaces.

Literature Resource Development

Peter Stoehr

MSc Applied Genetics 1978, Birmingham University.
Statistician in Agriculture Faculty, Newcastle University.
Analyst/Programmer with AFRC, Cambridge and Harpenden.
At EMBL since 1988, EBI since 1994

Team Members

Senior Software Engineers
Sharmila Pillai
Mark Rijnbeek*

Software Engineer
Nikolay Nikolov*

Visitor
Richard Easty

**Indicates part of the year only*

INTRODUCTION

The major goal of literature resource development is to integrate the scientific literature with the data in biological databases and provide public services to exploit this. We achieve this by implementing a state-of-the-art search engine, flexible web access, novel biomedical text mining methods and ontologies such as Gene Ontology (GO) and Unified Medical Language System (UMLS).

We develop literature resources for use in-house and in public services. A local copy of PubMed is maintained under lease from the US National Library of Medicine (NLM), supplemented by bibliographic data from other sources such as AGRICOLA (USDA-NAL) and Chinese Biology Abstracts (CAS-SICLS). Biological patent abstracts are captured from the European Patent Office (EPO) and the US Patent Office (USPTO).

CiteXplore has been developed as a tool for querying the scientific literature, showcasing text mining methods, and linking to biological databases. UMLS, GO, the NCBI taxonomy and gene synonyms from UniProt are used as thesauri. Text mining methods from the research community, such as the 'Whatizit' methods of the Rebholz-Schuhmann research group at EMBL-EBI (http://www.ebi.ac.uk/Rebholz-srv/whatizit/form.jsp) provide several filters for enrichment of text with annotation. Gene and protein names from UniProt and GO terms are examples of entities identified in text and hyperlinked to the underlying data resources.

ACTIVITIES DURING 2008

CiteXplore development

Mark Rijnbeek, Sharmila Pillai, Peter Stoehr

The service platform for CiteXplore was enhanced during 2008 by migrating the back-end Oracle database from Oracle version 9i to Oracle 11g, from a Tru64 Unix to RedHat Linux, and from a pair of Alphaservers to Intel-based Sun servers. The database servers remain in a redundant configuration, using Oracle Real Application Cluster (RAC) technology to protect against server failure, and a physical standby database protecting against storage failure.

CiteXplore features developed during 2008 include:

- enhancement of the content of article citations by using the full XML archive of PubMedCentral as well as via CrossRef Web Services. We currently have reference lists from nearly two million articles, about 60 million references in total, and match them to existing CiteXplore records in nearly 70% of cases. This percentage is steadily being improved as we analyse the non-matching references. In CiteXplore we not only show these references, but also show 'cited-by' articles, and where possible also show the text fragment from the full-text article where the citation is being made;
- indexing several additional fields for querying, as described in Table 1 (below).

Field	Description
has_fulltext	records which have link(s) to full-text articles (Y/N value)
has_free_fulltext	records which have link(s) to free full-text articles (Y/N value)
has_pdf	records which have link(s) to full-text articles in PDF (Y/N value)
has_html	records which have link(s) to full-text articles in HTML (Y/N value)
fulltext_site	records which have links to full-text articles at a specific site (e.g ScienceDirect, PubMedCentral, WileyInterScience, HighWire etc)
has_doi	records which have a link to full-text articles as a Digital Object Identifier (DOI) (Y/N value)
in_pmc	records which have link(s) to full-text articles in PubMedCentral (Y/N value)
in_ukpmc	records which have link(s) to full-text articles in UK PubMedCentral (Y/N value)
grant_id	articles supported by a specific grant number (e.g. '103HR960426P000-000')
grant_agency	articles supported by a specific grant agency (e.g. 'United States NHLBI')

Publications

2008
Nikolov, N. & Stoehr, P. (2008). Integrating biomedical publications with existing metadata. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 653-655

Web services

Sharmila Pillai, Mark Rijnbeek

Many bioinformatics resources are now available through programmatic interfaces, known collectively as web services, for example REST and SOAP which are accepted standards in the web community. Such web services are increasingly used in combinations or workflows. We have implemented our web service access to CiteXplore using JAX-WS technology (http://www.ebi.ac.uk/citations/webservices) to enable comprehensive querying of the database and downloading of data. In 2008, we have begun to implement REST web services in order to attain high performance for third party query systems such as UK PubMedCentral (UKPMC), on the assumption that the SOAP protocol imposed unnecessary overhead on response times.

Bibliographic data sources

Peter Stoehr, Mark Rijnbeek, Weimin Zhu

By October 2008, CiteXplore served over 18.4 million PubMed records from about 5,000 journals, over 475,000 AGRICOLA records, and over 1.6 million patents. Over 1,200 records have been added by the Literature Resource Development team itself.

We have continued to collaborate with the Shanghai Information Centre of Life Sciences (SICLS), a part of the Chinese Academy of Sciences, enabling us to incorporate over 134,000 records from the Chinese Biology Abstracts (CBA) into CiteXplore. These data include English abstracts and metadata that are not already covered by PubMed.

UK PubMedCentral (UKPMC)

Peter Stoehr, Dietrich Rebholz-Schuhmann

In 2006, a contract was awarded to a consortium comprising the British Library, University of Manchester and the EBI to operate a UK version of PubMedCentral (PMC). PMC is a free archive of biomedical literature hosted by NCBI/NLM. The five-year project is funded by a group of nine UK funders led by the Wellcome Trust. In January 2007, the UK PubMedCentral website was launched at www.ukpmc.ac.uk, establishing a mirror of the PubMedCentral site and a service to allow UK authors to submit full-text articles to the archive. The Wellcome Trust has published a policy requiring that all Trust-funded researchers make their peer-reviewed research available through PMC.

In July 2008, a three-year grant was awarded to the EBI by the Wellcome Trust (on behalf of the UKPMC Funders Group) to begin development of UKPMC and contribute biomedical domain knowledge to integrate the research literature with the underlying bioinformatics databases. The value of open access scientific literature is not only that it is free to read, but that it is also free to re-use and exploit electronically. Our work on the development of UKMPC has initially focused on two projects:

- **access to comprehensive bibliographic data:** the UKPMC archive itself contains around 1.5 million full-text articles, but we believe a realistic search of the literature must be comprehensive and include PubMed. We are customising the CiteXplore web service interface to provide this functionality;
- **semantic enrichment of UKPMC documents:** the research group of Dietrich Rebholz-Schuhmann at EBI, in collaboration with the UK National Text Mining Centre (NaCTeM) in Manchester will apply text mining techniques to the UKPMC corpus. Biological entities such as proteins, genes, GO terms, organisms, chemicals, drugs diseases, will be systematically identified and extracted from the text and presented as annotations and summaries of the documents.

Citation networks

Nikolay Nikolov, Peter Stoehr, Sharmila Pillai, Mark Rijnbeek

The feasibility of exploiting references cited in scientific articles is being explored. This information would naturally fit the browsing function of CiteXplore to enable users to conveniently navigate cited articles, but would also open possibilities for generating citation networks and enabling bibliometric analysis. Such references are not provided as part of the PubMed/AGRICOLA datasets and we have identified three main complementary approaches to gathering such data:

- **PubMedCentral:** PMC XML records contain structured citations and we have processed the open access section of PMC and incorporated the article citations into CiteXplore. In 2008 we gained

access to the complete UK PubMedCentral archive, and have processed all the available XML documents to extract citations and their contexts. Unfortunately, most of the UKPMC archive is in scanned image form and the number of articles with citations in XML form is only around 240,000;

- **CrossRef Web Services:** we have subscribed to the CrossRef Web Services (CWS) initiative (alongside Microsoft and Scirus) to gain access to centrally-deposited article metadata from publishers. In many cases, these metadata include citations that we can extract and re-use. During 2008, we systematically downloaded the CrossRef data and found the download mechanism unreliable and the quality of the bibliographic data variable. Nevertheless, by October 2008 we had very nearly processed the complete back-file of publications up to early 2008, and applied the extracted citations to CiteXplore;
- **web harvesting:** inspired initially by CiteSeer, the computer science bibliographic resource built automatically from the web, we have run a pilot project to identify biomedical articles freely available on the web as PDFs. Using the Yahoo API, we have systematically queried using PubMed article titles, and have accumulated over four million document URLs. Although a high percentage of these will not represent a copy of the actual article, we have begun building the pipeline tools necessary to download and extract citations from these documents. Approximately 12 million queries have been issued over the year, generating about one million 'hot' URLs (URLs which are hit by more than 3 queries). Approximately 33% of these URLs seem to be real PubMed articles, based on manual sampling. About 180,000 of these PDF documents have been downloaded and converted to HTML. Matching of these to actual PubMed articles is a little problematic, with currently less than 50% success rate. Similarly the success rate of extracting citations from these documents (HTML) and matching them precisely to existing PubMed records is quite low and work is ongoing.

Table 2 shows the current state of citation harvesting from UKPMC and CrossRef data. Figure 1 shows the display of cited/cited by information in a CiteXplore record.

Origin	Articles with citations	Total citations	Citations matched	Distinct citations matched	% citations matched
PMC XML	247541	9617135	8143527	2199722	84.68
CrossRef	1280117	36372879	23243574	5825315	63.90
Overall	1527658	45990014	31387101	6347611	68.25

Table 2. Summary of citations in CiteXplore, from UKPMC XML data and from partial CrossRef data (as of 10 October 2008).

Figure 1. Screenshot of a PubMed record in CiteXplore, showing mark-up of proteins and organisms in the text, links to the complete article (a free PDF at UKPMC), article citations and cross-references to EMBL nucleotide sequence databases.

Table 2 shows that of the 1.5 million articles which include reference lists in CiteXplore, there is a total of nearly 46 million references, 68% of which can be matched to an existing publication. 6.3 million articles in CiteXplore are found to be cited by at least one other, and within this group, the average article is cited five times. It is commonly stated that 80% of all citations reference just 20% of articles. This table also shows that PMC XML data provides a higher quality of data, resulting in a higher percentage of matched citations.

Firefox web browser plugin

Richard Easty, Nikolay Nikolov

We have explored the idea of developing a plugin for the Firefox web browser, which could help us gather citation-related data via end users. A plugin has been developed which will process the results on a Google Scholar search page, and provides the user with links to the corresponding CiteXplore record(s), as well as links to databases such as UniProt, retrieved via CiteXplore web services. In return for these user-related enhancements, the plugin returns some data to us, such as citation counts from Google Scholar. Similarly, a recent extension to this plugin will work on PubMed search results to add links to Google Scholar and CiteXplore.

Linking to full text

Peter Stoehr

We continue to maintain a database of links to full-text articles, synchronised with updates to PubMed and AGRICOLA. The links are either direct URLs, which (usually) follow publisher-specific rules, or Digital Object Identifiers (DOIs; www.doi.org). We are a member of CrossRef (www.crossref.org) as an affiliate member for the purpose of harvesting DOIs.

These links are exploited in CiteXplore and underlying data are also made available for other EMBL-EBI projects, as a standalone file or within an internal Oracle database. DOIs derived from our full-text link resource are incorporated into the publications cited in the EMBL Nucleotide Sequence Database, and the UniProt KnowledgeBase.

CiteXplore now includes over 17 million links to external sources, with over 6.4 million DOIs and links to 2.6 million PDFs.

Food composition – EuroFIR project

Peter Stoehr, Sharmila Pillai, Mark Rijnbeek

The European Food Information Resource (EuroFIR) project is a five-year EU FP6 Network of Excellence. It aims to build and disseminate a comprehensive, validated databank providing a single, authoritative source of food composition data for nutrients and bioactive compounds. We are involved in establishing IT standards and the internet deployment of EuroFIR databanks, including a common bibliographic database based on CiteXplore.

In 2008, we captured the results of a pilot literature indexing project conducted by several of the EuroFIR food composition database compilers. Scientific articles were indexed according to several characteristics such as whether the articles described an analytical method or country of origin etc. We implemented this in CiteXplore as a demonstration of how EuroFIR-specific annotations could be searched. Later, we implemented a specific field (subset) which can be used to restrict searches to just those articles in CiteXplore which are relevant to EuroFIR, by including the search term 'SB:eurofir'.

FUTURE PROJECTS AND GOALS

Text mining functionality and enhanced CiteXplore

We plan to accelerate exploitation of recent methodology into CiteXplore in areas such as GO and MeSH annotations, related articles, inclusion of semantic information in the indexing, and sentence/paragraph retrieval in addition to whole documents. We will explore the use of BioLexicon, a terminological resource generated from EBI databases, to facilitate interoperability of literature with the databases.

Additional bibliographic data

In collaboration with the British Library, we will identify additional content for UKPMC, for example UK National Health Service publications, NICE guidelines etc., and the bibliographic data for these will be exposed via CiteXplore.

Citation networks

We will begin to process UKPMC articles that are available only as scanned images, using the results of OCR and citation extraction from NCBI. An extensive task is to complete the pipeline of harvesting relevant scientific articles from the web, and accurately extract citations and their context from these.

As the citation network fills, we will include citation counts in the CiteXplore indexing and ranking function, in the same fashion as the Google PageRank method, to enable highly-cited articles to appear more prominently in search results.

Web services

Further web services interfaces to all of the CiteXplore functionality will be developed to make the bibliographic data aggregated at the EBI available to third party information systems, workflows and research tools. REST web services will be published towards the end of 2008, providing high performance for specific remote services such as UKPMC.

Section 3

Research in 2008

The Bertone Group: differentiation and development

Paul Bertone

PhD 2005, Yale University.
At EMBL–EBI since 2005.
Joint appointments in Gene Expression and Developmental Biology Units.

Group Members

Staff Scientist
Mali Salmon-Divon

Postdoctoral Fellows
Pär Engström
Remco Loos

PhD Students
Heidi Dvinge
Diva Tommei

Visitors
Miia Rõõm
Beatriz Rosón
Kairi Tammoja

**Indicates part of the year only*

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 preimplantation 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 also 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 increases the stability of RNA:DNA duplexes (Hakansson *et al.*, 2001, Wang *et*

Publications

2008

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

Kirstetter, P., *et al*. (2008). Modeling of C/EBP? Mutant Acute Myeloid Leukemia Reveals a Common Expression Signature of Committed Myeloid Leukemia-Initiating Cells. *Cancer Cell*, 13, 299-310

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

Battersby, A., *et al*. (2007). Comparative proteomic analysis reveals differential expression of Hsp25 following the directed differentiation of mouse embryonic stem cells. *Biochim. Biophys. Acta*, 1773, 147-156

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

Dunkley, T.P., *et al*. (2004). Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics*, 3, 1128-1134

Frith, M.C., *et al*. (2006). Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol*., 3, 40-48

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).

Proteomic analysis constitutes another dimension to global functional studies, and we are working on analytical methods to further the development and application of these approaches. In the LOPIT (localisation of proteins via isotope tagging) method, changes in protein expression are observed through the use of cleavable ICAT labelling, enabling relative quantitation of protein levels by mass spectrometry (Dunkley *et al*., 2004). This is performed following a series of gradient fractions to separate organelles, allowing their protein complements to be tagged for subsequent identification. The spatial organisation of eukaryotic proteins is related to their function, so the assignment of an uncharacterised protein to an organelle provides insight into its possible role.

The localisation of novel proteins is then determined using multivariate data analysis techniques to match their distributions with those of proteins that are known to reside in specific subcellular compartments. This analysis can be used to measure relative protein abundance between populations, as well as variation within a population. In particular, we aim to follow specific signalling events in a quantitative manner, applying this technique to analyse the abundance and compartmental localisa-

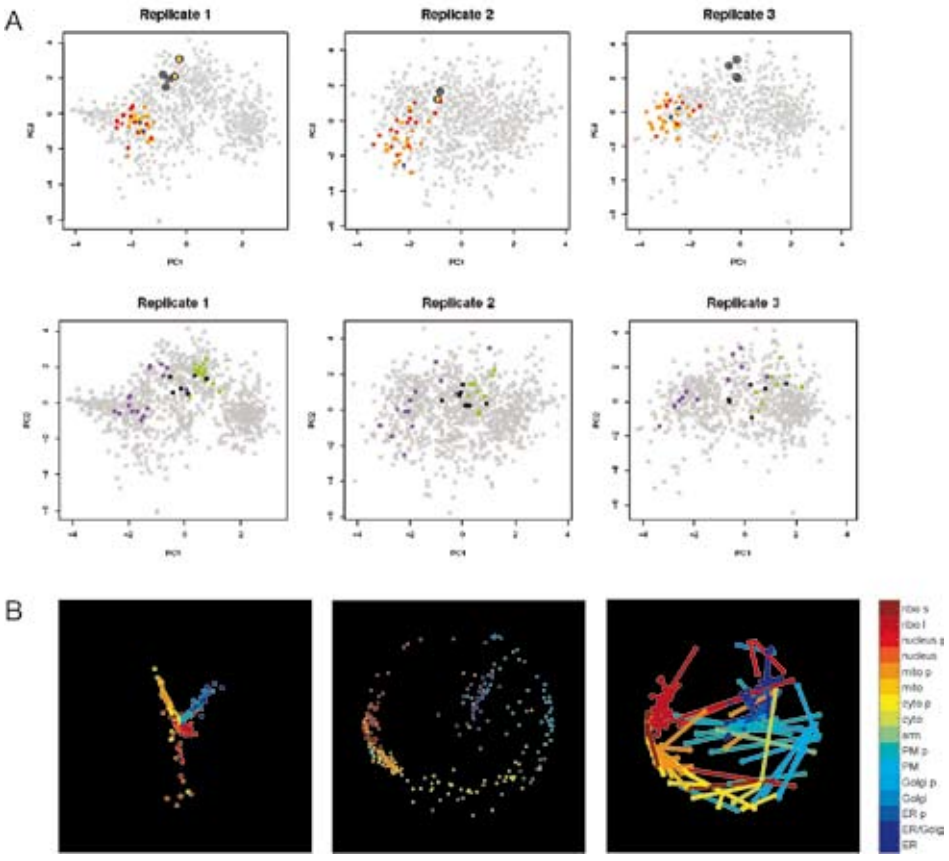


Figure 1. *Quantitation of organelle proteins by isotope tagging. Proteins are identified by mass spectrometry following analytical centrifugation to isolate subcellular components, allowing rapid determination of proteomic expression and localisation. A) LOPIT analysis yields a snapshot of the cell in action. Top: 40S ribosomal subunits (yellow), initiation factors (red) and elongation factors 1 alpha and 2 (blue) colocalise at the site of translation. In contrast, elongation factors 1 beta, gamma and delta (blue circles) colocalise with the proteins P40 and Rps21 (yellow circles) at a different site, possibly associated with the process of GTP exchange. Bottom: Endocytic proteins (black) are localised at the cytoplasmic (purple) edge of the plasma membrane cluster (green). B) Multiple cellular conditions are then compared to provide a dynamic view of the changing proteome. Here the data have been separated by principal components (a) and projected onto a sphere (b) for clearer visualisation of organelle subgroups. Superposition of data points taken across multiple conditions reveals proteins that are affected in response to pertubation experiments (c), implicating these molecules in various signalling pathways. (Data: Denise Tan, University of Cambridge; Analysis: Heidi Dvinge, EMBL-EBI).*

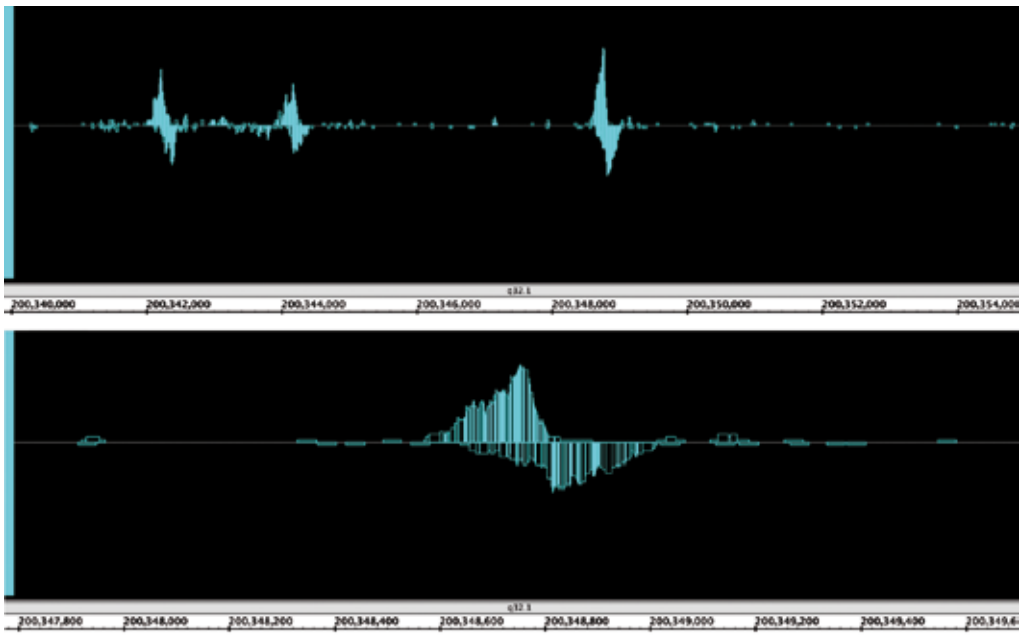


Figure 2. *Top: Transcription factor binding sites mapped across the genome using chromatin immunoprecipitation combined with high-throughput DNA sequencing (ChIP-seq). Bottom: Typical binding site peak after sense/antisense strand mapping of the trace reads to the reference genome and normalisation against a negative control dataset.*

tion of proteins as they change during various stages of cellular differentiation (Battersby *et al*., 2007). Monitoring trans-organelle dynamics of key components of particular signalling pathways allows us to identify proteins that undergo similar transitions together with these components (Figure 1).

