Towards European standards for training and education in Systems Biology

Report of the ERASysBio Summer School
Data Management for Systems Biology
28th-31st July 2008, Adeje, Tenerife, Spain

Background

The partners of the initiative 'Towards a European Research Area for Systems Biology' (ERASysBio) are committed to encourage the adoption of data standards, data management and sharing best practices across the European Research Area. This summer school is designed as a first step to take forward this action and is part of a series of training activities to be implemented over the next two years.

Aims

- to achieve effective communication between all participants;
- to raise awareness of the urgent need to adopt data standards, data management and sharing best practices in the ERA;
- to provide knowledge on the current tools, methods, practices, models, challenges and solutions in the management of data in systems biology;
- to equip participants with tools for decision-making.

Location

The summer school took place on 28th-31st July 2008 at the University of La Laguna Summer Campus, situated in Adeje, on the south-west coast of the island of Tenerife, only 10km away from Reina Sofia International Airport. The University of La Laguna is the oldest university in the Canary Islands, with its origin dating back to 1701 when Augustine monks established a centre for higher learning in the city of La Laguna. In 1992 the summer campus was founded in Adeje to promote culture, research, politics and philosophy among other disciplines, on the island.

Organisers and Sponsors

The summer school was organised by the Biotechnology and Biological Sciences Research Council (BBSRC). BBSRC, the ERA-NET ERASysBio and the Borough Council of Adeje co-funded this event.

Participants

Twenty-six postdoctoral researchers from academic institutions in ERASysBio partner countries attended the summer school, following a few last minute cancellations. The majority of participants were experimentalists currently working in a systems biology project. A small group (6) of bioinformaticians, software modellers and physicists also attended. The countries represented in the summer school were Austria, Germany, Netherlands, Italy, Slovenia, Spain and United Kingdom.

The following experts contributed to the summer school:

Session 1 – Introduction to Data Generation, Standardisation and Storage

Dr. Isabel Rojas, EML-Research, Villa Bosch, Germany. Isabel leads the Scientific Databases and Visualization groups since its foundation in 1999. She has over a decade of experience in the development of databases and related interfaces. She has worked as a lecturer in several computer science disciplines, such as Discrete Mathematics for Computer Science, Probabilities and Statistics, Performance analysis, Introduction to Databases, several programming courses, as well as in the area of databases for bioinformatics. She also has had industrial experience acting as a as database development consultant for a period of over 4 years, managing a group of developers and training personnel in multiple companies. Isabel gave a keynote lecture - 'Data management in a systems biology project' - via Skype using the SysMO Data Management Strategy as a case study.

Dr. Olga Krebs, EML-Research, Villa Bosch, Germany.
Olga is working in the Scientific Databases and Visualization group led by Isabel Rojas. She is one of the developers of SabioRK - System for the Analysis of Biochemical Pathways - Reaction Kinetics (http://sabio.villa-bosch.de/SABIORK/). Olga has a PhD in molecular genetics, 17 years professional experience in the "wet" lab and 8 years in the area of design and development of databases for bioinformatics. Olga ran the hands-on exercise on data modeling and integration.

Dr. Stefan Richter, EML-Research, Villa Bosch, Germany.
Stefan is working in the Molecular and Cellular Modelling Group led by Rebecca Wade. He is implementing workflows and software tools for usage on the www. Stefan has a PhD in biophysics and more than 10 years of professional experience in bioinformatics at biotech companies. Stefan ran the hands-on exercise on Standard Operation Procedures.

Dr. Sven Sahle, University of Heidelberg, Germany. Sven is a junior group leader at the Department for Modelling of Biological Processes. He's one of the lead software engineers of COPASI (www.copasi.org), a software tool for modelling of biochemical reaction networks. His interests include the dynamic properties of biochemical models as well as simulation and analysis methods for those models. Sven is also SBML (Systems Biology Markup Language) editor, elected by the SBML community. Sven gave an update on SBML as a keynote lecture within session 1 and ran the hands-on exercise on modelling and SBML.

Chris Lawerenz, German Cancer Research Center in Heidelberg, Germany. Chris and Juergen Eils have been leaders of the Data Management Group at the Department of Theoretical Bioinformatics for five years. Their research field is the development of databases and IT concepts in the field of translational research and systems biology. Chris is member of the national quality management consortia Telematic Platform for Medical Networks (TMF) and National Genome Research Network (NGFN). He has 15 years experience in project management and data management in industry. Chris ran the hands-on exercise on high throughput database iCHIP with Juergen Eils.

