Standards for the study of Metabolism: MetaboMeeting @ EBI, March 7th 2005

Chris Taylor acting chair and Susanna-Assunta Sansone recording sec'y.
The agenda and the list of attendees are provided at the end of this document.

Thanks to the EBI Industry Programme for generously underwriting the costs of this meeting (www.ebi.ac.uk/industry) and Liz Ford (EBI) for the logistics.

SCOPE AND INTRODUCTION

The overall goals of this meeting were twofold: Firstly, to bring together existing groups approaching the standardization of metabolic investigations from different angles; secondly, to foster interaction with standardization efforts in other 'omics' domains, optimizing synergy and avoiding duplications.

The wider functional genomics context for this work cannot be ignored; this 'multi-omics' approach to biology will significantly affect the structure and content of ontologies, file formats and reporting structures. Core biological descriptors need to be shared, as do descriptors relating to the design of investigations, sample generation and treatments. This requires extensive liaisons between communities that previously were only loosely connected, and will, as stated, heavily impact the structure and content of the relevant omics-based standards (being) developed by those communities. All those standards should stand alone, but should also function together to support functional genomics investigations.

Chris Taylor scoped the meeting, describing a roadmap involving the HUPO Proteomics Standardization Initiative (PSI, psidev.sourceforge.net) and the Microarray Gene Expression Data Society (MGED, www.mged.org), highlighting the activities of interest to the metabolic community. Chris leads the General Proteomics Standards working group within PSI, and is involved in the development of a format and a supporting ontology for investigations employing MS technology, which has relevance for the metabolic domain. The Reporting Structure for Biological investigation (RSBI, www.mged.org/Workgroups/rsbi) is a working group within MGED, moderated by Susanna Sansone. It acts a single point of focus for toxicogenomics, nutrigenomics and environmental genomics and is working with PSI towards a reporting structure and ontology for describing investigations that encompasses different technological domains.

PRESENTATIONS

Nigel Hardy, University of Aberystwyt. ArMet

Nigel presented the history and status of ArMet ('Architecture for Metabolomics', www.armet.org). The work started under the FSA G02006 project, to develop an internet-accessible database containing full information on the metabolic profiles (metabolomes) of Arabidopsis thaliana and Solanum tuberosum. A prototype database was built first the concepts abstracted from the data model and the ArtMet architecture was developed.

ArMet is a framework for the description of plant metabolomics experiments and their results. Two future projects are planning to deploy ArtMet-based systems; MeT-RO (Metabolomics at Rothamsted, www.metabolomics.bbsrc.ac.uk/MeT-RO.htm) and HiMet (users.aber.ac.uk/nwh/Research/Projects/himet.html). The initial FSA project prototype contained many data items, but most of these were optional. A balance was struck and the minimum requirements were identified. ArMet is a component based-structure and it comprises of a core set of data items and 9 components each of which describes part of the process of a plant metabolomics experiment. This component-based approach enables extensions, so that the complete architecture to be customised. Compliance to ArMet implies support for the core data and dependencies for these components. (1) Admin component: user responsible; (2) Biological Sources: genotype (provenance, species etc) and source; (3) Growth: treatment and environment; (4) Collection: harvesting; (5) Sample Handling: bulking, division, grinding, shipping; (6) Sample Preparation: preparation methods; (7) Analysis Specific Sample Preparation; (8) Instrumental Analysis: run on the analytical instrument, link to the raw datasets (NO RAW DATA is stored), to be expanded when multiple inttrsruments are used; (9) Metabolome Estimate: fingerprinting, metabolomics, metabolite profiling and targeted analysis.

ArMet does not have an ontology, but some data fields have a complex types, requiring a value and an authority; the authority identifies the ontology or the controlled vocabulary (CV) from which the value was taken. ArMet aims to be descriptive, not prescriptive! ArMet can be implemented as LIMS, RDMS, JCAMP etc; it promotes sharing of SOPs, experimental design and CVs. It supports data analysis managements and reporting ultimately. ArMet does not deal with data analysis. Currently there are several implementations in Oracle, Postgres and bits in Access. ArMet design has been translated into SQL and XML with the objective of supporting the exchange of data from different implementations. The current preoccupation is broadening acceptance of the model and the discovery possible flaws.