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 (Figure 2). 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.

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 probeset. 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 3). This analysis is greatly enabled by direct programmatic access to Ensembl (Flicek *et al*., 2008), facilitating the accurate integration of current genome annotation data.

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 complementatry sequneces. *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

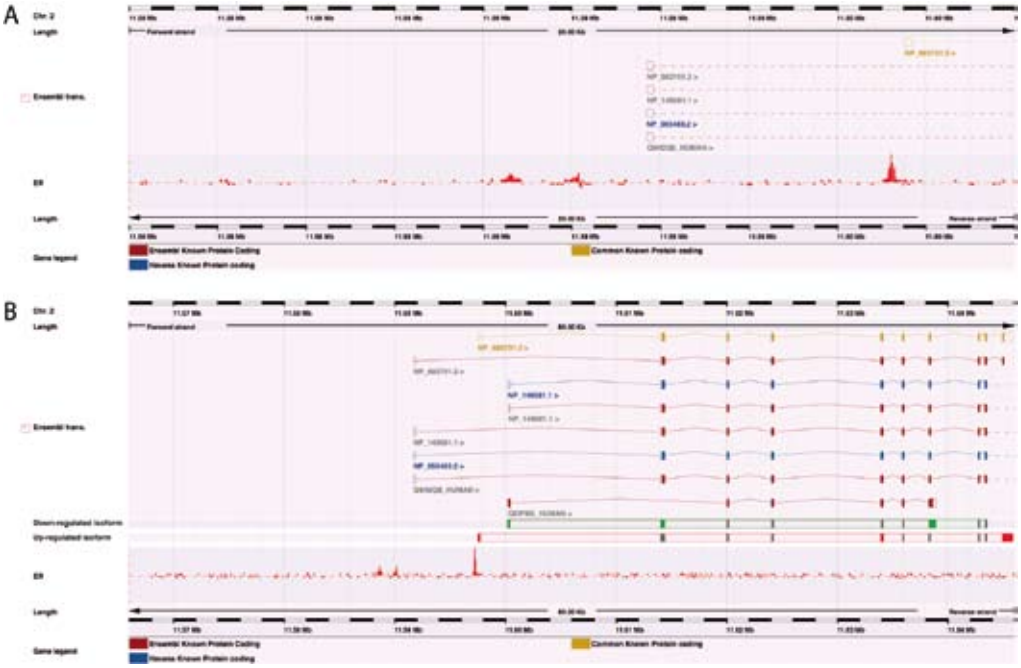


Figure 3. *Top: ChIP-seq analysis of transcription factor binding (red track) reveals occupancy of promoter elements 5' of a known protein-coding target gene. Another binding site is also identified within the gene, although its putative function cannot easily be inferred. Bottom: The same locus with differential expression information, derived from exon array time series analysis. Here we detect a shift in fluorescence levels from several exonic probesets 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.*

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/GP1 30 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 (Figure 4).

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.

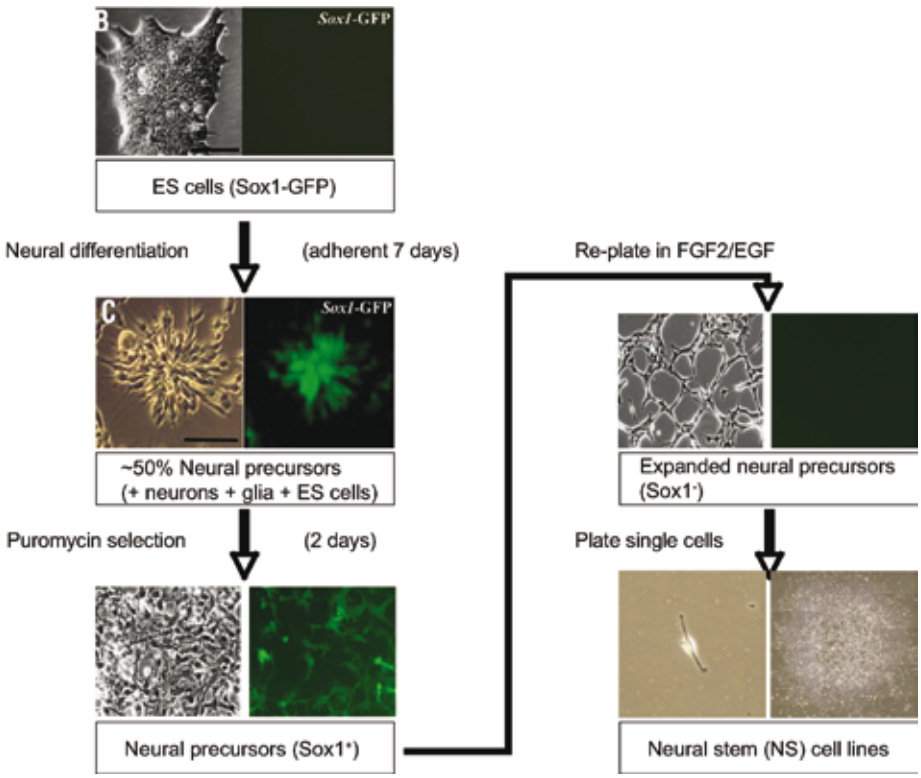


Figure 4. *Derivation of neural cell types from embryonic stem (ES) cells: neural commitment of ES cells and derivation of neuroepithelial precursors (NEPs) after LIF/BMP withdrawal (from embryonic day E4.5 + 0d), followed by selection and expansion of committed NEP to NS cells in the presence of EGF and FGF-2 (+7d).*

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 5). 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 tumor samples. Gliomas are driven by a 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 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.

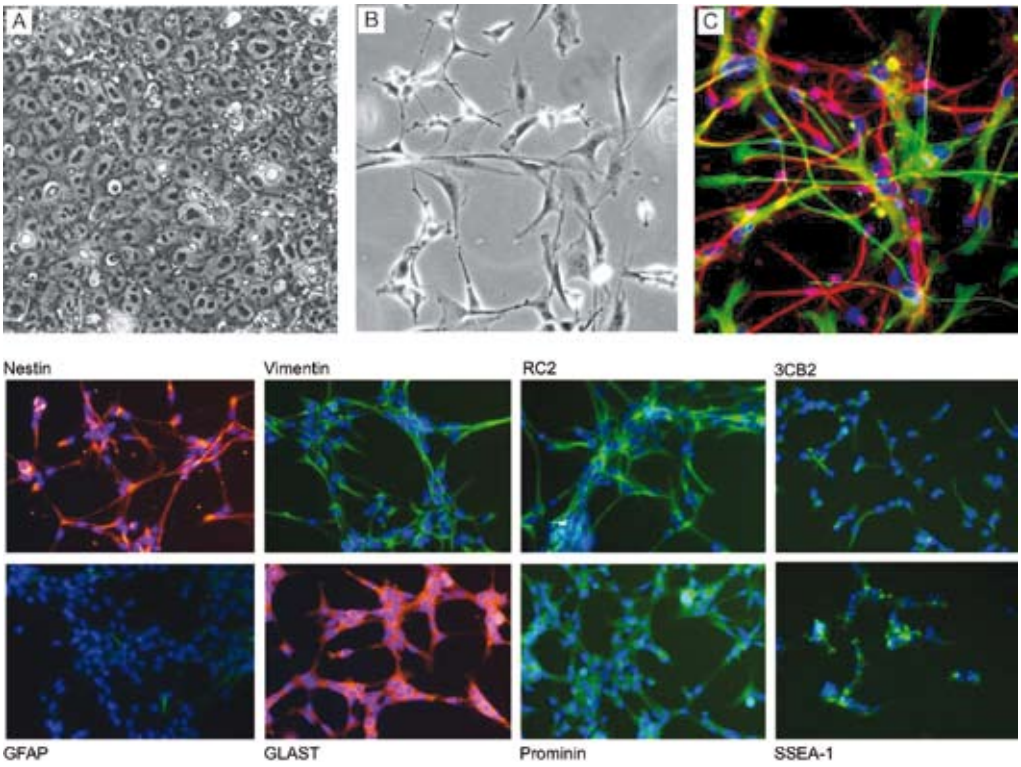


Figure 5. *Top: Differentiation into neural stem (NS) cells from neural-rosette structures. A) ES cell primary culture, B, C) immunostaining for specific surface markers. Bottom: NS cells express markers characteristic of radial glia, permitting both accurate identification of differentiation stages and efficient FACS selection of homogeneous cell populations for genomic analysis. (Images: Steve Pollard, University of Cambridge; adapted from Conti* et al.*, 2005).*

The processes that regulate stem cell differentiation are not well understood and are likely to be mis-regulated in cancer. We are now characterising the differentiation of GNS cells to oligodendrocytes, an event which is positively correlated with patient survival rates in cases of glioblastoma multiforme. During this project we will analyse both GNS and NS cell populations using a combination of real-time and microarray-based expression level measurements, to identify both protein-coding genes and 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 designing microarrays 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 over-expression and knock-down 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. In *Drosophila* 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 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 great 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.

The Goldman Group: evolutionary tools for sequence analysis

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 use this information to make predictions about the function of genomic sequence. One focus of the group is on comparative genomics and the bulk analysis of biological sequence data. Collaborations with major sequencing consortia remain fruitful, providing new data, challenges and a proving ground for new methods of sequence analysis. Intra-group collaborations between members developing theory and methods and those involved in the 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 2008 we have had a productive year, developing methods to infer and visualise evolutionary trees, and methods to use trees to improve sequence alignment. Our research aims to increase our understanding of the process of evolution and to provide new tools for biologists to elucidate the changing function of biological sequences.

PHYLOGENETIC METHODOLOGY

Visualising evolutionary trees

Greg Jordan and Nick Goldman

Greg Jordan has been working on the PhyloWidget software, a web-based tree visualisation program (Jordan & Piel, 2008). Although many tree-drawing programs are currently available, PhyloWidget implements a unique 'highly zoomable' interface, giving the user a constant global context of the tree's shape and composition (Figure 1). PhyloWidget is well-suited for integration with web-based

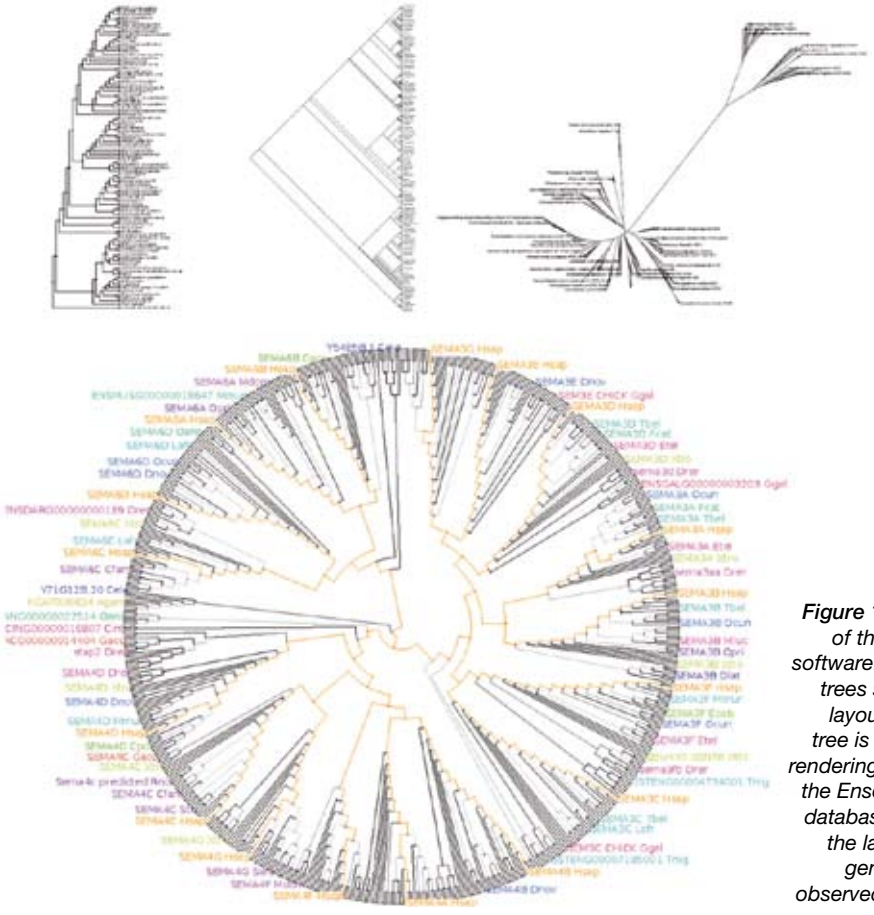


Figure 1. *Screenshots of the PhyloWidget software. The top three trees show standard layouts; the bottom tree is a PhyloWidget rendering of a tree from the Ensembl Compara database, highlighting the large number of gene duplications observed in this family.*

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.

Group Members

Scientists
Ari Löytynoja
Tim Massingham
Martin Taylor

Postdoctoral Fellows
Alexander Alekseyenko*
Emeric Sevin*
Stefan Washietl

PhD Students
Jacky Hess
Greg Jordan
Fabio Pardi

Visitors
Nicolas Canicatti
Benny Chor
Dezsõ Miklós

Scientific Programmer
Nicolas Rodriguez
(shared among several research groups)

**Indicates part of the year only*

Publications

2007
Massingham, T. & Goldman, N. (2007). Statistics of the log-det estimator. *Mol. Biol. Evol.*, 24, 2277-2285

2008
Faller, B., *et al*. (2008). Distribution of phylogenetic diversity under random extinction. *J. Theor. Biol.*, 251, 286-296

Gesell, T. & Washietl, S. (2008). Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics*, 9, 248

Gruber, A.R., *et al*. (2008). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9, 19

Huggett, J.F., *et al*. (2008). Development and evaluation of a real-time PCR assay for detection of Pneumocystis jirovecii DNA in bronchoalveolar lavage fluid of HIV-infected patients. *Thorax*, 63, 154-159

Jordan, G.E. & Piel, W.H. (2008). PhyloWidget: Web-based visualizations for the tree of life. *Bioinformatics*, 24, 1641-1642

Ke, X., *et al*. (2008). Singleton SNPs in the human genome and implications for genome-wide association studies. *Eur. J. Hum. Genet.*, 16, 506-515

Löytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320, 1632-1635

Massingham, T. (2008). Detecting the presence and location of selection in proteins. Pp. 311–329 in *Bioinformatics, vol. 1: Data, Sequence Analysis and Evolution* (J.M. Keith, ed.). Humana Press, Totowa, NJ, USA

Tress, M.L., *et al*. (2008). Determination and validation of principal gene products. *Bioinformatics*, 24, 11-17

databases, giving researchers a simple method for visualising the quickly growing number of phylogenetic trees being generated from large-scale analyses.

The 'Tree of Life' is the name given to the ideal representation of the evolutionary relationships of every species that has lived on earth (Cracraft & Donoghue, 2004; Maddison & Schulz, 2008). As a result of the work on PhyloWidget, we were contacted by the Wellcome Trust and asked to design a model structure of the Tree of Life. Our visualisations are to be used at the heart of a 3D animation to be used in an upcoming David Attenborough documentary celebrating the 200th anniversary of Darwin's birth. In order to strike a balance between scientific realism and visual aesthetics, we created a pipeline that grows, shapes, and prunes the tree based on parameters defined by a team of animators. These animators will subsequently turn the tree into an animated 3D model of the history of life, forming a four to five minute sequence that will be a centrepiece of the documentary, to be released in early 2009.

Inferring evolutionary trees

Jacky Hess, Fabio Pardi, Nick Goldman

A recent subject of the group's research has been the analysis of distance methods for the inference of phylogenetic trees. These are the methods of choice when computational efficiency is critical, for example when there are large numbers of sequences to analyse (a scenario that is becoming more and more common in the genomic era). While the group's past work has focused on the methods for estimating pairwise distances (Massingham & Goldman, 2007), Fabio Pardi directed his research to consider the second step – reconstructing a phylogenetic tree from the estimated distances. A successful criterion for this task is 'balanced minimum evolution' (BME; Desper & Gascuel, 2004), where the aim is to minimise the total length of the reconstucted tree based on a 'balanced' schema for branch length estimation.

The work on BME is being directed both towards the practice and the theory. For the practical side, a 'branch and bound' algorithm, guaranteed to find optimal trees with respect to BME, has been developed and has been used to evaluate the performance of existing heuristic methods such as NJ (Saitou & Nei, 1987) and FastME (Desper & Gascuel, 2002). On the theoretical side, Fabio Pardi proved that simple (although potentially hard to attain) conditions on the precision of the input distances guarantee that the BME-optimal tree is correct. Both these studies are the subjects of manuscripts in preparation.