Session 2 – Introduction to Data Exchange and Data Sharing

Dr. Katy Wolstencroft, School of Computer Science, University of Manchester, United Kingdom. Katy is a researcher on the myGrid project. She runs the myGrid outreach programme, providing regular training courses and consultancy for workflow design. Her background is in bioinformatics, bio-ontologies and her research interests involve workflows, e-science and ontologies, and the application of cutting-edge computer science technologies to biological problems. Katy gave a keynote lecture on data exchange and data sharing within this session and ran the hands-on exercises with Franck Tannoh (see below).

Dr. Franck Tannoh, School of Computer Science, University of Manchester, United Kingdom. Franck is a member of the myGrid team and is responsible for the curation and annotation of web services used by Taverna. He provides regular training courses on Taverna. He has a background in bioinformatics and is currently involved in the creation of a registry of annotated biological web services. Franck ran the hands-on exercises within this session with Katy Wolstencroft.

Session 3 – Final Session

Prof. Alfonso Valencia, Centro Nacional de Investigaciones Oncológicas (National Cancer Research Centre), Madrid, Spain. Alfonso is Director of the Department of Bioinformatics and Computational Biology and Director of the Spanish Bioinformatics Institute. Alfonso gave an overview on Bioinformatics, Computational Biology and Data Management for Systems Biology as a closing keynote lecture.

Pre-summer school questionnaire

Participants were asked to complete a questionnaire pre-summer school. The questionnaire included the questions regarding the type of data and the way is generated and stored, the access to data by internal/external users, resources to linking data and data linking computing models, generation and storage of models.

The questionnaire served as a survey and provided trainers with the background of the group participating in the summer school. It also prompted participants to prepare for the summer school.

The programme

The programme included three main sessions and is at **Annex 1**.

The event was initiated with the welcoming remarks by Julio Barbas, as ERASysBio Spanish partner, from the Ministry of Science and Innovation in Madrid. This was followed by a welcome from Gracia Rodríguez, pro-chancellor at University of La Laguna in Tenerife.

A networking exercise, led by Gabriela Pastori from the Biotechnology and Biological Sciences Research Council in the UK, introduced participants in a structured environment. Participants interacted in a speed-dating exercise comprising brief introductions and rapid discussion.

Session 1 – Introduction to Data Generation, Standardisation and Storage

This session was led by Isabel Rojas, Olga Krebs and Stefan Richter in collaboration with Sven Sahle and Chris Lawerenz.

Following to Isabel's keynote lecture, the exercises were introduced to participants in three 15 min presentations: 1) Standard Operation Procedures, by Stefan Richter; 2) Data Modelling and Integration, by Olga Krebs and Chris Lawerenz; and 3) Modelling and SBML, by Sven Sahle. Participants divided in three breakout groups joining the topic of preference and had the opportunity to join another group during the second part of the hands on exercises.

Session 2 - Introduction to data exchange and data sharing

This session was led by Katy Wolstencroft and Franck Tanoh. Following to Katy's lecture, the group divided in two and the exercises were run in parallel. Participants received training in the exchange and sharing of data using Taverna workflows and myExperiment in 15 successive exercises. There was an opportunity for participants to use their own data and to discuss issues and problems that had arisen throughout the day or in their work.

Session 3 - Final Session & Feedback

This session was led by Gabriela Pastori, Olga Krebs and Katy Wolstencroft. The Final Session provided participants an opportunity to sit back and reflect on what they learnt and also to think of the future. Participants were presented with a few questions for them to answer in breakout groups. There were 5 breakout groups of 6 participants each. Participants answered questions individually, using the breakout group for discussion. The questions were:

1) Look at the questionnaire results that you submitted to us before attending the summer school. Is there anything that you are going to do differently as a result of this training?

o To implement **Standard Operation Procedures (SOPs)**: record experiment design in a way that allows the scientific community to use it and benefit from it. Standard protocols, all members within a group to agree on procedures. Common templates and catalogue of methods in the lab.
o To optimise communication between groups to agree formats.
o To consider and implement **annotation** formats and standards.
o To implement **SMBL** and to **standardise data storage** in groups.
o To try to **adapt SBML** kinetic models to human models.
o To increasingly use **workflows** and encourage wet scientists to learn and use workflows.
o To establish a **WIKI**.
o To **share data**.
o To implement **Taverna** wherever suitable; for certain tasks may not be possible.

2) Are you planning to implement any changes in the way you work? If yes, what barriers will you have to overcome to implement them?

- o **Cultural barriers**: people to be convinced, there is a culture barrier in the way we work. Requires open mind to look around for data and people doing similar experiments and using similar tools and methods.
- o **Facilities**: WIKI, internet, IT resources, centralised storage.
- o **Collaboration barriers**, data ownership: shared knowledge of techniques, tools, formats rather than results.
- o **Time and productivity**: implementing and maintaining these changes is time-consuming and may affect, at least temporarily, productivity.
- o **Distance between experimentalists and modellers**: the gap between these two different languages must be overcome. Factors that could have a positive effect: 1) co-location; 2) listening to users´ demands and preferences, talking rather than learning a new discipline; 3) working together at least once - job shadowing; 4) motivation to cross-talk; 5) leadership.