Nigel Hardy, University of Aberystwyth. Other efforts: a summary

MIAMET (Minimum Information About a Metabolomics Experiment) is a checklist of the information necessary to provide context for metabolomics data that is to be published (Bino, R.J. *et al*. (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci. **9**, 418–425). It represents a first positive step in the direction of a standard, but does not provide the complete formal data description of the specific required data items necessary for the development of supportive data handling systems.

The **Metabolomics Society** is a group founded from academia, government and industry, announced in March 2004, which aims to address this issue (metabolomicssociety.org). To date the society has only minimal members on any of its Boards from major industry, although it is a stated aim to improve this discrepancy.

The **Standard Metabolic Reporting Structure** (SMRS, www.smrsgroup.org) is a group started in 2003, aiming to develop and recommend standards for conducting and reporting metabonomics and metabolomics studies. Participants in the SMRS include strong group of practitioners from both industry and academia (mainly in pre-clinical toxicology) with clear working practices and targets. A draft policy document has been produced and posted on the public website for comments. The document is divided into: biology, analytical techniques,

chemometrics reporting terms. It aims to support regulatory submission, database reposition, and reporting data to journals.
Nigel clarified the difference between the minimal requirements, like MIAME and MIAPE, SMRS and MIAMET, and the designs, such as ArtMet, MAGE-OM, PSI-OM and PEDRo. Additional distinctions were made are that ArMet is a project driven initiative developed from the 'bottom up', whereas MIAME and SMRS are 'top down' approaches. Clearly there is some overlap between ArtMet, MAGE and PSI-OM (on experiments, sample descriptions) which needs to be examined through a group activity.

Martin Yuille, MRC Geneservice. MRC infrastructure for post-genomics research: two cases

MRC Geneservice (http://www.hgmp.mrc.ac.uk/geneservice/) is a molecular biological resource centre supported by the Medical Research Council, which manages samples of recombinant DNA (clones) and of genomic DNA from human subjects and which has capacity to type these samples. MY is grant holder in the MRC DNA Banking Network (http://www.dnabank.mrc.ac.uk/), whose remit is to collect large numbers of (nucleic acid) samples from patient cohorts and case-control studies of common diseases of major public health impact, to be managed as national resources and be made widely available.

Martin stressed that reference clone sets are key in this post-genomics era. These are huge in number and widely shared. Standards for handling these samples are required urgently. He brings two examples of common mistakes and possible solutions:

— RNAi clone set: 15k clones, 200 copies and 1 contaminant! The preparation of RNAi samples needs clean rooms and a centralized service.
— Genomic DNA: 1k sample, 2 stock sets: 10% error rate! In this case the samples are taken on site (wherever the donor is), but other procedures need to be in place to avoid mistakes like mislabelling of tubes, or the over-dilution of working stocks.

Infrastructures for both curation and sharing are required. There is a clear need for common standards and secure reference copies, updated annotations, SOPs in place, an efficient distributions and a system ensuring ethical use and confidentiality for samples of human origin. Similarly, this could apply to biological material, such as metabolites.

This year, an application was put forward for a coordination action to the EU by MRC Geneservice, Sanger, EBI and many others, to discuss issues including a possible clone archive . Currently there is a UK network for DNA banking and genotyping, coordinated by MRC, with a central LIMS (referenced above). The LIMS has an inventory of the existing samples from which users can identify and 'book' samples for their needs. A database is proposed that would contain descriptions of studies that are carried out; data (on both genotypes and phenotypes) is stored and can be searched (where permissions allow).

Chris Taylor, PSI and EBI. Overview of PSI activities

The Human Proteome Organisation (HUPO, www.hupo.org) was created to federate national and regional proteomics societies, assist in the coordination of public proteome initiatives, and engage in 'scientific and educational activities'. The HUPO Proteomics Standards Initiative (PSI, psidev.sourceforge.net) has been in existence for some four years now; its remit is to generate data format standards, standard term definitions (as controlled vocabularies / ontologies) and minimum reporting requirements (MIAPE). PSI has built collaborations and cooperative arrangements with public and private 'bench scientists' and informaticians, tech vendors, funders and publishers related to the field.