In times of publicly available full genome sequences covering large parts of the spectrum of organismal diversity, the fundamental phylogenetic problem of reconstructing inter-species relationships has been presented with both new perspectives and new challenges. As opposed to the classical approach where one or a few genes that were considered phylogenetically informative were analysed, datasets of hundreds or even thousands of genes are now at our disposal. 'Phylogenomics', the analysis of such increased amounts of data, promises resolution where single-gene analyses have failed to elucidate relationships between species by overcoming the stochastic effects arising from the sampling error encountered when a small number of characters is analysed.

It is well known, however, that heterogeneity of the evolutionary process, even within a single gene, can affect the accuracy of phylogenetic reconstruction. By using data from different loci across the genomes of several different species, we can only expect this heterogeneity to increase. In a classic study, Rokas *et al.* (2003) concluded that a relatively simple phylogenetic analysis of 106 genes found a well-supported, congruent species tree for seven species of yeast where analysis of individual genes had failed. As increasing amounts of data from yeasts and other organisms have become available, we considered it an appropriate time to review whether this methodology is still valid and practical when applied to larger datasets, incorporating more genes, and more difficult phylogenetic problems, spanning a larger range of evolution.

Jacky Hess collected a dataset of 351 one-to-one orthologues from 18 ascomycetous yeasts, increasing the phylogenetic range to about 250 million years since their last common ancestor. She analysed the coding nucleotide sequences of those genes on a gene-by-gene basis as well as in concatenated forms using different models of evolution and maximum likelihood methods. The single gene analyses confirmed the necessity for a multi-gene approach; indeed, 345 distinct gene trees were recovered from the 351 genes, failing to provide a congruent solution for the species tree. Subsequent analyses used two approaches. The simple concatenation approach, where the entire dataset is analysed with the same parametrisation of the evolutionary model, assumes a constant level of heterogeneity of the

evolutionary process across the different genes in the alignment. In order to account for potential differences between the different genes analysed, we also carried out a partitioned approach, where each gene forms a partition and the parameters of the model are estimated separately for each partition, thereby allowing for heterogeneity between different genes. In addition, we investigated the impact of the particular evolutionary model used both in concatenated and partitioned analysis (Figure 2).

We found the choice of evolutionary model to have a significant impact on the inferred tree, with a different result found for every model tested. All trees had high bootstrap values throughout, suggesting that bootstrap values are not a reliable indicator of the quality of a tree when the dataset analysed is large and that great care should be taken to choose the best possible model. The best model was found to be the most complicated one tested and, to our surprise, even though partitioned analysis clearly outperformed concatenated analysis in terms of yielding much better model fit (see Figure 2), the choice of concatenated vs. partitioned analysis did not affect which tree was found to be optimal in the nucleotide analyses.

We thus obtained a well-supported species tree relating 18 ascomycetous yeasts, which was also confirmed by analysis of the amino acid sequences of the 351 genes. Whilst more data are useful to resolve difficult nodes in the phylogeny, they also reinforce inconsistency of phylogenetic reconstruction when the model used to analyse the data is inappropriate. Thus, emphasis should be placed on better models and understanding of the nature of heterogeneities acting on datasets, rather than on simply incorporating even more data.

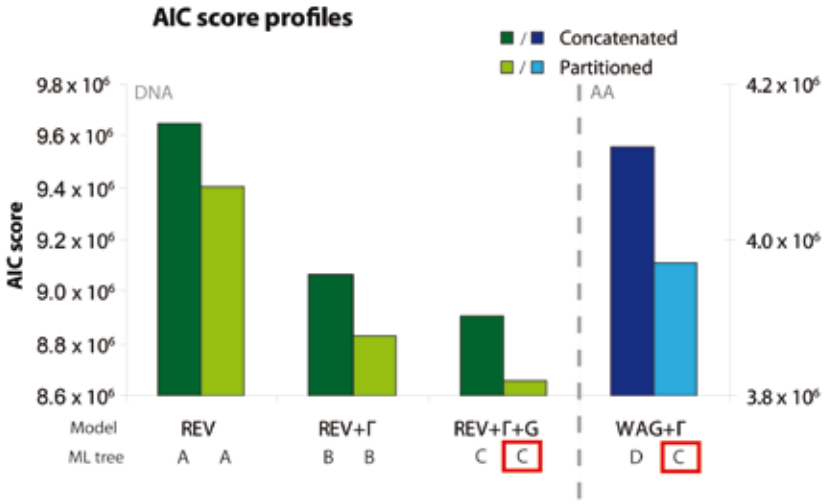


Figure 2. *Results of phylogenomic studies of 18 ascomycetous yeast species. The lower the AIC score, the better the model fits the data. For DNA sequence analyses (left; green), increasingly complex models give better results. The preferred tree, 'C', is also confirmed by analysis of amino acid sequences (right; blue), again using the preferred, more complex, model.*

Phylogenetically-inspired weighting schemes

Alexander Alekseyenko and Nick Goldman

The study of 'identity by descent', the equality of some trait in different organisms because it has been inherited unaltered from their common ancestor, has found many uses in statistical genetics. However, the utility of this important quality has not been as fully explored in phylogenetic and sequence analysis applications. Alexander Alekseyenko and Nick Goldman have revisited the concept of identity by descent, and shown its applicability for problems such as motif description and profile building where full evolutionary analyses are mathematically or computationally difficult, but where we do not wish to ignore the evolutionary relationships of the sequences under study. We have derived efficient algorithms for computing the expected number of sequence characters identical by descent, via a forward-backward dynamic programming approach (Durbin *et al.*, 1998).

In analysis of sets of related nucleotide and protein sequences, we are interested in how much independent information each sequence contributes to this set. For example, if two homologous characters are an infinite evolutionary distance apart then they represent two independent observations, each contributing one unit of information. At the other extreme, two observations separated

Whelan, S. (2008). New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Systematic Biology*, 56, 727-740

Whelan, S. (2008). Inferring trees. In *Bioinformatics* (Keith, J.M., ed) 287-309, Humana Press

Other EMBL publications

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature,* 450, 203–218

Flicek, P., *et al*. (2008). Ensembl 2008. *Nucleic Acids Res.,* 36, D707–714

Massingham, T. & Goldman, N. (2005). Detecting amino acid sites under positive and purifying selection. *Genetics*, 169, 1753–1762

Massingham, T. & Goldman, N. (2007). Statistics of the log-det estimator. *Mol. Biol. Evol.,* 24, 2277–2285

Washietl, S. *et al*. (2007). Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*. 17, 852–864

Other publications

Altschul, S. F. & Erickson, B.W. (1985). Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, 2, 526–538

Cracraft, J. & Donoghue, M.J., eds. (2004). *Assembling the Tree of Life.* Oxford University Press, USA

Desper, R. & Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comp. Biol.,* 9, 687–705

Desper, R. & Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.*, 21, 587–598

Durbin, R., *et al*. (1998). *Biological Sequence Analysis.* Cambridge University Press, Cambridge, UK

Gesell, T. & von Haeseler, A. (2006). In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics,* 22, 716–722

Maddison, D.R. & Schulz, K.-S., eds. (2008). *The Tree of Life Web Project.* http://tolweb.org

Rokas, A., *et al*. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425, 798–804

Saitou, N. & Nei, M., (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4, 406–425

Taylor, M.S., *et al*. (2006). Heterotachy in mammalian promoter evolution. *PLoS Genet*., 2, e30

Washietl, S. & Hofacker, I.L. (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, 342, 19–30

Whitehouse, I., *et al*. (2007). Chromatin remodelling at promoters suppresses antisense transcription. *Nature,* 450, 1031–1035

Wray, G.A. (2003). Transcriptional regulation and the evolution of development. *Int. J. Dev. Biol.*, 47, 675–684

by exactly zero evolutionary divergence are by necessity identical and contain in total a single unit of information. Identity by descent provides an effective metric for the amount of information in a set of homologous characters. Two characters are said to be identical by descent if there has been no mutation along the evolutionary path that separates them. If a character is expected to be identical by descent to many other characters in the set due to common ancestry, then this character carries much less information than a character with a smaller expected identity. We have proposed the use of the reciprocal of the expected number of characters identical by descent as a measure of the amount of independent information a character contributes.

This quantity may serve as a natural weighting scheme for sequence profile building and other analyses that need to account for similarity biases, such as the representation of shared motifs where similarity in motif elements between divergent sequences is more informative than similarity in closely related sequences. The particular interest of this weighting scheme for profile building is that this is the first model-based way of incorporating phylogenetic information into this problem.

Accurate null models for comparative RNA gene prediction

Stefan Washietl

Any experiment is only as good as its controls. What is true for experimental biology clearly holds also in the field of comparative genomics; the value of even the most sophisticated algorithm remains unclear if the significance of the results cannot be assessed properly. For many applications in comparative genomics, an accurate null model in the form of randomised alignments is helpful to assess the significance of the results. For example, in screens for non-coding RNAs, evolutionarily conserved RNA structures are predicted in multiple alignments of whole genomes. To estimate the false discovery rate in such screens, the analysis is repeated with randomised alignments. The detailed analysis of the ENCODE regions (Washietl *et al.*, 2007) drastically demonstrated that overly simplistic randomisation strategies (e.g. by näive shuffling of alignment columns) can lead to underestimation of the true false-discovery rate.

RNA folding algorithms use the so-called nearest neighbour energy model to predict thermodynamically favourable secondary structures of low free energy. Under this model, energies are not assigned to single base pairs but rather to neighbouring base pairs that stack on each other. As a consequence, the folding stability of genomic sequences does not only depend on the mononucleotide content but also the dinucleotide content. While there have been algorithms to randomise single sequences preserving dinucleotide content for more than 20 years (e.g. Altschul & Erickson, 1985), the problem remained challenging for multiple alignments.

In collaboration with Tanja Gesell (Center for Integrative Bioinformatics, Max F. Perutz Laboratories, Vienna), Stefan Washietl developed a new algorithm to produce dinucleotide-controlled random alignments (Gesell & Washietl, 2008). Building upon Tanja Gesell's previous work (Gesell & von Haeseler, 2006), random alignments are simulated along a phylogenetic tree using a substitution model that considers overlapping dependencies between sites. This allows exact control of the equilibrium dinucleotide content in the generated datasets. In addition, the model considers specific rates for each site to account for rate heterogeneity in genomic data, e.g. caused by conserved motifs. Fast heuristics and a distance-based approach are used to estimate a phylogenetic tree under this complex evolutionary model. The approach is implemented in a program called SISSIz (http://sourceforge.net/projects/SISSIz) that was shown to generate accurate randomisations for genome-wide alignments typically used in comparative studies. In particular, the novel approach leads to more accurate (and more conservative) estimates for the false-discovery rate in screens for structural RNAs.

Following a similar strategy as used by Washietl & Hofacker (2004) in the program AlifoldZ, the new simulation algorithm was directly combined with an RNA consensus folding algorithm. SISSIz thus not only serves as a randomisation program but also as a new variant of an RNA gene finder – the first that is not biased by the dinucleotide content of the input sequences.

Insertions, deletions and structure in multiple sequence alignments

Ari Löytynoja and Nick Goldman

The starting point of any evolutionary sequence analysis is the sequence alignment that aims to represent the ancestral homology among the characters. Many widely-used alignment methods are, however, designed to match structurally corresponding, conserved protein regions and often produce evolutionarily implausible solutions for variable regions, inferring incorrect homologies and seri-

ously underestimating the true numbers of insertions and deletions. Ari Löytynoja has developed a phylogeny-aware alignment algorithm that is less affected by these problems (Löytynoja & Goldman, 2008) and, according to simulation studies, produces nearly unbiased alignments regarding the numbers and placements of alignment gaps (Löytynoja & Goldman, in press). This method has been implemented in Löytynoja's programs PRANK and PRANKSTER, the latter providing an easy-to-use graphical interface for running the program and visualising the resulting alignments (Figure 3).

As described above, in phylogenetic inference the modelling of heterogeneity across sequence sites is widely used. Methods for sequence alignment, in contrast, have mostly assumed homogeneous sequences with all sites evolving under the same constraints. Ari Loÿtynoja has developed an alignment method that allows for defining multiple evolutionary processes using probabilistic models (Löytynoja & Goldman, in press). In addition to more realistically describing the evolution of different sequence regions, the approach provides a simultaneous inference of sequence features across the sequence sites (Figure 4). Furthermore, the algorithm is useful not only for creating alignments but can also be used to analyse existing alignments. Whereas phylogenetic inference methods typically ignore alignment gaps and thus discard significant information, Löytynoja's algorithm takes these into account and is highly suitable for comparative analyses of genomic sequences such as promoter regions and control elements, and allows the study of selective changes across evolutionary time and lineages.

The impact of deletions on phylogenetic analysis

Martin Taylor and Nick Goldman

Phylogenetic analysis of biological sequences has typically focused on substitution changes, with sites involved in insertions or deletions being discarded. This is an approach that has worked extremely well in the analysis of protein-coding sequence evolution, for example in the analysis of the ratio (typically denoted ω) of non-synonymous (amino acid changing) to synonymous (non-amino acid changing) substitutions in order to study selective pressures acting on protein-coding DNA. Such methods are increasingly being generalised for application to non-coding sequence. However, some of the assumptions that are valid in the analysis of protein-coding sequence are often violated in the analysis of non-coding sequence. In particular, deletions in non-coding sequence have the potential to disrupt such tests and could lead to spurious estimates of selection.

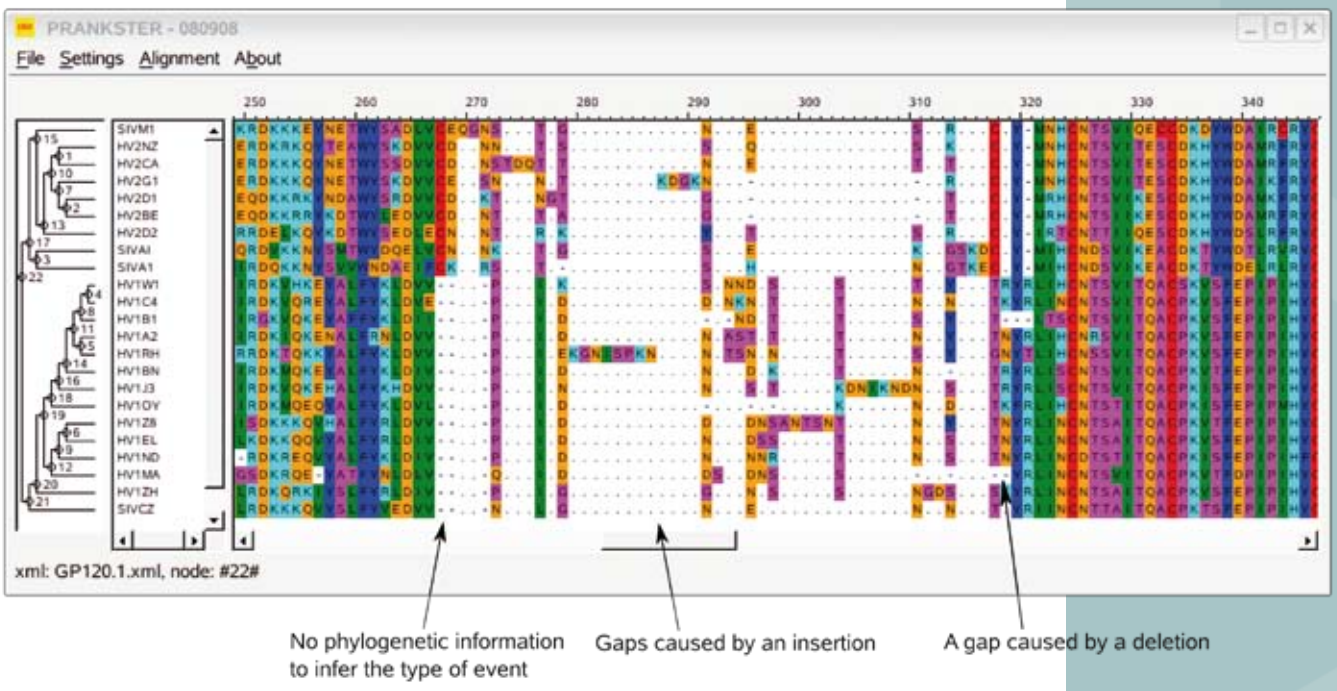


Figure 3. *Screenshot of the PRANKSTER software for visualising multiple sequence alignments. The evolutionary tree relating the sequences, without which the evolutionary meaning of the alignment cannot be fully interpreted, is visible to the left, and various interesting insertion and deletion patterns are indicated (bottom).*

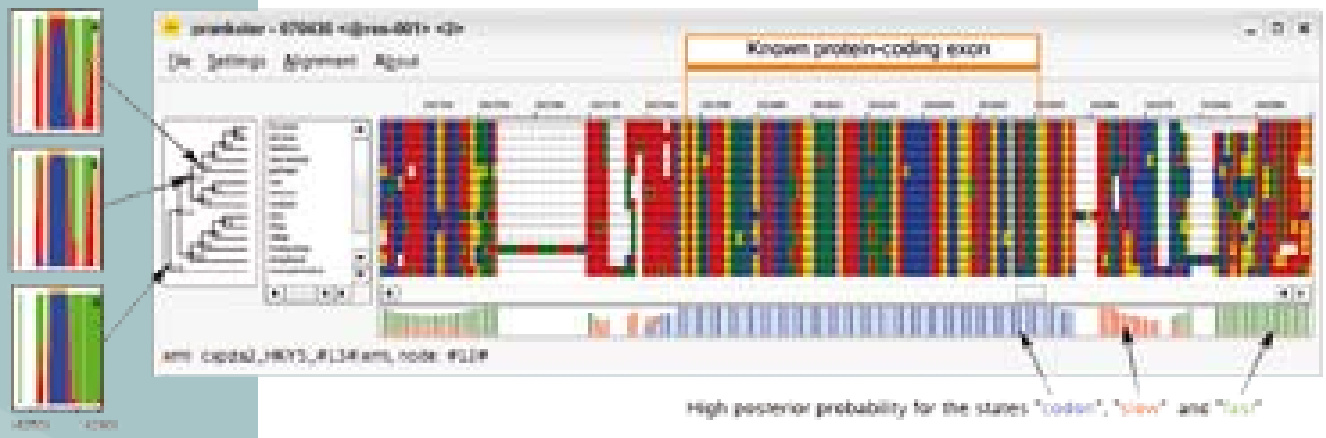


Figure 4. Simultaneous alignment and structural analysis of the CAPZA gene. Below the central alignment is the inferred structure of protein coding (codon; blue), slow non-coding (red) and fast (green) non-coding sequence in a genomic region. Notice that the inferred coding region is in excellent agreement with the position of a known exon. To left, similar structural inferences are shown for different points in the evolutionary history of the sequences.

Using a combination of simulation studies and real biological data, Martin Taylor and Nick Goldman are exploring the impact of deletions on ω-like tests of selection. We find that under some biologically plausible scenarios, ω-like estimates can be dramatically distorted by deletions. Understanding this problem has allowed us to formulate two possible corrections and led to a new test of neutral evolution for non-coding sequence.