3) Please draw an action plan for the management of your data in the next 6 months.

- o To implement **SOPs** and to create and manage studies in **iCHIP**. Go back to data, convince own group of the need.
- o To use the vast resource of **tools available** through Google
- o To use **WIKI** to collate all info.
- o To use **Laboratory Information Manuals** (LIMs) to record experimental procedures and protocols. Reproducibility and reliability of results.
- o To improve **communication** between modellers and experimentalists, to take time to learn/interact.

4) Are there any wider issues that you would like to comment on and/or be included in future summer schools? Would you like to see this type of training more readily available, e.g. e-learning platforms?

What can be included

- o Additional examples of the use of data management tools and comparison to other systems. Overviews on databases available for proteomics, metabolomics, etc. More examples on use of Taverna. Issues in managing experimental data, public repositories, success stories.
- o Extra day to do more hands on, or parallel sessions, or using own data on 3rd day, more lectures
- o Follow up the progress of participants. Face to face is important but also use web-seminars, and other ways to continue networking - WIKI.
- o Advanced courses (next steps).
- o Data modelling for biologists and viceversa.

Closing keynote lecture

The summer school ended with a keynote lecture by Alfonso Valencia on Bioinformatics, Computational Biology and data management for systems biology, and closing remarks by Gabriela Pastori, who thanked everyone involved in the organisation and sponsoring of this event.

Have we achieved our aims?

To achieve effective communication between all participants

The group worked well through the networking and hands-on exercises and the final session. Participants were open to share their experiences in data management. In doing this, they identified areas of common interest as well as common barriers and challenges. This created a suitable environment for participants to communicate effectively between them and with the experts. As a result, participants expressed their wish to continue interacting and remain in contact through e.g. a network. ERASysBio offered the possibility to set up a Wiki in its website to include all documents and presentations related to the summer school and any additional material that the experts consider appropriate. The SysBio Data Management Wiki would serve as a communication/consultation tool and be expanded to others as the summer school series develops.

To raise awareness of the urgent need to adopt data standards, data management and sharing best practices in the ERA

In the Final Session, participants concluded that the summer school had been successful in raising awareness. The specific areas where the influence of the training on participants' views and future plans can be observed are:

o   All participants expressed clearly their awareness of the new working philosophy underpinning systems biology with regards to communication, data standards and data sharing.
o   Participants were keen to promote effective communication within their team and/or within their consortium, and to find vehicles to maximise communication between experimentalists and modellers.
o   Participants unanimously expressed their commitment to implement and promote Standard Operation Procedures (SOPs) in their current projects.
o   Participants highlighted the importance of recording and storing data appropriately, e.g. LIMS (Laboratory Information Managing System) in establishing SOPs.
o   Participants expressed their will to use SBML and other existing community standards for data annotation.
o   Many participants were keen to implement the use of workflows in their projects using Taverna wherever possible.

To provide knowledge on the current tools, methods, practices, models, challenges and solutions in the management of data in systems biology

The sessions were designed with the aim to provide an introduction on the tools and methods available. The first keynote lecture on day 1 by Isabel Rojas provided participants with an overview of the range of systems and tools available on the internet for data management in systems biology, all with free access. As defined by Isabel, the main messages participants were encouraged to consider and embrace were:

- systems biology projects involve groups with a wide spectrum of resources, and data management and integration expertise;
- it is crucial that the users' views are incorporated early in the implementation of any data management solution and that the users are alerted of the benefits of such solution;
- it is recommended that project participants use (partial) solutions already in place;
- it is crucial that data annotation is performed at all levels to facilitate exchange, comparison and integration;
- it is essential that project participants use defined exchange standards and controlled vocabularies;
- it is important to carefully consider ownership and versioning issues in the storage of data, models and processes.

Within session 2, participants were equipped with a step-by-step guidance to exchange and share their data using Taverna workflows and myExperiment. The keynote lecture by Katy Wolstencroft provided an overview of the range of systems and tools available on the internet for data exchange and sharing, all with free access, and examples of experimentalists applying these tools in their projects. Participants learnt about the range of applications of such workflows in 15 practical exercises.

To equip participants with tools for decision-making

The introductory Summer School in Data Management for Systems Biology has given participants the sufficient tools to make some basic decisions on:

- what information is available and with free access, and where to find it;
- which parameters to consider in setting up the data management of a new systems biology project or the measures to take for improving an existing systems biology project;
- who to contact.

BBSRC, August 2008