Biggest win for PSI to date; the Molecular Interaction Format (MIF) from the PSI: MI workgroup. Several PPI databases exist (BIND, DIP, MINT, Hybrigenics, IntAct) but the data they contain is provided in many different formats, and is not synchronised, which means lots of tedious work repeatedly combining data sets. The standard (MIF) defines a minimal data model that allows scientists to provide core data, with back references and importantly simplifies synchronisation between resources (cf. EMBL + GenBank + DDBJ). MIF data can now be viewed in the Cytoscape viewer; a commercial product that draws interaction graphs and allows expression data to be layered on top.

A number of projects are being run under the PSI: GPS (General Proteomics Standards) and PSI: MS (mass spectrometry) banners: MIAPE (Minimum Information About a Proteomics Experiment) is our reporting requirement for proteomics to be enforced by journals, repositories and funders; it is composed of a number of technology-specific modules associated with a parent document. The first of these, for mass spectrometry, is now undergoing expert review. The parent/core document states the underlying principles (sufficient information to be useful for the contextualisation of data and conclusions, but not requiring so much as to be impracticable).

We are also generating a number of XML formats for data exchange. PSI-ML, derived from the PSI Object Model (PSI-OM), is the 'umbrella' format designed for data exchange and submission from any proteomics workflow. This format 'wraps up' a number of modular formats: mzData from the PSI: MS workgroup, which captures the use of and data generated by a mass spectrometer, is in a mature state (stable production version 1.05) and is being implemented by all the major mass spec vendors and a number of other companies (for example, Matrix Science who produce the Mascot software); note that raw data capture is not a primary role for mzData (although it is technically feasible). The mzIdent format captures the use of software tools to assign peptides and proteins to mass spectra; it will help to unify the results from different search engines and will facilitate cross-comparisons. PSI will also soon begin working on GelML — a format built around gel electrophoresis and related technologies.

Lastly, our ontology work: The development of a rich and extensive ontology is crucial for the construction of unambiguously worded data files. GPS will begin with a proteomics-specific resource initially focusing on supporting the mzdata and mzIdent formats. It will ultimately contribute to an ontology for functional genomics under development as a collaborative effort between the MGED Ontology, the MGED RSBI working groups (which deals with high level descriptors for projects, studies, study groups, biomaterials *etc.*).

DISCUSSION

Standards – the carrot and the stick dilemma!

When integrating data, the issue of data sharing is getting more complex and more metadata would be required to interpret such complex multi-technology investigations. The attitude is changing, funding agencies and developers are aware of the payback and willing to invest (see MGED and PSI). Prospective standards for metabolic study should be embraced for the same reasons as in other omes; facilitating comparison, reposition and exchange; enabling the extraction of maximum value from data sets; enabling an assessment of the quality and relevance of a piece of work.

SMRS Group and the proposed document

Hector Keun reported on this topic. This is a policy document particularly developed in the context of submission to regulatory bodies. The group has come together to develop and recommend standards for conducting and reporting metabonomics and metabolomics studies. The document comprises all the aspects crucial to any interpretation and therefore subject to consideration for standardisation and reporting, such as; analytical data acquisition, the data on the biological material (metadata) and the statistical, chemometric or bioinformatic analysis. Time and funding constraints have 'limited' the scope of SMRS group to initiate other activities such as developing ontology and an XML format.

The discussion focused on whether or not SMRS is truly an 'open source' effort. Ultimately the attendees agreed that, even if the first meeting was composed of only a small number of academic collaborators (plus a large number of industrial participants), the second meeting was more open. Susanna Sansone brought to the attention of the attendees that since the last meeting SMRS has developed a public website with an open mailing list. This resolution is intended to open the group and allow others to participate in this initiative. In addition the group has sought inclusion of journal editors and recently an editor from Science and one from Nature Biotechnology have accepted to be on board of the SMRS group.

XML format

The attendees agree that it is hard to standardize around implementations. A draft UML is helpful to get users feedback, but ultimately XML is what we need to develop. The way to go would be to develop and agree on an XML format for data exchange based on a general UML model. The XML format(s) should be flexible, expansible and modular.