GENOME EVOLUTION

Genome-wide turnover of transcription factor binding sites in *Drosophila*

Emeric Sevin, Ari Löytynoja, Martin Taylor

It is now widely accepted that evolution of regulatory elements driving gene expression has a major role in the variation of morphological features of species (Wray, 2003). Experimental studies of specific, well-studied systems have given excellent anecdotal examples of the evolutionary processes in play, such as the dynamic gain and loss of transcription factor binding sites ('TFBS turnover'). However, a genome-scale view of the influence of binding sites structure and evolutionary changes on function remains to be given.

The recent availability of twelve sequenced drosophilid genomes (*Drosophila* 12 Genomes Consortium, 2007) provides a good means to assess the evolution of regulatory sequences along a given lineage. Focusing on the core promoters of all genes of the *D. melanogaster* genome, we are investigating the turnover of TFBS through the course of drosophilid evolution. Using various multiple alignment methods, one of which (developed in the group and described above) handles insertion/deletion events more appropriately than former algorithms (Löytynoja & Goldman, 2008), we were able to produce various sets of alignments of the promoters and reconstruct their corresponding ancestral sequences. In collaboration with Jüri Reimand (from the Luscombe group at the EBI), known TFBS were then mapped in the orthologous regions for each species and at ancestral nodes in the tree. While assessing how the choice of alignment method affected the subsequent analysis, we considered insertions, deletions, substitutions and rearrangements to estimate their impact on predicted TFBS, with particular interest in the sequence of evolutionary events leading to the replacement or conservation of functional TFBS. This work will also allow us to investigate spatial constraints within drosophilid promoter regions and relate both TFBS turnover and more general promoter constraints to the function and expression of the downstream gene.

Mammalian promoter evolution

Greg Jordan, Tim Massingham, Martin Taylor, Nick Goldman

Promoter regions provide key regulatory sequences and a molecular staging post necessary for the expression of genes. They are also a category of non-protein coding genomic sequence that can be readily identified using functional genomic data. This makes them an ideal testing ground for the development and application of comparative genomic methods to non-coding sequences. We have

continued a long-standing interest in promoter evolution by investigating how estimates of selection in promoter regions can be distorted by inappropriate neutral evolutionary rate estimates. This work has supported earlier conclusions that primate promoter regions have a higher rate of mutation than other regions of the genome (Taylor *et al.*, 2006), and that an elevated mutation rate could be misinterpreted as the signal of positive selection in comparative genomic analyses (Taylor *et al.*, in press).

Continuation of this work has investigated the biological processes that can lead to the apparent fluctuation in mutation rate between even closely separated genomic regions. Factors such as genomic context, sequence composition and expression pattern of the downstream gene have all been considered, as well as substitution rate heterogeneity between lineages. Insight from this study and results from the short transcript clustering and analysis work reported on last year have contributed to an internationally collaborative manuscript that has been submitted for publication. Results from our study of the impact of deletions on ω-like estimates described above will also feed into the further continuation of this work. More generally, the use of promoter regions as a model system to investigate non-coding sequence evolution has been explored in a review that is soon to appear (Taylor *et al.*, in press).

Genome-wide detection of selective constraint

Greg Jordan, Tim Massingham, Nick Goldman

The SLR method and software for inferring the pressures of natural selection in peptide-coding DNA sequence was previously developed in the group (Massingham & Goldman, 2005). It operates by analysing the patterns of synonymous and non-synonymous substitutions in coding DNA, embodied by the ratio ω as described above. Tim Massingham has worked with the Ensembl team at the EBI to integrate SLR into the Ensembl annotation pipeline (Flicek *et al.*, 2008), so that estimates of the level of purifying selection (conservation) and positive selection are now automatically made for all vertebrate protein alignments. Currently this information is accessible via the Ensembl databases, but in future it will be integrated directly into the genome browser to allow easy comparison with other peptide features. Greg Jordan has been using this information for a genome-wide survey of selective evolutionary pressures at the level of individual sites of protein sequences, relating the results to protein features such as domains, secondary structure and active sites (Figure 5).

The impact of nucleosome positions on the evolution of genomic DNA

Stefan Washietl and Nick Goldman

A driving force behind the group's research is the belief that for phylogenetic modelling and comparative genomics it is essential to have a good understanding of the evolutionary forces that change DNA over time and shape the genomic landscape. Recently published high-resolution maps of nucleosome positions in the yeast genome (Whitehouse *et al.*, 2007) made it possible to address an as-yet unexplored question in this context – is the molecular evolution of DNA affected by its packaging in the nucleus? Stefan Washietl and Nick Goldman, in collaboration with Rainer Machné (Institute for Theroretical Chemistry, Vienna) analysed substitution rates in the yeast genome dependent on the location within maps of nucleosome positions (Washietl *et al.*, in press). A statistically significant

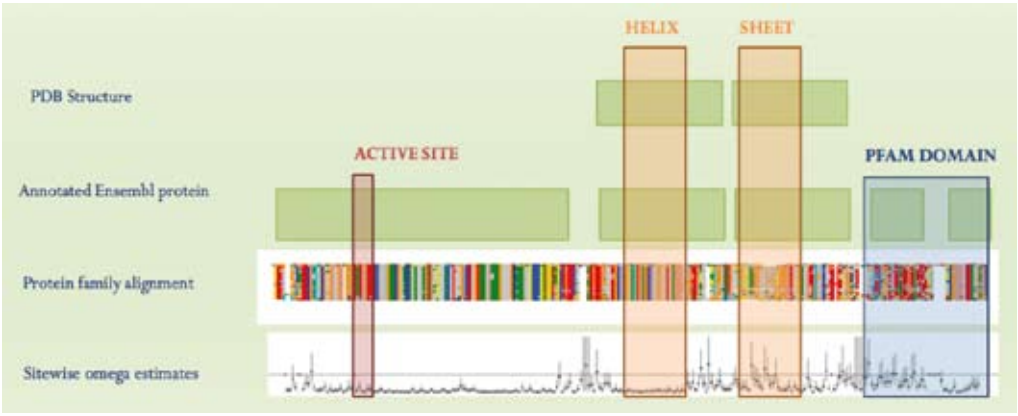


Figure 5. *Inferring selection in a protein, and comparison with Ensembl peptide annotations. Selection ('sitewise omega estimates') is determined for every column in the protein family alignment, unusually conserved or variable sites are indicated, and comparisons made with annotations regarding the peptide's function derived from other sources. Selection is shown as a point estimate with confidence interval.*

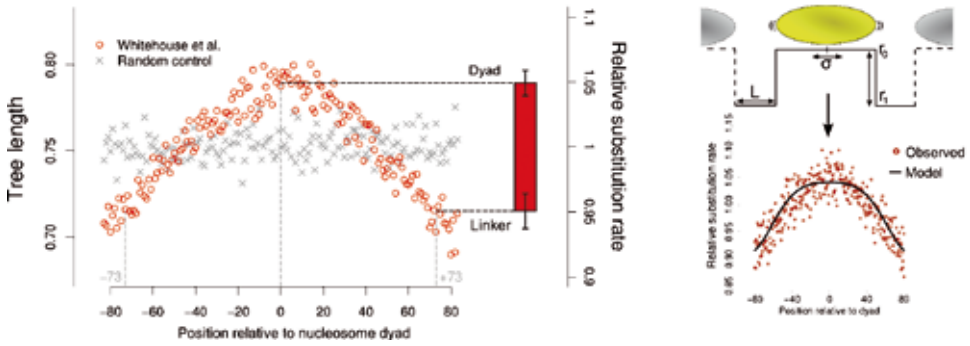


Figure 6. *Left: The length of the phylogenetic tree of* S. cerevisiae *and four closely-related yeast species was estimated using sequence data that correspond to specific relative positions in the nucleosome map of Whitehouse* et al. *(2007). The tree length (red circles) is used as measure for the substitution rate. A strong positional dependency ($P<10^{-25}$) can be observed. Substitution rates are highest around the dyad (the equidistant centre point of nucleosomal DNA) and lowest towards the linker (around positions +/-73). The difference is approximately 10% (red bar). Right: A simple model assuming two different rates for the linker (r_0) and nucleosomal DNA (r_1) and a level of uncertainty for the experimental nucleosome positions (σ) can explain the observed patterns.*

dependency could be observed (Figure 6); 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. It thus cannot be explained by the fact that transcription factor binding sites are enriched in the more accessible linker regions, which has been observed before. Another possible explanation is that selective constraints from nucleosome positioning signals (the 'nucleosome code') are responsible for higher conservation in linker regions. However, no obvious connection to any known positioning signals could be observed. Given the current knowledge, the most likely explanation for the nucleosome-related rate heterogeneity seems to be general differences in mutation rates between linker and nucleosomal DNA. A simple model assuming two distinct rates and taking into account the error in the nucleosome maps from biological or experimental noise, can explain the pattern of the observed substitution rates (Figure 6).

Interestingly, it is well-established that most known DNA repair mechanisms are impaired by nucleosomes and are more efficient in naked DNA. This would be a very intuitive explanation for this effect backed by experimental evidence. However, many other molecular or cellular mechanisms could be involved. Regardless of the mechanisms, the observation has some interesting implications. The fact that these differences can be observed when comparing contemporary yeast species that diverged *c.*20 million years ago implies that nucleosome positions are conserved. 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.

HIGH-THROUGHPUT SEQUENCING

Improved quality of high-throughput sequencing by computational methods

Tim Massingham and Nick Goldman

Next-generation sequencing machines have provided a huge increase in sequencing capacity, both for large genome centres and for smaller laboratories worldwide. Vastly less time (and money) is needed to sequence a genome: whereas the Human Genome Project took over ten years to produce a draft sequence, next-generation technology is allowing the 1000 Genomes Project (http://www.1000genomes.org/) to sequence the genomes of 1,000 individuals within two years. Small institutes or large research groups can now afford sequencing capacity that outstrips that available at the major sequencing centres during the Human Genome Project.

But next-generation sequencing technology is not without problems. For one popular next-generation sequencing technology, the Illumina Genome Analyzer (GA) platform, rather than producing a few long reads of sequence, the output is millions of short reads, generally less than 50 nucleotides long, with an error rate that increases considerably as the length of the read increases. A large contribution to this error is termed 'phasing'. The platform relies on repeated cycles of incorporation

of a single new nucleotide by synthesis onto the next position of each molecule in a large cluster of homogeneous DNA but, as sequencing progresses, chemistry errors mean that individual molecules get out of step, falling behind or getting ahead, and so the overall signal loses coherency.

We have developed new statistical techniques to describe how phasing error changes from cycle to cycle. These allow the error to be corrected for with considerably more reliability than the software provided with the GA platform. Reads processed with our techniques, embodied in the software 'AYB', show a marked increase in quality (Figure 7). This increases the number of good reads that can be extracted from the machines and reduces the overall error rate, resulting in more, higher quality sequence from the same data. The AYB software is currently being tested by the EMBL Genomics Core Facilities and several other laboratories.

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, and the group is confident that we will continue to contribute to this theoretical field of research. 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.

In 2009, next-generation sequencing technologies will produce almost as much sequence data as was produced in total prior to that date. For groups like ours that work on methods for sequence analysis, it is necessary to remember that the questions biologists want to ask of these data will change as more diverse experiments become practical. We have started to address this through our work on understanding the actual generation of data from next-generation sequencing platforms, and we will continue to work to allow the greatest benefit from modern molecular biology.

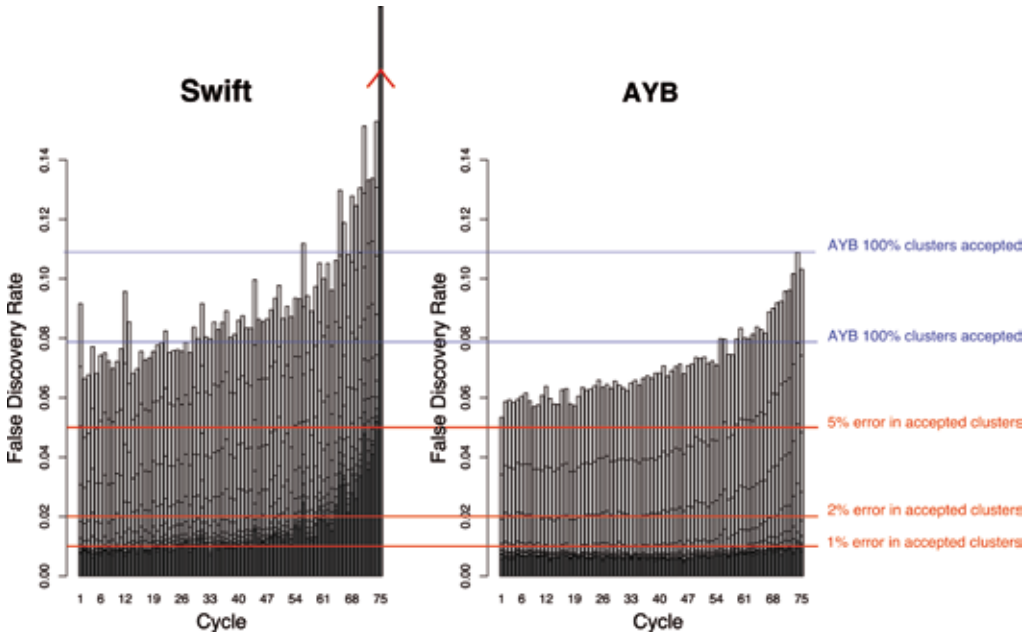


Figure 7. Comparison of error rates from different analyses of data from an Illumina GA sequencer. Left: Results from the 'Swift' analysis package devised at the Sanger Institute. Right: Results from our AYB software. The x-axes represent nucleotide incorporation cycles or positions along a read. At each cycle, the height of the tip of the bar represents the percentage of reads that contain an error in that cycle; here, AYB is strictly better than Swift for all cycles. The subdivision of the bars shows how the error rate decreases when uncertain reads are discarded (in groups of 10%; e.g., the first bar below the tip is the error when 90% of the most certain reads are kept).

The Huber Group: functional genomics

Wolfgang Huber

PhD, 1998, University of Freiburg.
Postdoctoral researcher at IBM in San Jose, California and at the German Cancer Research Centre (DKFZ), Heidelberg.
At EMBL-EBI since 2004.
Joint appointment with Gene Expression Unit, EMBL Heidelberg.

INTRODUCTION

The Huber group develops advanced mathematical and statistical methods for the understanding of functional genomics data and the elucidation of complex phenotypes from genetic networks. We work with experimental labs in systems genetics and functional genomics to develop the best methods for designing and analysing genome-wide experiments whose aim is to unravel the mechanisms of genetic inheritance, gene expression, signal transduction and how they shape phenotype. Most phenotypes, including human diseases, are complex, i.e., they are governed by large sets of genes and regulatory elements. Our aim is to map these complex networks and eventually, to devise strategies for designing phenotypes by engineering combinatorial perturbations.

Our research is driven by new technologies, and we employ data from high-throughput sequencing (ChIP-seq, RNA-seq), tiling microarrays, large-scale cell-based assays, and automated microscopy, as well as the most advanced methods of computational statistics. We are a regular contributor to the Bioconductor project (www.bioconductor.org).

HIGH-RESOLUTION MAPPING OF MEIOTIC CROSSOVERS AND NON-CROSSOVERS IN YEAST

Richard Bourgon and Alessandro Brozzi in collaboration with Eugenio Mancera and Lars Steinmetz at EMBL Heidelberg

Meiotic recombination plays a central role in the evolution of sexually reproducing organisms. The two recombination outcomes, crossover (CO) and non-crossover (NCO), increase genetic diversity, but have the potential to homogenise alleles by gene conversion. While CO rates are known to vary considerably across the genome, NCOs and gene conversions have only been identified in a handful of loci. To examine genome-wide recombination at high spatial resolution, we generated maps of COs, CO-associated gene conversions and NCO gene conversions using dense genetic marker data collected from all four products of 56 yeast meioses. Our maps reveal differences in the distributions of COs and NCOs, showing more regions where either COs or NCOs are favoured than expected by chance. Furthermore, we detect evidence for interference between COs and NCOs, a phenomenon previously only known to occur between COs. Up to 1% of the genome of each meiotic product is subject to gene conversion in a single meiosis, with detectable bias towards GC nucleotides. The maps represent the first high-resolution, genome-wide characterisation of the multiple outcomes of recombination in any organism. In addition, because NCO hot spots create holes of reduced linkage within haplotype blocks, our results stress the need to incorporate NCOs into genetic linkage analysis (Figure 1).

SYSTEMS ANALYSIS BY QUANTITATIVE CELLULAR ASSAYS, NETWORK MODELLING AND INTEGRATION OF METADATA

Elin Axelsson in collaboration with Thomas Horn and Michael Boutros at DKFZ Heidelberg

Genetic experiments associate genotype with phenotype. However, similar phenotypes can arise from different pathways, multiple phenotypes can originate from the same pathway, and gene buffering can obscure the relationship between a gene and a pathway. The goal of our project is to perform a coherent and comprehensive examination of genetic interactions for model pathways of cell death, differentiation, and morphogenesis. We are conducting an integrated series of experiments to identify synthetic genetic interactions. We are also interested in identifying synthetic and antagonistic genetic interactions between specific gene pairs. For this we perform large-scale systematic combinatorial perturbations. Computational methods have been developed for quality assessment, analysis and statistical modelling of the resulting data (Figure 2).