Ontology

The attendees agree that this should be developed in a functional genomics context. This should be done in an open way as a collaborative effort with PSI, the MGED RSBI and Ontology working groups and other interested parties.

SMRS - ArMet

What is the level of communality among these? Nigel Hardy reports that ArMet fulfils all the SMRS requirements and vice versa SMRS requirements are almost all in ArMet architecture. Is there any 'conflicts' between ArMET and SMRS? Yes, there is, but it is just the inclusion of data process and experimental design and this is a matter of implementation. There is a need for optimizing and maximizing the collaboration between ArMet and SMRS groups and Nigel Hardy is the right person for this task, being involved in both efforts.

NMR developments

The attendees agree that the developments should be divided amongst the contributing persons/groups. Before initiating any work, existing models will be surveyed to avoid re-inventing the wheel (we suspect though that no relevant models exist). Jules Griffin and group will make a first attempt at an NMR model; Mark Viant, Hector Keun, Chris Taylor and Andrew Nicholls will then assist in refining that working model. All will examine the work of the Proteomics Standards Initiative to reuse materials where possible. Assistance will also be sought from Mike Beale at Rothamsted (lead on the BBSRC-funded MeT-RO project, which in addition to Rothamsted includes Aberystwyth and Manchester); this joint activity will be important in fostering cooperation between SMRS, ArMet and MeT-Ro. Bruker has already output compliant to existing regulatory approved formats and it should be taken on board too.

MS developments
As with NMR, the attendees agree that there is already work in PSI-OM and ArMet; Chris Taylor and Nigel Hardy will make a first survey.

Analysis methods
Do we need to be descriptive or prescriptive in this case? The latter is not yet possible, even if most of the time scientists get this wrong! PSI has opted provided some model for this, but not in response to demand... The attendees agree that we should have a common high-level analysis model, as there are mechanisms that are applied across the board, enabling the standardization of the reporting/description of such analyses.

Project page on sourceforge
For the development of standards, openness and due process must apply. To ensure that all parties have the opportunity to express their views, the attendees agree to create a website at SourceForge named smrsgroup.sourceforge.net, and collate existing/relevant data models, useful use cases and links to relevant initiatives. Community buy-in is vital. Mailing lists will be set up where interested parties can subscribe. To promote acceptance and avoid authorship issues, representatives of the different groups (SMRS, ArMet, etc) will have access to the project page as project developers. They will serve as a 'gatekeepers' and solicit input from relevant contributors.

Follow-up meetings
The attendees agree that a follow-up meeting to coordinating this developmental activity is needed. Rothamsted this summer is a good option, perhaps in combination with the ArMet developers' meeting. MeT-RO and BBSRC to be consulted on this (*addendum by CFT*).

Other relevant meetings this year include; Metabolomics Society meeting in Japan, June 20th-23rd; SMRS + Metabolomics society + FDA meeting in September (date yet to be finalised); CHI Metabolic Profiling meeting in December in Orlando, which garners interest from the US metabolomics companies and East-coast pharma.

Dissemination
Active contacts at upcoming conferences should be made, in addition to reporting the above proposal. Relevant meetings include: the joint SMRS - Metabolomics Society (US, mid summer, TBS), Uni of Cambridge meeting (September, TBA), Japanese metabolomics meeting (Japan, June, TBA).

SUGGESTED ACTIONS

— Jake Pearce will set up the SourceForge site.
— Nigel Hardy will act as a bridge between ArMET and SMRS groups.
— Jules Griffin, Mark Viant and Hector Keun will collaborate on model/formats for NMR.
— Chris Taylor and Nigel Hardy will survey PSI model and formats and ArMet for MS .
— Chris Taylor will approach BBSRC and Rothamsted with a view to holding a July follow up meeting.

— Mike Beale to be approached, both w.r.t. the potential July meeting, and to build bridges with the MeT-RO consortium.

— Susanna Sansone will report the above proposals to the MGED Board and working groups.

— Chris Taylor will report the above proposals to the HUPO-PSI group.