SYSTEMATIC CHARACTERISATION OF GENE FUNCTION BY MULTIPARAMETRIC CELLULAR DESCRIPTORS AND GENOME-WIDE RNAI

Gregoire Pau, Oleg Sklyar, Wolfgang Huber, in collaboration with Florian Fuchs, Christoph Budjan, Sandra Steinbrink, Thomas Horn, Angelika Pedal and Michael Boutros at DKFZ Heidelberg

While an increasing number of genomes are sequenced, the functions of many genes remain unknown. Genetic screens for phenotypes on the level of the organism have been successfully used

Group Members

Staff Scientists
Richard Bourgon
Bernd Fischer*
Audrey Kauffmann
Daniel Murrell*
Gregoire Pau
Oleg Sklyar*

Marie Curie Fellow
Simon Anders

PhD Students
Elin Axelsson
Tony Chiang
Jörn Tödling

Visitors
Remy Clement
Kristen Feher
Ramona Schmid

** Indicates part of the year only*

Acknowledgements
EMBL
EU
Human Frontier Science Programme

Publications

2007

Casneuf, T., *et al*. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics*, 8, Article 461

Chiang, T., *et al*. (2007). Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol*., 8, Article R186

Huber, W., *et al*. (2007). Graphs in molecular biology. *BMC Bioinformatics*, 8, Article S8

2008

Chiang, T., *et al*. (2008). Rintact: Enabling computational analysis of molecular interaction data from the IntAct repository. *Bioinformatics*, 24, 1100-1101

Fischer, J.J., *et al*. (2008). Combinatorial effects of four histone modifications in transcription and differentiation. *Genomics*, 91, 41-51

Hahne, F., *et al*. (2008). *Bioconductor Case Studies.* Springer

Li, X.Y., *et al*. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.*, 6, 0365-0388

Lin, S.M., *et al*. (2008). Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res*., 36, Article e11

Mancera, E., *et al*. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454, 479-485

Sarasin, A. & Kauffmann, A. (2008). Overexpression of DNA repair genes is associated with metastasis: A new hypothesis. *Mutation Research*, 659, 49-55

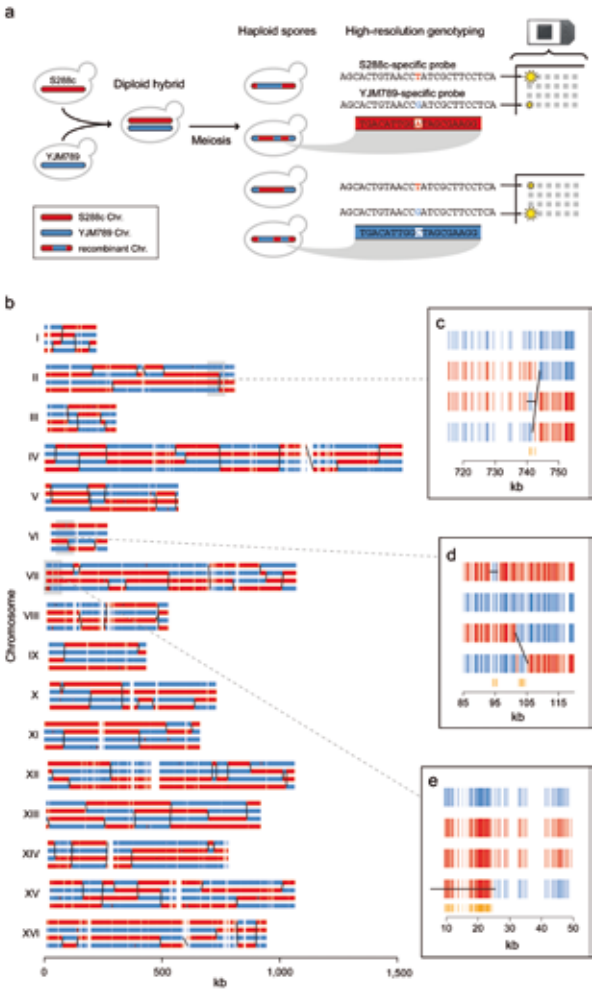


Figure 1. *High-resolution mapping of meiotic recombination along the yeast genome. (a) Schematic description of recombination mapping approach. Meiotic products from a hybrid derived from highly polymorphic strains were individually genotyped using microarrays. (b) Genotype calls at ~52,000 markers for all four spores resulting from a single meiosis. Diagonal/vertical black lines are inferred COs, and horizontal lines, NCOs. Close-ups of (c) a CO overlapped by an independent NCO in a third spore; (d) a CO with complex gene conversion tract, and a nearby, independent NCO; and (e) a long NCO at the end of the chromosome. In close-ups, orange vertical segments denote markers with non-Mendelian ratios (1:3 or 0:4).*

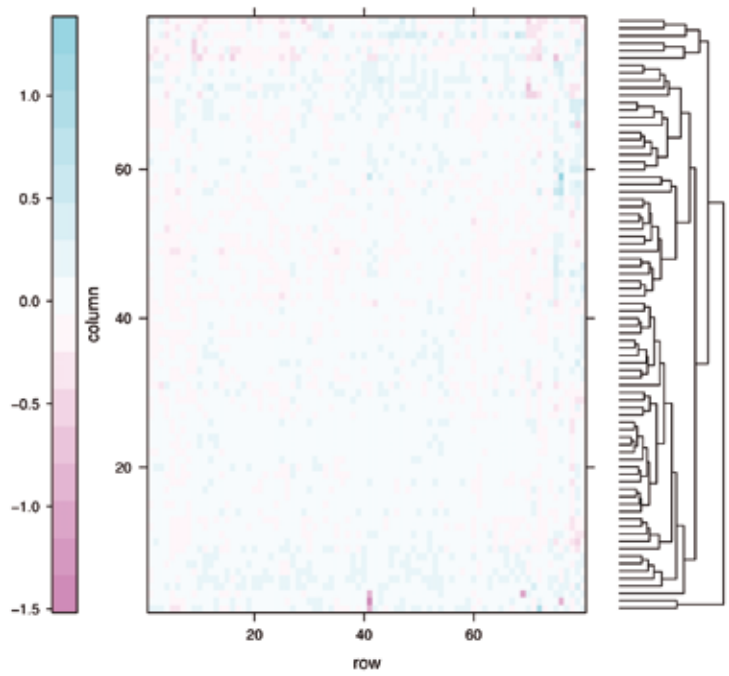


Figure 2. *Heatmap of genetic interaction data. An embryonal fly cell line was treated with all pairwise combinations of 84 RNAi reagents (dsRNAs), each specifically targeting a phosphatase, and a continuous-valued measure of cell viability was registered as the phenotype. Each element of the matrix corresponds to a pair of reagents, and the colour encodes the synthetic effect: white corresponds to no interaction (the effect of the double knock-down is the same as the sum of the effects of the individual knock-downs), red to synergistic interaction (the effect of the double knock-down is stronger than the sum of the single gene effects), blue to antagonistic interaction. Rows and columns of the matrix were arranged by a clustering method.*

to characterise the function of genes and order their action into cellular pathways. Perturbation tools, such as RNAi, allow the phenotypic characterisation of genes on a genome-wide scale in cellular models. However, a rapid and robust method for measuring multi-parameter phenotype profiles suitable for large-scale perturbation experiments has been lacking. Here, we mapped the phenotypic profile space of gene perturbations based on morphological signatures of single cells. After depletion of almost every gene in the human genome, more than six million cells were individually classified to generate a multidimensional phenotypic descriptor for each RNAi experiment. A machine-learning technique was developed to extract networks from the phenotypic descriptors. We found that the perturbations of several uncharacterised genes lead to multivariate phenotypes similar to those observed for genes involved in DNA damage response pathways. This study shows that multiparametric phenotyping by imaging is an efficient approach to associate new genes with known functional modules on a genome-wide scale (Figure 3).

Scholtens, D., *et al*. (2008). Estimating node degree in bait-prey graphs. *Bioinformatics*, 24, 218-224

Other EMBL publications

Toedling, J. & Huber, W. (2008). Analyzing ChIP-chip data using bioconductor. *PLoS Comput Biol*., 4, e1000227

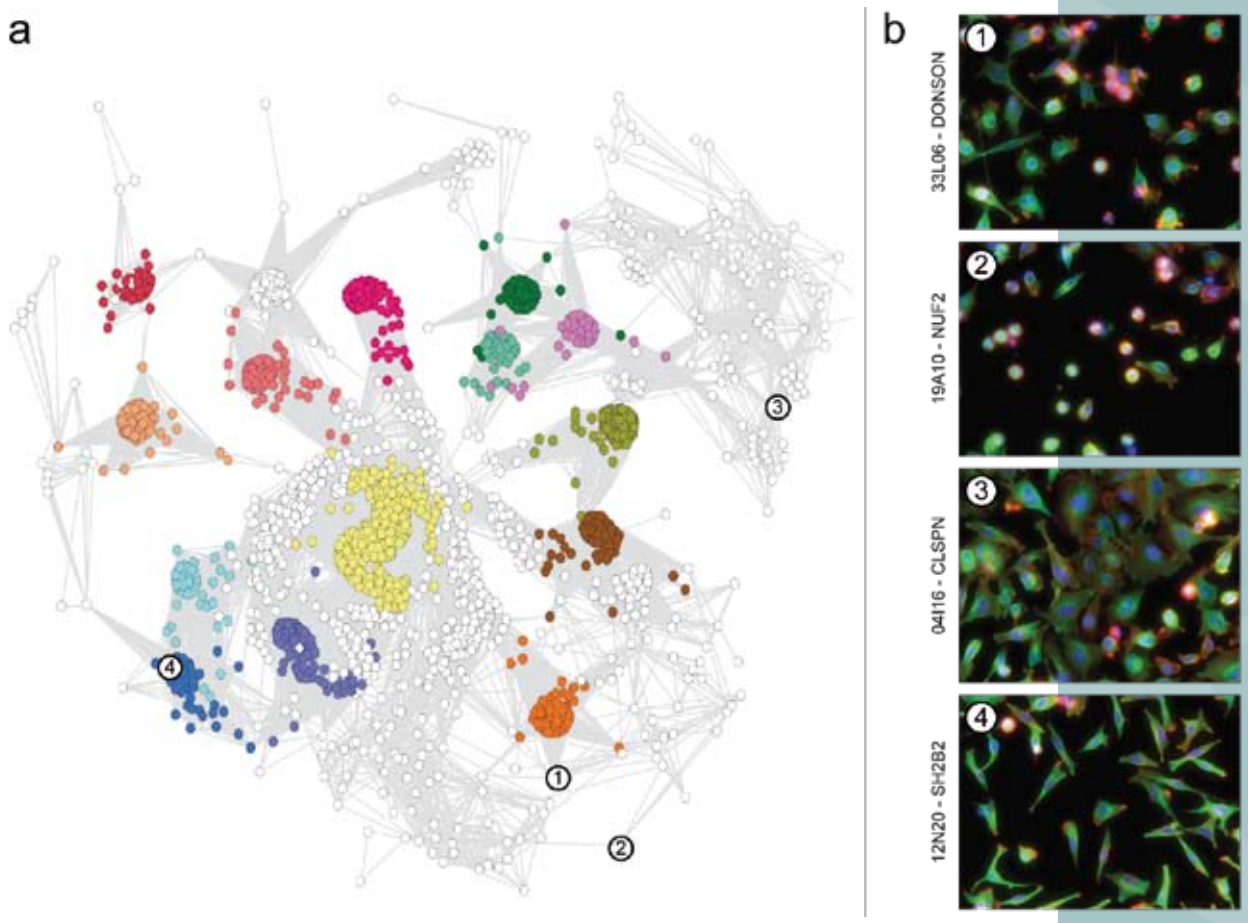


Figure 3. *Genome-wide phenotypic map. (a) Each of the 1,839 nodes represents a perturbation. Nodes are linked by a grey edge when they have a small phenotypic distance. Cliques of nodes with small distances form clusters and are coloured according to their representative phenoprint. This 2D graph rendering provides an overview over the variety of phenoprints and helps define groups of phenotypically similar genes. (b) Representative images of cell populations from different regions of the graph show the similarity of phenotypes of neighbouring genes, and the gradual variation of phenotypes on paths through the graph.*

CELLHTS2: END-TO-END ANALYSIS OF RNAI-SCREENING DATA

In collaboration with Florian Hahne and Robert Gentleman at FHCRC, Seattle, Michael Boutros, DKFZ Heidelberg and Amy Kiger, UC San Diego

The Bioconductor software package cellHTS2 provides a comprehensive solution for the preprocessing, normalisation, quality assessment, hit scoring and reporting of experimental data from RNA interference or molecular compound screens on cell-based assays conducted in multi-plate format. It is widely used in academic and commercial laboratories. We are continually improving its functionality and user-friendliness.

ANALYSIS OF CHIP-CHIP DATA USING BIOCONDUCTOR

Joern Toedling

In response to requests from the community and based on our experience with the tools of the Bioconductor project and with ChIP-chip analysis, we have written a comprehensive tutorial that explains in detail how to perform an end-to-end analysis of ChIP-chip data using Bioconductor. ChIP-chip – chromatin immunoprecipitation combined with DNA microarrays – is a widely used assay for DNA–protein binding and chromatin plasticity. The interpretation of ChIP-chip data poses two major computational challenges: primary statistical analysis and integrative bioinformatic analysis. In the tutorial, we present how the freely available software systems R and Bioconductor can be used for the analysis of a ChIP-chip experiment for histone modifications, starting from reading the raw scanner output data and culminating in ways to relate identified ChIP-enriched regions to annotated genome features and other experimental results. The tutorial consists of a detailed explanation of the analysis steps and the R source code used for each step. In addition, the example dataset is provided as a Bioconductor package. This setup allows the reader to easily reproduce the complete analysis. We presented the tutorial in the form of an article in the Education section of *PLoS Computational Biology* and as a lab exercise at the CSAMA course in Brixen, Italy, in June 2008 (Toedling & Huber, 2008).

QUALITY METRICS

Audrey Kauffmann with Alvis Brazma, Misha Kapushesky and Helen Parkinson, Microarray Informatics group

The objective of this project is to develop and disseminate quality metrics and tools for determining data quality. The Bioconductor package, arrayQualityMetrics, proposes quality metrics for assessment of individual array quality, homogeneity, signal to noise ratio, and it conducts outlier detection. Removing outlier arrays from the dataset before performing the analysis reduces the noise, and can increase the statistical power and lead to a more accurate biological understanding of the studied system.

As a second part of the project, we are developing a new Bioconductor package, named ArrayExpress, that converts ArrayExpress TAB-MAGE data into a Bioconductor object. By allowing such automated and easy access for the Bioconductor data analysis software packages to the vast amount of datasets provided by ArrayExpress, we aim at facilitating further large-scale and systematic analysis of public data and meta-analyses. We are assessing and evaluating data quality, using the arrayQualityMetrics package, on every dataset of the database to identify the major factors affecting data quality.

TOOLS FOR THE ANALYSIS OF NEXT-GENERATION SEQUENCING DATA

Simon Anders, with Martin Morgan and Robert Gentleman at FHCRC, Seattle and Julien Gagneur and Lars Steinmetz, EMBL Heidelberg

Next-generation sequencing technologies, such as Solexa and 454, are revolutionising the approaches taken to almost every genomics experiment. They are used for studies of DNA–protein binding (ChIP), chromatin plasticity, allele- and strand-specific transcription, and sequence and copy number variation in individual genomes, etc. Tools for data analysis, however, are currently still poorly developed and many aspects, including quality assessment, read mapping and assembly, and quantification are currently not well understood. We are contributing to an effort to leverage the analytical and statistical capabilities of the R and Bioconductor environments for this task (Figure 4).

BIOCONDUCTOR

with the Bioconductor core developers, see http://www.bioconductor.org

Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data. The project was started in autumn 2001 with members from various US and international institutions.

Publication of a new computational method is incomplete and of limited use to others without a software implementation whose source code is available for review, criticism and improvement. Most data analysis and modelling problems require the use of many tools, which need to be able to work together in a flexible way that can be easily configured by the user. One system that offers this flexibility is provided by the programming language and statistical computing environment R and, more

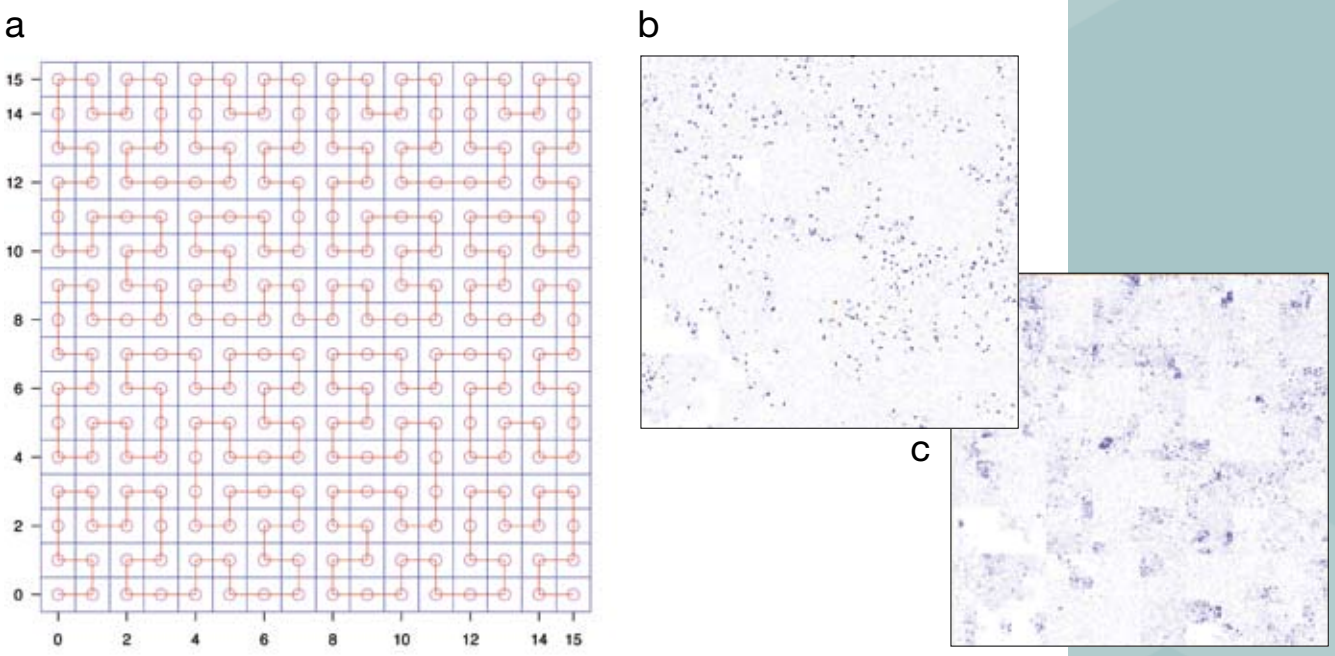


Figure 4. *(a) Fourth-order approximation of the Hilbert curve, a continuous, fractal, space-filling curve in the unit square. (b) False colour representation of along-chromosome read mapping density from a ChIP-experiment using an H3K4-monomethylation specific antibody. Each pixel of the 512x512 pixel image corresponds to a region whose size is 1/262,144 of the length of a whole chromosome, and the pixels are connected according to the ninth order approximation of the Hilbert curve. Darker colours correspond to higher read density. (c) Similar to (b) but for H3K4-trimethylation. Panels (b) and (c) show the uneven coverage of the target chromosome and visualise the differences in the sizes of the enriched regions, which tend to be long and contiguous for the monomethylation and short for the trimethylation.*

specifically, by the functional genomics-inspired infrastructure of the Bioconductor software project. It enjoys great popularity in the microarray data analysis field, but is also extending into the analysis and modelling of other large-scale experiments such as protein interactions, mass spectrometry, phenotype screens and genetic interactions.

In 2008, Bioconductor has seen two major releases, and a further expansion of the user base and the number of software and data packages it provides. The number of software packages is now 294 and the user base is in the thousands (exact numbers are not known since there is no registration). In July 2008, we held a two-day conference in Seattle, which was attended by 200 participants. A developer meeting in Lausanne, Switzerland, in April 2008 attracted 35 participants. An internationally attended advanced course was held in Brixen, South Tyrol, in June 2008.

FUTURE PROJECTS AND GOALS

Biology and its applications to human health will continue to be driven by advances in experimental technologies. Of particular interest are next-generation sequencing, genetic screens for phenotypic consequences of DNA sequence and copy number variation, and high-content phenotyping using automated microscopy. To make these advances fruitful for systematic models of biological processes, we aim to stay at the forefront of developments in experimental design, data analysis, statistical software and mathematical modelling. An emphasis lies on project-oriented collaborations with experimenters.

The Le Novère Group: computational systems neurobiology

INTRODUCTION

The Le Novère group's research interests revolve around signal transduction in neurons, ranging from the molecular structure of membrane proteins involved in neurotransmission to modelling signalling pathways. In particular, we focus on the molecular and cellular basis of neuroadaptation in neurons of the basal ganglia. The supra-macromolecular structure of the postsynaptic membrane strongly influences signal transduction (Figure 1). Moreover, the whole structure is dynamic and evolves, for example, under the control of neuronal activity. By building detailed and realistic computational models, we try to understand how neurotransmitter-receptor movement and clustering, interactions between membrane and cytoplasmic proteins, and spatial location 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. As editors of the language, we also work on increasing its coverage, and are developing software to support SBML usage. The Systems Biology Graphical Notation 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), and controlled vocabularies to improve the models (the Systems Biology Ontology, the TErminology for the Description of DYnamics etc.). Finally, a model is only useful if it can be easily accessed and reused. BioModels Database (www.ebi.ac.uk/biomodels/) is now the reference resource where scientists can store, search and retrieve published mathematical models of biological interest.

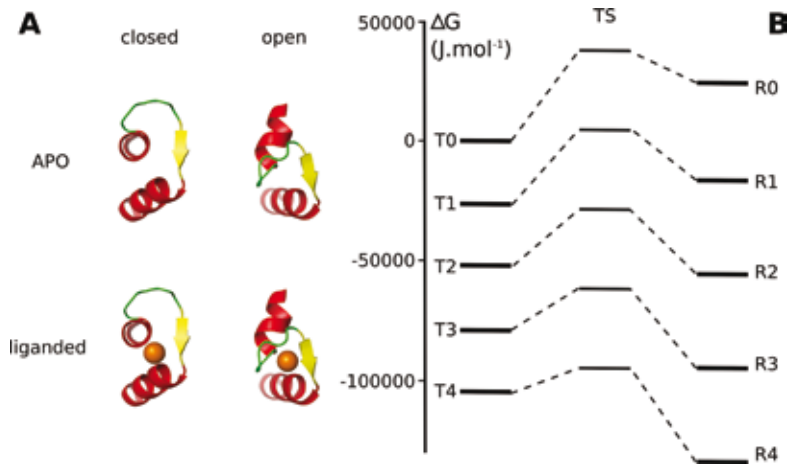


Figure 1. *(A) Representative structures of a calmodulin EF, showing residues 49–75. The closed apo structure and the open liganded structure come from experimental observations. The calcium-liganded structure is inferred from a zinc-bound structure. The open apo structure is a prediction of our model. (B) Summarised free energy diagram for the different states of calmodulin. Each level of energy represents the average of all the forms carrying the same number of calcium ions. Free energy differences between T state and corresponding R state relate to the allosteric isomerisation constant. Between corresponding T and R states, a hypothetical transition state is depicted based on estimates of rate constants. Closed T state is shown on the left, open R state on the right, and the transition state in the middle.*

Nicolas Le Novère

PhD. 1998, Pasteur Institute, Paris.
Postdoctoral research at the University of Cambridge.
Research fellow, CNRS, Paris.
At EMBL-EBI since 2003.

Group Members

Visiting Scientist
Stuart Edelstein*

Postdoctoral Fellow
Noriko Hiroi*

Software Engineers
Chen Li
Camille Laibe
Daniel McGreal*
Nicolas Rodriguez
(shared among several research groups)

Scientific Database Curators
Vijayalakshmi Chelliah*
Lukas Endler*
Nick Juty*

PhD Students
Lu Li
Michele Mattioni
Melanie Stefan
Dominic Tölle

Trainees
Ranjita Dutta-Roy*
Kedar Nath Natarajan*
Duncan Berenguier*

Marie Curie Students
Koray Dagan Kaya*
Dagmar Koehn*

**Indicates part of the year only*

COMPUTATIONAL SYSTEMS BIOLOGY OF DENDRITIC SPINE SIGNALLING

The glutamatergic synapse is one of the main cellular components of the mammal 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 multimolecular 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 in 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. 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; Figure 1). 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, 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, Michele Mattioni

The position and movement of neurotransmitter receptors in and around synapses influences neuronal signal processing. Moreover, it has long been know that LTP is, in part at least, due to the appearance of new neurotransmitter receptors at the postsynaptic site. The source of the newly acquired receptors is still not fully known, nor is the mechanism by which the receptors eventually end up in the synapse. 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. The model of the dendritic spine includes receptors diffusing within the membrane, scaffold molecules within the synaptic specialisation capable of binding receptors and a molecular picket-fence surrounding the postsynaptic density. 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. Additionally, confinement of receptors to a micro-domain at the synapse and the release location of receptors within the spine affect the time course of incorporation. We are now developing a spatial model of the entire medium-spiny neuron that will assist us in understanding signal integration and plasticity of many dendritic spines. These models incorporate kinetic descriptions of signalling pathways in each compartment and electrical behaviour of synapses, spines and neuron.

Publications

2007

Laibe, C. & Le Novère, N. (2007). MIRIAM Resources: Tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.*, 1, Article 58

2008

Le Novère, N. (2008). Neurologic diseases: are systems approaches the way forward? *Pharmacopsychiatry*, 41, S28-S31

Le Novère, N., *et al*. (2008). DARPP-32: Molecular integration of phosphorylation potential. *Cell. Mol. Life Sci.*, 65, 2125-2127

Le Novère, N., *et al*. (2008). Systems Biology Graphical Notation: Process Diagram Level 1. *Nature* Preceedings DOI:10101/npre.2008.2320.1

Shimizu, T.S. & Le Novère, N. (2008). Looking inside the box: Bacterial transistor arrays. *Mol. Microbiol.*, 69, 5-9

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

Taylor, C.F., *et al*. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nat. Biotechnol*., 26, 889-896

Other EMBL publications

Fernandez, E., *et al*. (2006). DARPP-32 is a robust integrator of dopamine and glutamate signals. *PLoS Comput. Biol*., 2, e176

Le Novère, N., *et al*. (2005). Minimum Information Requested In the Annotation of biochemical Models (MIRIAM). *Nat. Biotechnol*., 23, 1509-1515

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. Their importance is illustrated by the conditions that arise when they fail to function properly, such as in Huntington's disease, schizophrenia and drug addiction. The main inputs to these neurons come from cortical glutamatergic terminals. Dopamine modulates this transmission, providing a measure of the internal (hedonic) state. In the mammalian brain, a protein phosphatase inhibitor, DARPP-32, has been identified as a major target for both dopamine and glutamate signalling. We have developed a detailed quantitative model of the regulation of DARPP-32 phosphorylation and dephosphorylation by both signals (Fernandez *et al.*, 2006, Le Novère *et al.*, 2008). Dynamic simulations show that the function of DARPP-32 depends on the delay between the two signals, and therefore the protein not only measures the intensity, but also the coincidence between signals. This model has now been extended to accurately account for the variety 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. A forthcoming version of the model will encompass the transduction of neurotrophin signals and the subsequent intracellular cascades, in particular, the signalling triggered by activation of TrkB receptor and mediated by PLCgamma, which have been indicated as important for associative learning.

COMPUTATIONAL SYSTEMS BIOLOGY

The practise of systems biology relies on interfaces; interfaces between the entities we study (a paradigm shift from a physical object-centric view toward a relationship-centric one), interfaces between tools (several combined tools are used from the retrieval of primary data to the detailed analysis of a model's behaviour), and more importantly interfaces between individuals (to build a non-trivial mechanistic model requires researchers to merge existing work and gather external expertise). 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 tools and parts kits. 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 the 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 for the community because they are insufficiently characterised to allow them to be reused. With today's increased interest in detailed biochemical models, it was necessary to define a minimum quality standard for the encoding of those models. The Minimal Information Requested in the Annotation of Models (MIRIAM) is a set of rules for curating quantitative models of biological systems (Le Novère *et al.*, 2005). 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 and composition into large subcellular models. An important part of the standard consists in controlled annotation of model components, based on Uniform Resource Identifiers (URIs). The MIRIAM database is an online resource created to enable interoperability of this annotation (Laibe & Le Novère, 2007; www.ebi.ac.uk/miriam/). It is a catalogue of data resources, whether controlled vocabularies or primary data resources, and provides the means to generate and resolve MIRIAM URIs. The use of MIRIAM annotations by the community is now growing, and software has been developed that uses those URIs as a glue to merge models and to integrate other datasets. MIRIAM's guidelines deal mostly with the structure of the models. However, in order to run simulations and obtain numerical results using these models, one needs additional information. The Minimum Information About a Simulation Experiment (MIASE) is a fledging effort to agree upon a set of mandatory information to add to relevant publications. Both MIRIAM and MIASE are part of MIBBI, a more general effort to coordinate the development of reporting guidelines (Taylor *et al.*, 2008).

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. SBO is currently made up of five different vocabularies: 1) a controlled vocabulary for parameter roles in quantitative models (e.g. 'forward unimolecular rate constant', 'Michaelis constant'); 2) a taxonomy of the roles of reaction participants (e.g. 'catalyst', 'competitive inhibitor'); 3) a list of modelling frameworks that specify how to interpret a mathematical expression (such as 'continuous' or 'discrete'); 4) a classification of mathematical expressions used in biochemical modelling (e.g. 'mass action kinetic', 'Henri-Michaelis-Menten kinetics'); and 5) a branch containing the classification of events represented by biochemical models ('binding', 'transport'). 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 below), or semantically enriched computing formats to represent biochemical knowledge such as BioPAX. To complete SBO, which is meant 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 for instance (see below), and an ontology to characterise numerical descriptions of dynamic behaviours (TEDDY; http://www.ebi.ac.uk/compneur-srv/teddy).

Formal languages to encode models and simulations

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 the BioModels Database. While bringing minor corrections and clarifications to the current specification of the language (Level 2, Version 3), 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; www.ebi.ac.uk/compneur-srv/sed-ml/). 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 process flow diagrams 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 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.sbgn.org), 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 Diagrams, the Entity Relationship Diagrams and the Activity Flow Diagrams. 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. The first level of SBGN Process Diagram

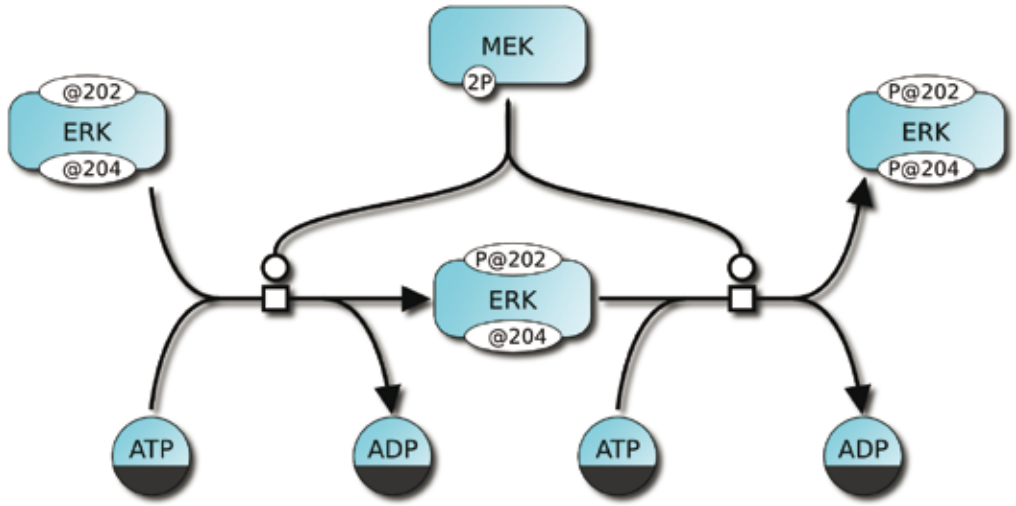


Figure 2. *An example of SBGN in action, showing the catalysis of ERK phosphorylation by MEK in the growth factors signalling pathway.*

has been publicly released (Figure 2; Le Novère *et al.*, 2008). Software support for SBGN Process Diagram was developed concurrently with its specification in order to speed up public adoption. 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 Rodiguez, Vijayalakshmi Chelliah

For computational modelling to become more widely used in biological research, modellers must be able to exchange and share their results. Simple model repositories already exist, but one of the persistent requests from the user community has been for a fully-fledged model database. 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. 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 was developed in collaboration with the California Institute of Technology and is now the largest database of curated models worldwide (containing more than 250 models and 27,000 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 recently released a new version of the database, with a faster and improved interface, offering additional services such as online simulations.

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 the development of standards and resources for systems biology will continue, with the goal of completing the puzzle of representations and ontologies so as to efficiently integrate the different levels of description of biochemical and cellular processes, qualitative, quantitative and experimental.

The Luscombe Group: genome-scale analysis of regulatory systems

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.

Group Members

Staff Scientists
Annabel Todd
Juanma Vaquerizas

PhD Students
Florence Cavalli
Aswin Sai Narain Seshasayee
Judith Zaugg

Visitors
Jonathan Landry
Inigo Martincorena
Jordan Peccia
Juri Reimand

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.

Much of our work so far has focused on the regulatory system in the yeast *Saccharomyces cerevisiae*. By integrating diverse data sources – from genome sequence to the results of functional genomics experiments – we can study the regulatory system at a whole-organism level (Figure 1). Since the start of the group in 2005, we have also expanded our interests to understanding regulation in enterobacteria and humans. Below we describe some of our findings in these new areas as well as our continued work in yeast.

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;

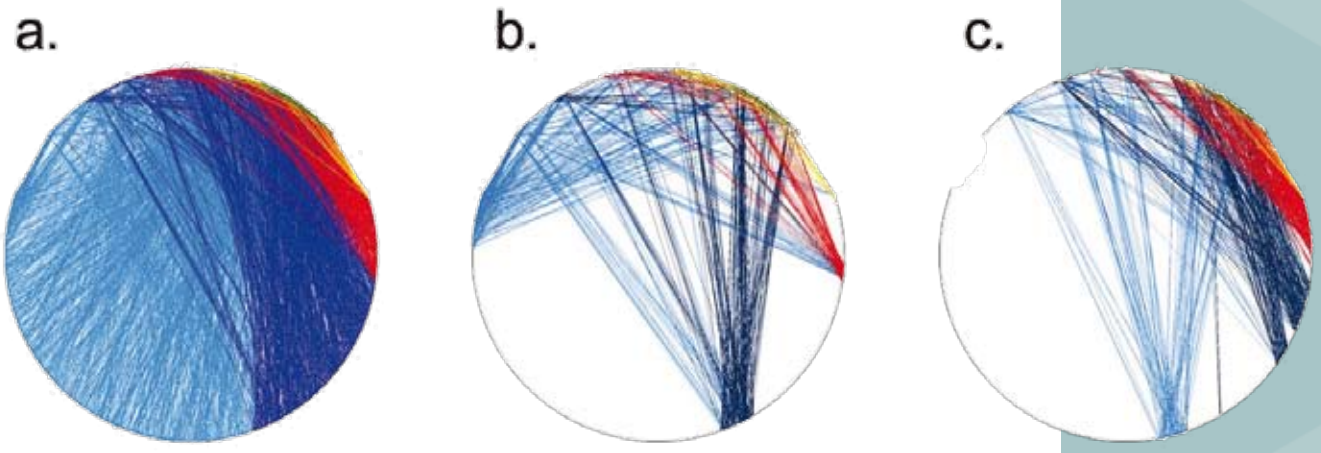


Figure 1. *A network representation displays the usage of the yeast regulatory system during (a) the cell cycle and (b) stress response. Distinct regions of the network are clearly used and this is accompanied by fundamental changes in the underlying network structure.*

Publications

2008

Hancock, V., *et al*. (2008). Transcriptomics and adaptive genomics of the asymptomatic bacteriuria Escherichia coli strain 83972. *Mol. Genet. Genomics*, 1-12

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

Pearson, M.M., *et al*. (2008). The Complete Genome Sequence of Uropathogenic Proteus mirabilis, a Master of Both Adherence and Motility. *J. Bacteriol.*, June:190(11), 4027-4037

Reimand, J., *et al*. (2008). GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res*., 36, W452-459

Seshasayee, A.S., *et al*. (2008). Principles of transcriptional regulation and evolution of the metabolic system in E. coli. *Genome Res*., Epub ahead of print

Sivaraman, K., *et al*. (2008). Codon choice in genes depends on flanking sequence information - Implications for theoretical reverse translation. *Nucleic Acids Res*., 36, Article e16

Other EMBL publications

Luscombe, N.M., *et al*. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.*, 1, reviews001

Other publications

Barabasi, A.L. & Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5, 101-113

Beer, M.A. & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, 117, 185-198

- 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 2009 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, Nicholas Luscombe in collaboration with Gillian Fraser, Department of Pathology, University of Cambridge

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 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.

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 2). 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.

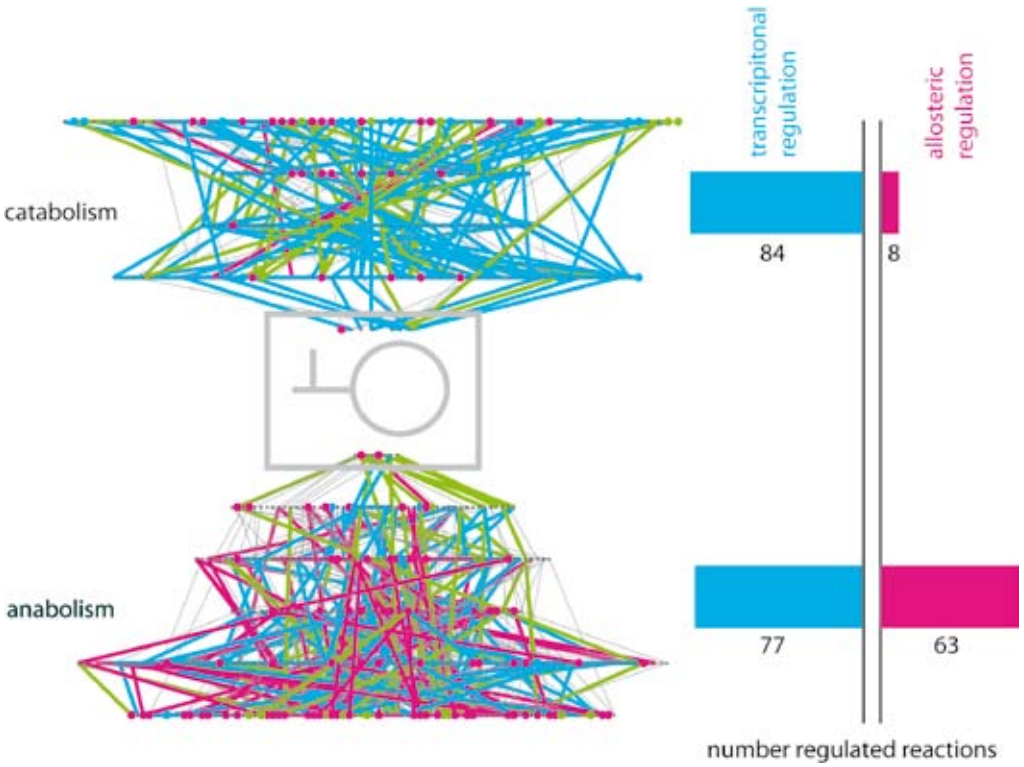


Figure 2. *A network representation displays the* E. coli *metabolic system. Nodes represent small molecules and edges depict enzymatic reactions. The reactions are coloured according to whether they are controlled transcriptionally (blue), allosterically (cyan) or by both methods (green). Allosteric feedback predominantly regulates anabolic pathways, whereas transcriptional feedback controls both anabolic and catabolic pathways.*

Regulation of complex and infectious bacterial behaviour

Our studies of bacterial regulation form part of a broader collaboration with Gillian Fraser's group 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 (Figure 3a; 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 (Figure 3b). 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.