— Hector Keun will report to above proposals to the SMRS group.

Agenda:

10.00 Welcome and coffee in Sanger building.
10.30 ArMet: A database for plant metabolomics experiment data
-- Nigel Hardy, Aberystwyth.
11.30 The need for (wet and dry) standards in biobanking
-- Martin Yuille, MRC Geneservice, Hinxton
12.00 Move across to EBI, first floor landing...
12.30 Lunch (EBI, first floor landing).
13.30 Open discussion (EBI, A2-33).
15.30 Coffee break.
16.30 Summary and distribution of action items.
17.00 Close.

Discussion points for afternoon session (summary of):

What is needed, what purpose would that serve:
Data exchange? Reposition? Submission? Validation? Mining?
Formats -- flexible (requires ontology), expansible, modular?
Ontology -- should allow for functional genomics context

Different requirements:
pharma / medicine / biological science / publishers /
regulators / statisticians / bioinformaticians; FuGEs?

What is desirable from the above, versus what is feasible...

Reuse of existing metabolomics resources --
SMRS, Imperial groups / COMET, ArMet, Kell group,
Griffin group, CEBS-MAGE, ..?

Reuse of relevant non-metabolomics resources --
PSI, MGED (including RSBI)

Meeting at Rothamsted over Summer (BBSRC? SMRS Pharmas?)

SourceForge project as single point of focus for format development --
Mailing lists, XMLs, ontology, (requirements,) not DB schemata
-> smrsgroup.sourceforge.net?

Collaborations --
UK metabolomics and metabonomics
Global equivalent
Non-metabolomics (technology transfer, functional genomics)

Attendees:

Given name	Family name	Affiliation	Email
Theo	Arvanitis	University of Birmingham	t.arvanitis@bham.ac.uk
Conrad	Bessant	Cranfield University	c.bessant@Cranfield.ac.uk
Kevin	Brindle	University of Cambridge	k.m.brindle@bioc.cam.ac.uk
David	Croft	EBI	croft@ebi.ac.uk
Kirill	Degtyarenko	EBI	kirill@ebi.ac.uk
John	Easton	University of Birmingham	jme071@bham.ac.uk
Marcus	Ennis	EBI	mennis@ebi.ac.uk
David	Grainger	Addenbrookes Hospital	djg15@cam.ac.uk
Jules	Griffin	University of Cambridge	jlg40@mole.bio.cam.ac.uk
Nigel	Hardy	University of Aberystwyth	nwh@aber.ac.uk
Henning	Hermjakob	EBI	hhe@ebi.ac.uk
Hector	Keun	Imperial College London	h.keun@imperial.ac.uk
Jyoti	Khadake	EBI	jyoti@ebi.ac.uk
Maria	Krestyaninova	EBI	mariak@ebi.ac.uk
Mahon	Maguire	University of Cambridge	mlm23@hermes.cam.ac.uk
David	Mosedale	Royal Papworth Hospital	dem@mole.bio.cam.ac.uk
Steffen	Neumann	IPB Halle Gemany	sneumann@ipb-halle.de
Andrew	Nicholls	GlaxoSmithkline	andrew.w.nicholls@gsk.com
Jake	Pearce	Imperial College London	jake.pearce@imperial.ac.uk
Philippe	Rocca-Serra	EBI	rocca@ebi.ac.uk
Denis	Rubtsov	University of Cambridge	dvr22@mole.bio.cam.ac.uk
Reza	Salek	University of Cambridge	salek@biochem.ucl.ac.uk [old]
Susanna	Sansone	EBI	sansone@ebi.ac.uk
Ugis	Sarkans	EBI	ugis@ebi.ac.uk
Larissa	Soldatova	University of Aberystwyth	lss@aber.ac.uk
Irena	Spasic	University of Manchester	i.spasic@manchester.ac.uk
Chris	Taylor	EBI	christ@ebi.ac.uk
Imre	Vastrik	EBI	vastrik@ebi.ac.uk
Mark	Viant	University of Birmingham	m.viant@bham.ac.uk
Martin	Yuille	MRC Geneservice, Hinxton	myuille@geneservice.mrc.ac.uk