Boiani, M. & Scholer, H.R. (2005). Regulatory networks in embryo-derived pluripotent stem cells. *Nat. Rev. Mol. Cell Biol.*, 6, 872-884

Borneman, A.R., *et al*. (2007). Divergence of transcription factor binding sites across related yeast species. *Science*, 317, 815-819

Chua, G., *et al*. (2004). Transcriptional networks: reverse-engineering gene regulation on a global scale. *Curr. Opin. Microbiol*., 7, 638-646

Collins, S.R., *et al.* (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell Proteomics*, 6, 439-450

Durbin, R., *et al*. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. *Cambridge University Press*

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

Harbison, C.T., *et al*. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431, 99-104

Jansen, R., *et al.* (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* 302, 449-453

Jensen, L.J., *et al*. (2006). Co-evolution of transcriptional and post-translational cell cycle regulation. *Nature*, 443, 594-597

Keseler, I.M., *et al.* (2005). EcoCyc: a comprehensive database resouve for Escherichia coli. *Nucleic Acids Res*., 33, D334-337

Kschischang, F.R., *et al*. (2001). Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47, 498-519

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

Odom, D.T., *et al*. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet*., 39, 730-732

Palin, K., Taipale, J. & Ukkonen, E. (2006). Locating potential enhanver 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

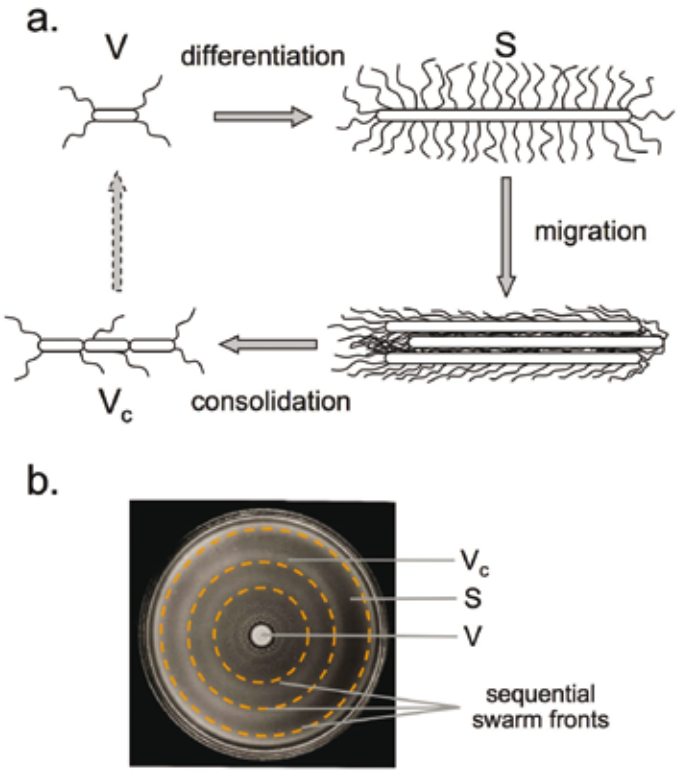


Figure 3. *The* Proteus mirabilis *swarm cycle (a) schematic and (b) photographic image. There is periodic cycling between vegetative (V) and swarming cell types (S); outward bacterial migration results in a colony with a bull's eye appearance.*

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.

REGULATORY NETWORKS IN YEAST

Jonathan Landry, Juri Reimand, Annabel Todd, Nicholas Luscombe in collaboration with Jacky Hess and Nick Goldman

A graphical model-based approach for transcription regulation

Discrete biological function can rarely be attributed to an individual molecule. Instead, most biological behaviours arise from complex interactions between a cell's components including DNA, RNA, proteins and small molecules. High-throughput techniques (such as microarrays) allow routine interrogation of the status of a cell's components at a given time. In turn, new technology platforms (such as protein chips and tandem-affinity mass spectrometry screens) permit us to determine how these molecules interact with each other (Collins *et al*., 2007).

The yeast *Saccharomyces cerevisiae* has extensive experimental data describing the transcription regulators, binding sites and target genes. We have collected large quantities of publicly available genome-scale datasets including: ~200 DNA binding transcription factors identified by sequence homology searches; over 14,000 regulatory interactions between transcription factors and target genes reported from the literature and ChIP-chip experiments (Harbison *et al*., 2004); and gene expression data for more than 70 published microarray studies covering many cellular conditions (Chua *et al*., 2004).

Numerous recent studies have demonstrated the potential of analysing molecular interactions from a network perspective. Furthermore, many features of molecular biological networks (such as the scale-free topology) are shared across disparate systems (Barabasi & Oltvai, 2004). Thus networks provide a unifying language, allowing us to adapt analytical methods from diverse fields for molecular biological investigation, including transcription regulation (Figures 1 and 2).

Most current network approaches reduce biological interactions to simplistic binary links between nodes, despite the fact that these interactions represent complex relationships (such as physical contacts or small molecule transfer) depending on the molecules involved. We continue to develop graphical models techniques (GMs) to quantitatively analyse the robustness of these networks. GM methods were developed concurrently in several fields of mathematics and computer science (Kschischang *et al*., 2001) including artificial intelligence, signal processing and digital communications. GMs are simple and flexible, allowing modelling of complex systems. They are also powerful tools for network inference and learning, underpinned by their probabilistic nature.

In bioinformatics, GM methods (such as Hidden Markov Models) are now routinely used for biological sequence analysis (Durbin *et al*., 1998), and they have also recently been successfully introduced to network analysis (Jansen *et al*., 2003; Beer & Tavazoie, 2004). Similar approaches will enable us to integrate noisy datasets, go beyond correlations to determine causal and quantitative relationships, and predict outcomes that can be tested experimentally.

A graphical models approach for data integration

The collective analysis of many types of biological information is prerequisite to investigating systems-level behaviour; however, integrating these data remains a challenge. Experimental data are derived from a variety of methods that record diverse information to differing levels of accuracy. Here, GMs provide the ability to progress beyond the simple aggregation of biological data; using a combination of symbolic and numerous data types, GMs provide a powerful framework for integrating heterogeneous, noisy and incomplete information from disparate sources.

We are now comparing datasets against each other (such as ChIP-chip with binding site information) to assign quantitative weights reflecting their reliability and agreement. Within the GM framework, these weights will be represented as probabilities quantifying confidence levels for interactions between biological entities. This method for data integration will provide a solid foundation for future work, as we will be able to assess our confidence in the conclusions we draw. In effect, the implementation of robust data integration methods from the outset increases the reliability of the predictions drawn from these data.

Combinatorial regulation by transcription factors

One of the major differences between prokaryotic and eukaryotic organisms is the extent to which combinatorial regulation is utilised. By combining two, or often many more, transcription factors, and by switching regulatory partners depending on circumstances, eukaryotes achieve complex and intricate control of target gene expression. Furthermore, it allows organisms to encode fewer transcription factor genes; thus 6.2% (280 of 4,485) of *E. coli* genes are DNA binding factors, whereas in yeast the proportion falls to 3.1% (~180 of 5,794).

Though many complementary models for combinatorial regulation have been suggested, so far there has not been a comprehensive analysis of a 'code' for transcriptional regulation on a genomic scale. Most recently, Harbison and colleagues proposed several possible models, including the use of single regulators, multiple regulators and repetitive upstream DNA binding motifs. We are currently analysing these regulatory modules on a genomic scale.

Evolution of the regulatory system

In order to use *S. cerevisiae* as a model for eukaryotic regulatory systems we must assess the extent to which the same network is maintained in other organisms. By comparing across organisms, we are also able to assess how the evolution of regulatory components drives speciation (Jensen, 2006; Tirosh, 2006).

So far, cross-species comparisons of transcription regulation have mainly focused on DNA sequence analysis – searching for conserved transcription factor binding sites in multiple genomes using phylogenetic footprinting. More recently, ChIP-chip experiments have been performed in multiple species, demonstrating that transcription factor binding evolves rapidly even between closely related organisms (Borneman *et al*., 2007; Odom *et al*., 2007). In collaboration with Jacky Hess and Nick Goldman, we are currently examining the evolution of the transcription factor repertoire across numerous yeast genomes.

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

Tirosh, I., *et al*. (2006). A genetic signature of interspecies variations in gene expression. *Nat. Genet*., 38, 830-834

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

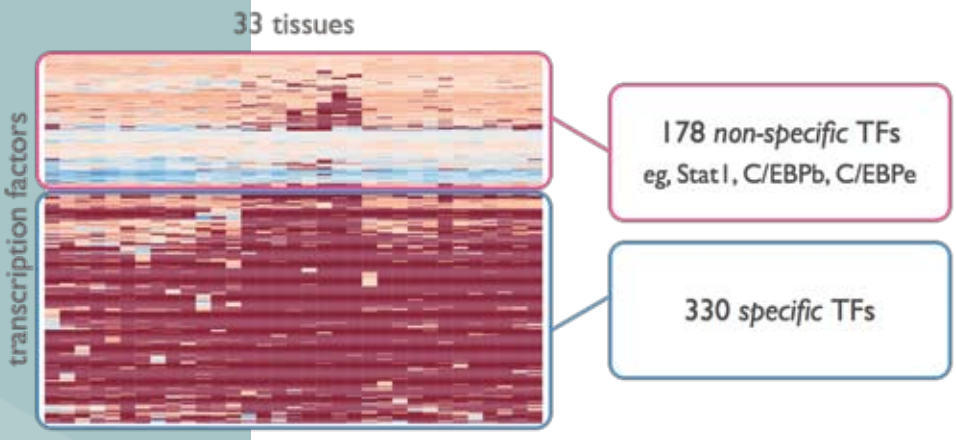


Figure 4. *A heatmap displaying the pattern of expression of human transcription factors. The regulators are either specifically expressed in a small number of tissues or ubiquitously expressed across all tissues.*

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 4). 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 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 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, Nicholas Luscombe in collaboration with Jop Kind, Ritsuko Suyama, Asifa Akhtar, EMBL Heidelberg

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 all together: males have an X0 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 up-regulate 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.

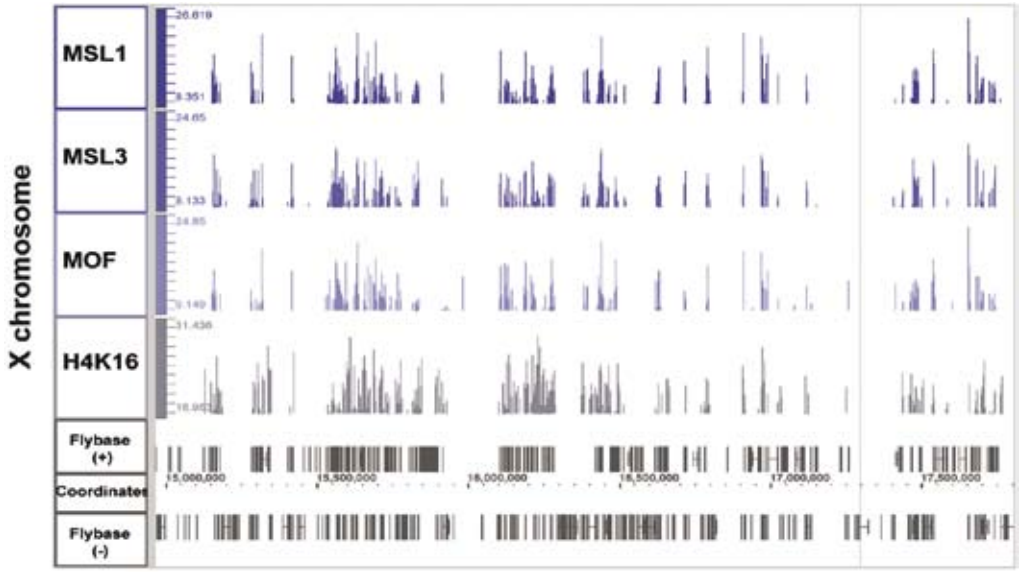


Figure 5. *Genomic mapping of X-chromosome binding by the members of the Dosage Compensation Complex (MSL1, MSL3, MOF) and the effect on histone H4 acetylation (H4K16).*

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 (Figure 5). 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.

The Rebholz-Schuhmann Group: facts from the literature and biomedical semantics

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.

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. Over the past two years, the team has generated several public resources: a lexicon of biomedical terms, an ontology for gene regulatory events and a semantic retrieval engine (MedEvi).

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

The BioLexicon: a terminological resource for information extraction

Vivian Lee, Jung-Jae Kim, Piotr Pezik, Anika Oellrich with project partner Sophia Ananaidou, University of Manchester

To provide full coverage of domain knowledge in molecular biology, the Rebholz-Schuhmann group has undertaken research to generate a complete terminological resource 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 the BioLexicon, for example, the BioThesaurus (Liu *et al.*, 2006), to cope with nonsense names and identify ambiguous terms. Finally, 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.

Identification of chemical named entities in patent texts

Tiago Grego, 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.

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) and OSCAR3 (Corbett, Batchelor & Teufel, 2007), and a machine-learning (ML) classifier for NCE recognition was trained and tested (using Conditional Random Fields). This classifier has shown that the gold standard corpus 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).

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.

Group Members

Staff Scientists
Vivian Lee
Piotr Pezik
Antonio Jimeno Yepes

Postdoctoral Fellows
Alexander Griekspoor*
Jung-Jae Kim

Software Engineer
Anika Oellrich

PhD Students
Adam Bernard*
Sylvain Gaudan*
Kevin Nagel

Visitors
Tiago Grego
Dolf Trieschnigg
Andra Waagmeester
Pinar Yildirim

Visiting Students
Alexandre Baillif
Christoph Grabmüller
Silvestras Kavaliauskas
Alejandro Pironti

**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.

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.

Our analysis compares 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 have been assessed with regards to their capability 1) to reproduce the original manual MeSH annotations and 2) to generate meaningful annotations that complement the manual annotations (verfied by a curator). KNN showed the best performance in all tests.

Furthermore, 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.

Annotation of protein residues based on a literature analysis: cross-validation against UniProtKB

Kevin Nagel, Antonio Jimeno Yepes, Tom Oldfield

The Universal Protein Resource (UniProt) is a valuable resource for the validation and interpretation of 3D structure patterns in proteins that have been extracted from a systematic analysis of the Protein Data Bank (PDB) content. Previous research work was concerned with the identification of point mutations from the biomedical literature in order to support curators in the time-consuming work of manual database curation. However, these methods were restricted to point mutations and did not identify features for the annotation of proteins at the residue level.

This ongoing research work introduces a system that identifies protein residue sites in the text of Medline abstracts and annotates them with contextual features. The performance of all text mining modules was evaluated against a manually-annotated corpus. The identified annotation features can be attributed to at least one of six targeted categories, e.g. enzymatic reaction. Extracted results were cross-validated against UniProtKB and for 13 annotations of residues that have not been confirmed in UniProtKB, a manual assessment was performed. Altogether, the proposed solution delivers annotations for proteins that have been identified in the scientific literature.

Extending the carotenoid pathway with terms from text mining

Andra Waagmeester, Piotr Pezik

The carotenoid metabolism is relevant to the prevention of various diseases, e.g. cardiovascular diseases, cancer and eye diseases. Although the main actors in this metabolic pathway are known, our understanding of the pathway is still incomplete. This analysis analysed a large amount of scientific documents to enrich the pathway with novel and relevant concepts. The proposed text mining workflow has been validated on the vitamin A metabolism pathway, which is a well-studied part of carotenoid metabolism (Network of Excellence Nutrigenomics, NuGO). The workflow uses an initial set of publications cited in a review paper (1,191 references), enlarges this corpus with Medline abstracts (13,579 documents) and then extracts the most relevant terminology from the total number of publications. The results have been validated by domain experts.

With our approach we were able to enrich the vitamin A pathway with 37 new and relevant concepts from a total of approximately 89,086 terms. 13 new concepts were added to the pathway and another 14 concepts were identified. The newly identified concepts belong to different pathways and thus could form relevant links between carotenoid metabolism and other pathways. The remaining ten concepts (from a total of 37) were classified as biomarkers or nutrients.

Confirmation of gene–disease associations with GO annotations from the literature

Christoph Grabmüller, Alejandro Pironti, Antonio Jimeno Yepes

Disease relevant genes are identified by association studies and remain hypothetical until the molecular mechanisms of the association have been verified experimentally (Butte et al., 2006). At the same time, evidence for modifications of cellular mechanisms and their relevance to diseases and genes is abundant in the scientific literature, i.e. in Medline abstracts. Therefore, we have processed the complete Medline database to associate genes to diseases based on their concept vectors of GO terms in Medline abstracts. MeSH concept profiles served as a complementary approach. All identified gene–disease associations were assessed against the Mendelian in Man (MIM) database and the genetic association database (GAD).

For the diseases we identified amongst the top-ranked genes using MeSH concept profiles, 26 associations were contained in GAD but not confirmed by MIM and therefore require experimental verification. GO concept profiles showed improved performance over MeSH concept profiles, but the best results for the extraction task were 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 MIM. 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). Altogether, GO concept profiles are a promising approach for associating genes to diseases based on experimental findings described in the scientific literature.

Ontological support for information extraction

Vivian Lee, Jung-Jae Kim, Piotr Pezik, Anika Oellrich

For the identification of gene regulatory events from text, members of the research group have composed the Gene Regulation Ontology (GRO), in collaboration with Elena Beisswanger, Jena University. It covers processes linked to the regulation of gene expression as well as the physical entities involved in these processes (such as genes and transcription factors). GRO has a particular focus on the relationships between processes and the molecules (participants) involved and as a consequence, the taxonomic backbone has been further enriched by additional semantic relation types ('part-of', 'from-species', 'participates-in' with the two sub-relations 'agent-of' and 'patient-of').

The GRO classes have been manually defined and cross-referenced to other ontologies. The definitions and cross-references make use of existing biological knowledge in the sense that terms from the following ontological resources have been integrated: Molecule Role Ontology (IMR), Sequence Ontology (SO), Gene Ontology (GO), ChEBI Ontology (ChEBI) and the Event Ontology (IEV). In addition, gene regulation-specific resources (e.g. TransFac) have been considered as well as protein domain databases (e.g. InterPro, Pfam, Panther), textbooks and online biomedical dictionaries. The GRO has been submitted to the Open Biomedical Ontologies (OBO) library.

FURTHER DEVELOPMENT OF THE IT INFRASTRUCTURE FOR INFORMATION EXTRACTION

Collaborative annotation of a large-scale corpus

Alexandre Baillif, Antonio Jimeno Yepes

Ongoing work is concerned with the harmonisation of annotations in the scientific literature by proposing a universal schema for annotations ('IeXML', Rebholz-Schuhmann *et al.*, 2006). This is an important step to enable semantic enrichment of scientific literature using text mining components from different research groups. In collaboration with two project partners, this annotation standard has been used to compare and align sets of annotated documents to identify diverging annotations provided from the different teams.

A novel IT infrastructure for the automatic submission and evaluation of annotated documents has been established (internship of Alexandre Baillif). This solution will be used to support a collaborative approach to the annotation of a large-scale corpus. Participants will receive a corpus in an open collaboration with the biomedical text mining community. The task is the annotation of named entities in a large biomedical corpus for a variety of semantic categories. The project will produce a large, collaboratively annotated corpus, linked to details of biomedical entities.

Publications

2008

Beisswanger, E., *et al*. (2008). Gene Regulation Ontology (GRO): Design principles and use cases. In *Stud. Health Technol. Inform.*, 9-14

Gaudan, S., *et al*. (2008). Combining evidence, specificity, and proximity towards the normalization of gene ontology terms in text. *Eurasip Journal on Bioinformatics and Systems Biology*, 2008, Article No 342746

Jaeger, S., *et al*. (2008). Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics*, 9, Article S2

Jimeno, A., *et al*. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9, Article S3

Kim, J.J. & Rebholz-Schuhmann, D. (2008). Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. *Brief. Bioinform.*, 9, 452-465

Kim, J.J., *et al*. (2008). MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline. *Bioinformatics*, 24, 1410-1412

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

Spasic, I., *et al*. (2008). Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC Bioinformatics*, 9, Article S5

Other EMBL publications
Rebholz-Schuhmann,D., Kirsch,H. & Nenadic,G. (2006). IeXML: 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

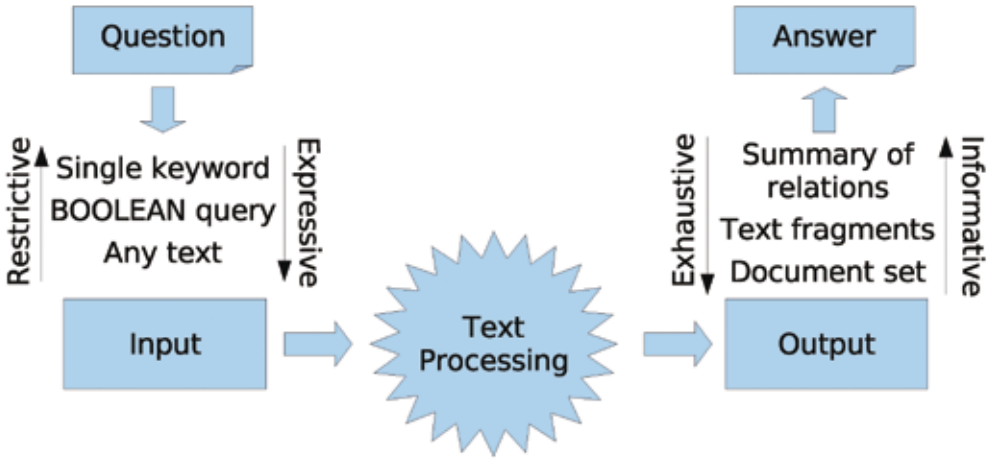


Figure 1. *Overview on the categorisation of information retrieval tools on the basis of their input and output formats.*

Categorisation of services for seeking information in biomedical literature

Jung-jae Kim

Biomedical researchers have to efficiently explore the scientific literature in the area of their research. This goal can only be achieved if the tools for accessing the literature meet the researchers' retrieval needs and if they understand how to use these tools to filter the perpetually increasing number of documents available. We have examined existing web-based services for information retrieval in order to give users guidance to improve their everyday practice of literature analysis. We propose two dimensions along which the services may be categorised: 1) categories of input and output formats, and 2) categories of behavioural usage (see Figure 1). The categorisation would be helpful for biologists to understand the differences in the input and output formats and the tasks they fulfil in information-retrieval activities. Also, they may inspire future bioinformaticians to contribute to innovative developments in this field.

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. We will develop this by focusing on automatic means to measure and evaluate existing options 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 scientific literature. Disambiguation of semantic types requires special solutions.

The Thornton group: computational biology of proteins – structure, function and evolution

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.

Group Members

Staff Scientists
Dan Andrews
Matthew Bashton
Angelo Favia
Nicholas Furnham
Gemma Holliday
Roman Laskowski
Rafael Najmanovich
Marialuisa Pellegrini Calace
Syed Asad Rahman
Gabrielle Reeves*
Eugene Schuster
David Talavera
James Watson*
Daniela Wieser

PhD Students
Julia Fischer
Fabian Gerick
Abdullah Kahraman
Nicola Kerrison
James Torrance

Personal Assistant to the Director
Helen Barker-Dobson
Assistant
Stacy Schab

Visitors
Claudia Andreini
Lorenzo Baldacci
Barbara Brodsky
Gabriele Cavallaro
Franz Fenninger
Noriko Hiroi
Stephanie Juniat
Oleg Lenive
Tim Maiwald

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 2008 our major contributions have been in the following five areas:

- enzyme structure and function;
- using structural data to predict protein function and to annotate genomes;
- evolutionary studies of genes, their expression and control;
- functional genomics analysis of ageing;
- development of tools and web resources.

ENZYME STRUCTURE AND FUNCTION

Julia Fischer, Gemma Holliday, Asad Rahman

In our attempts to understand how enzymes work, we have studied the chemistry of catalysis (Holliday *et al.*, 2007) and coenzymes. We have analysed the statistics of chemical mechanisms and amino acid residue functions that occur in enzyme reaction sequences using the MACiE database of 202 distinct enzyme reaction mechanisms as a knowledge base. We show that the most catalytic amino acid residues are histidine, cysteine and aspartate, which are also the residues whose side-chains are more likely to serve as reactants, and that have the greatest versatility of function. We show that electrophilic reactions in enzymes are very rare, and the majority of enzyme reactions rely upon nucleophilic and general acid/base chemistry. However, although rare, radical (homolytic) reactions are much more common than electrophilic reactions. Thus, the majority of amino acid residues perform stabilisation roles (as spectators) or proton shuttling roles (as reactants). The findings provide a better understanding of the mechanisms of enzyme catalysis and may act as an initial step in the validation and prediction of mechanism in an enzyme active site.

This year we have looked in more detail at which side chains perform which functions within the mechanism and also started an in-depth study of metals and coenzymes. In addition we have devoted considerable effort to developing new cheminformatics tools to compare small molecules and to handle reactions. This work is vital for developing a better process to store and compare mechanisms (Rahman *et al.*, private communication).

In summer 2008, we again updated the Catalytic Site Atlas, with the assistance of four undergraduate summer students. In total they contributed annotations for a further 100 enzymes.

USING STRUCTURAL DATA TO PREDICT PROTEIN FUNCTION AND TO ANNOTATE GENOMES

Roman Laskowski, Marialuisa Pellegrini, Gabrielle Reeves, Matthew Bashton, Rafael Najmanovich, James Watson, Abdullah Kahraman, Angelo Favia, Nicholas Furnham, 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.

Current computational methods for the prediction of function from structure are restricted to the detection of similarities and subsequent transfer of functional annotation. In a significant minority of cases, global sequence or structural (fold) similarities do not provide clues about protein function. In these cases, one alternative is to detect local binding site similarities. We have explored whether, within a dataset of non-homologous proteins, it is possible to discriminate those that bind similar

Ben McLeod
Ferdinando Spagnolo
William Pearson
Judith Reeks
Mauno Vihinen
Kazuto Yamazaki

** Indicates part of the year only*

ligands based on their binding site similarities. To address this challenge, we have implemented a graph matching-based method for the detection of 3D atomic similarities introducing some simplifications that allow us to extend its applicability to the analysis of large all-atom binding site models (Najmanovich *et al*., 2008). This method, called IsoCleft, does not require atoms to be connected either in sequence or space.

When we apply the method to a cognate-ligand bound dataset of non-homologous proteins we find that it is possible to discriminate between different ligand-binding sites to some extent. Discrimination ability decreases with decreasing knowledge about the identity of the ligand-interacting binding site atoms. The decrease is quite abrupt when considering size and chemical composition alone, but much slower when including geometry. We also observed that certain ligands are easier to discriminate.

As an alternative approach, protein–ligand docking has also been investigated as a tool for protein function identification. We have used docking to identify the substrates of a single protein family with remarkable substrate diversity, the short-chain dehydrogenases/reductases (Figure 1). We have examined different protocols for identifying candidate substrates for 27 short-chain dehydrogenase/reductase proteins of known catalytic function. 900 metabolites from the human metabolome were docked to each of these proteins together with their known cognate substrates and products (Favia *et al*., 2008). We investigated the ability of docking to (a) reproduce a viable binding mode for the substrate and (b) to rank the substrate highly amongst the dataset of other metabolites. We compared two different docking methods and two alternative scoring functions for one of the docking methods. In summary we found that the cognate ligand occurred in the 'top' 7% of all ligands for three quarters of the tests. This was encouraging and further work is in progress to improve this approach further and to apply it to proteins with unknown ligands.

In collaboration with Professor Christine Orengo at University College London, we are studying 'Functional Families for Functional and Comparative Genomics'. At the EBI we (Nicholas Furnham) are developing approaches to automatically explore how large families evolve to bind multiple different ligands, combining the phylogenetic analyses with analysis of the chemical structure of the enzyme and its ligands.

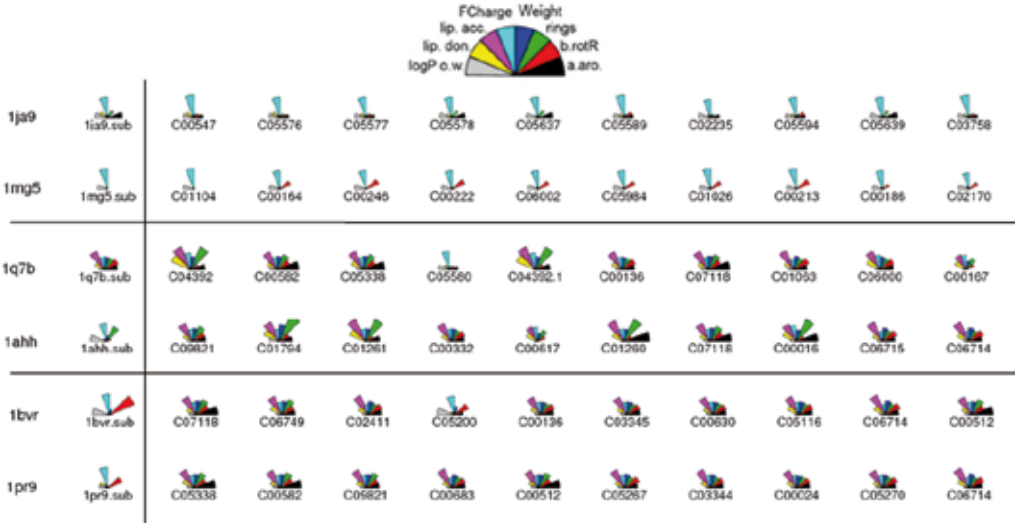


Figure 1. *Docking results for dehydrogenase/reductase enzymes. Physical chemical characterisation of the top ten hits from docking approximately 1,000 human metabolites to six members of the short chain dehydrogenase/reductase family of enzymes. The plots show eight 1D descriptors as colours, described in the hemisphere at the top of the figure, where the size of the sector reflects the value of the descriptor. These desriptors are: LogP, # H-bond donors; # H-bond acceptors, Molecular Weight, Charge, ~ rings, # rotatable bonds,~aromatic atoms. The same descriptors are also shown for the 'known' cognate substrate (first column) for comparison. The plot highlights that in the first two rows all ten top hits are similar and resemble the substrate. The middle two examples show hits which are all different to each other and different from the substrate. However these two enzymes are known to be promiscuous. In the bottom two examples, all the hits look alike but are different from the known substrate. This is probably due to the inaccuracies of the scoring functions, and these results improve if the energy is recalculated with more sophisticated energy functions.*

Publications

2007

Cuff, M.E., *et al*. (2007). Crystal structure of an acetyltransferase protein from Vibrio cholerae strain N16961. *Proteins: Structure, Function and Genetics*, 69, 422-427

Freeman, T.C., *et al*. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.*, 3, 2032-2042

Holliday, G.L., *et al*. (2007). The Chemistry of Protein Catalysis. *J. Mol. Biol.*, 372, 1261-1277

Holliday, G.L., *et al*. (2007). Evolution of enzymes and pathways for the biosynthesis of cofactors. *Nat. Prod. Rep.*, 24, 972-987

Kahraman, A., *et al*. (2007). Variation of geometrical and physicochemical properties in protein binding pockets and their ligands. *BMC Bioinformatics*, 8, S1

Laskowski, R.A. & Thornton, J.M. (2007). Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet*., 9, 141-151

Ward, J.J. & Thornton, J.M. (2007). Evolutionary models for formation of network motifs and modularity in the Saccharomyces transcription factor network. *PLoS Comput. Biol.*, 3, 1993-2002

2008

Bashton, M., *et al*. (2008). PROCOGNATE: A cognate ligand domain mapping for enzymes. *Nucleic Acids Res*., 36, D618-622

As part of the Midwest Center for Structural Genomics (MCSG) we have helped in the analysis of a newly determined protein structure for an acetyltransferase protein from *Vibrio cholerae* strain N16961 (Cuff *et al*., 2007). As part of the BioSapiens Consortium, we have been analysing cancer mutations (Talavera, private communication), participated in writing a book and are developing a new ontology for protein sequence annotation (Reeves, private communication). A review was also written on understanding the molecular machinery of genetics through 3D structures (Laskowski & Thornton, 2008).

EVOLUTIONARY STUDIES OF GENES, THEIR EXPRESSION AND CONTROL

Shiri Frielich

A function-driven approach to the analysis of metabolism has been developed which takes into account the phylogenetic origin of biochemical reactions to reveal subtle lineage-specific metabolic innovations, undetectable by more traditional methods based on sequence comparison (Frielich *et al*., 2008). The origins of reactions and thus entire pathways are inferred using a simple taxonomic classification scheme that describes the evolutionary course of events towards the lineage of interest. We have investigated the evolutionary history of the human metabolic network extracted from a metabolic database, and constructed the taxonomic categories representing eukaryotes, metazoa and vertebrates. This demonstrated that lineage-specific innovations correspond to reactions and pathways associated with key phenotypic changes during evolution, such as the emergence of cellular organelles in eukaryotes, cell adhesion cascades in metazoa and the biosynthesis of complex cell-specific biomolecules in vertebrates. This phylogenetic view of metabolic networks puts gene innovations within an evolutionary context, demonstrating how the emergence of a phenotype in a lineage provides a platform for the development of specialised traits.

FUNCTIONAL GENOMICS ANALYSIS OF AGING

Dan Andrews, Nicola Kerrison, Eugene Schuster, Daniela Wieser

In collaboration with the 'Functional Genomics of Ageing' Consortium at University College London, we have analysed transcriptome data to provide evidence for lifespan extension and delayed age-related biomarkers in insulin receptor substrate (IRS) 1 null mice (Selman *et al*., 2008). Recent evidence suggests that alterations in insulin/insulin-like growth factor 1 (IGF1) signalling (IIS) can increase mammalian life span. For example, in several mouse mutants, impairment of the growth hormone IGF1 axis increases life span and also insulin sensitivity. However, the intracellular signalling route to altered mammalian aging remains unclear. We therefore measured the life span of mice lacking either IRS 1 or 2, the major intracellular effectors of the IIS receptors. Our provisional results indicate that female Irs1-/- mice are long-lived. Furthermore, they displayed resistance to a range of age-sensitive markers of aging including skin, bone, immune, and motor dysfunction. These improvements in health were seen despite mild, lifelong insulin resistance. Thus, enhanced insulin sensitivity is not a prerequisite for IIS mutant longevity. Irs1-/- female mice also displayed normal anterior pituitary function, distinguishing them from long-lived somatotrophic axis mutants. In contrast, Irs2-/- mice were short-lived, whereas Irs1+/- and Irs2+/- mice of both sexes showed normal life spans. Our results therefore suggest that IRS1 signalling is an evolutionarily conserved pathway regulating mammalian life span and may be a point of intervention for therapies with the potential to delay age-related processes.

DEVELOPMENT OF TOOLS AND WEB RESOURCES

Matthew Bashton

We have developed a public data resource, PROCOGNATE, which is a database of protein cognate ligands for the domains in enzyme structures as described by CATH, SCOP and Pfam (Bashton *et al*., 2008). It can be accessed at http://www.ebi.ac.uk/thornton-srv/databases/procognate/.

FUTURE PROJECTS AND GOALS

We will continue our work on understanding more about enzymes and their mechanisms, including a study of how the enzymes, their families and their pathways have evolved. We will develop new computational tools to improve the handling of mechanisms and their reactions, which will allow improved chemistry queries across our databases. We are looking more closely at drug–protein interactions, membrane proteins (Pellegrini-Callace in collaboration with Professor David Jones at UCL) and allosteric effects. In the ageing project we are interested in tissue specificity and using human public transcriptome datasets to explore effects related to human variation and age.

Eswaran, J., *et al*. (2008). Structure of the Human Protein Kinase MPSK1 Reveals an Atypical Activation Loop Architecture. *Structure*, 16, 115-124

Favia, A.D., *et al*. (2008). Molecular Docking for Substrate Identification: The Short-Chain Dehydrogenases/ Reductases. *J. Mol. Biol.*, 375, 855-874

Freilich, S., *et al*. (2008). Metabolic innovations towards the human lineage. *BMC Evol. Biol.*, 8, 247

Kahraman, A. & Thornton, J. (2008). Methods to Characterize the Structure of Enzyme Binding Sites. In *Computational Structural Biology - Methods and Applications* (Schwede, T., & Peitsch, M.C., eds), 189-221, World Scientific Publishing Co

Laskowski, R.A. & Thornton, J.M. (2008). Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet*., 9, 141-151

Najmanovich, R., *et al*. (2008). Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 24, i105-111

Patel, P., *et al*. (2008). Solution Structure of the Inner DysF Domain of Myoferlin and Implications for Limb Girdle Muscular Dystrophy Type 2B. *J. Mol. Biol.*, 379, 981-990

Pietras, Z., *et al*. (2008). Structure and mechanism of drug efflux machinery in Gram negative bacteria. *Current Drug Targets*, 9, 719-728

Selman, C., *et al*. (2008). Evidence for lifespan extension and delayed age-related biomarkers in insulin receptor substrate 1 null mice. *FASEB J.*, 22, 807-818

Torrance, J.W., *et al*. (2008). Evolution of binding sites for zinc and calcium ions playing structural roles. *Proteins: Structure, Function and Genetics*, 71, 813-830

Index

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.