Minutes of the 3rd Annual Spring Workshop of the HUPO-PSI, held in association with the HUPO Publications Committee

Proteomics and Beyond
April 21-23, 2006
San Francisco

Sponsored by: EMBL Industry Programme, Thermo Electron Corporation, Applied Biosystems, Agilent Technologies, DKFZ German Cancer Researcg Inst., Genologics Life Sciences, Serono, Roche, Syngenta, Novartis, BBSRC.

Photo: Sang Yun Cho

Introduction The meeting was opened by a welcome from the local organisers of the event, Robert Barkovich (Thermo Corporation), David Horn (Agilent Technologies) and Sean Seymour (Applied Biosystems). This was followed by a brief statement of the goals of this workshop by the Chair of HUPO-PSI, Henning Hermjakob (EMBL-EBI) who explained that with increased participation, increased investment into the adoption of standards by both commercial and academic organisations and a growing involvement by scientists, journals and funders, the HUPO-PSI needed to formalise its organisational structure. Discussions to this end would be a part of this workshop in addition to the more traditional activities of pushing forward the existing standards to the point of publication. In addition, with the increasing

interaction between the proteomics community and other closely related efforts, it had been decided to invite other organisations to this meeting, not only to give them an opportunity to meet amongst themselves but also to explore and potentially exploit areas of commonality. Representatives of the Metabolomics Society, MIACA (Minimum Information About a Cellular Assay, FuGE (Functional Genomics Experiment working group) and FuGO (Functional Genomics Ontology) were present and would not only be chairing specific sessions but would be actively involved in other meeting tracks where there was an area of related interest.

A brief update of achievements to date was then made by the session chairs. PSI-MI 2.5 has now been released and several databases are already making data available in this format. A plug-in for Cytoscape to allow import in this format has also been written and will soon to sent to their site for publication. The Minimum Information About a Molecular Interaction Experiment (MIMIx) document has been completed and submitted for publication. During this meeting, in addition to minor updates to both the schema and controlled vocabularies (CVs), the plan was to work on tools to utilise this format, and also to respond to user requests for a simple tabular representation of the data in addition to the more complex XML download.
Randy Julian (Indigo BioBiosystems) then summarised the achievements of the PSI-MS group in developing specifications for the interchange of raw and processed mass spectrometry data. The remit of the group was to create a specification made resistant to obsolescence through the use of controlled vocabularies and support the schema through documentation. mzData 1.05 is now stable and the specification ready for submission to the newly appointed PSI editor prior to publication. Priorities at this meeting would be to focus on improving the CVs, to investigate integration of the standard with the Institute of Systems Biology mzXML and to address missing features, for example the ability to describe directed proteomics using multiple reaction monitoring techniques. Angel Pizarro (Univ. Penn), chair of the Proteomics Informatics (PI) workgroup who would be concentrating their initial efforts on the analysisXML standard for mass spectrometry based informatics, such as protein identification, gave a brief overview of the current status of that project. John Garavelli (EMBL-EBI) outlined the progress made to date in developing a hierarchical CV to express protein modifications (PSI-MOD), followed by Chris Taylor who gave a brief summary of the progress of the Separations and Sample Generation (SP) workgroup which have several reporting guideline documents in the final stages of preparation. The CVs developed as part of this process will also become part of the FuGO, and the FuGE data model will be used to link the models created by this group to more downstream processes in the workflow of an experiment. The activities of the Gel-based methods of analysis (GEL) group was described by Andy Jones (Univ.Manchester) and is in an advanced state, with the minimum reporting guidelines complete and the interchange format currently under test. The controlled vocabularies need further development work. Oliver Fiehn (UC Davis) representing the Metabolomics Standards Initiative (MSI) emphasised their intention to work closely with the HUPO-PSI on the technical standards whilst themselves working on the biological standardization of phenotype. Dawn Fields (NERC) described MIGS (Minimum Information About a Genome

Sequence) which extends the core information that is common to describing genomic sequences beyond that traditionally captured in genome annotations and proposed that a central registry of such reporting requirement documents be established, to avoid redundancy and allow the scientist to access all the documents required to describe an experimental workflow. Finally, Stefan Wiemann (DKFZ) described MIACA, a reporting standard for the description of cell based assay projects.

The attendees then listened to a series of presentations describing the use of proteomics standards in the laboratory workflow. Randall Julian described a comparison of plasma samples taken from both Zucker obese and lean rats to look for biomarkers for the development of diabetes. Proteins were identified and quantified from LC-MS/MS analysis of proteolytic protein digests. Raw data was exported in HUPO-PSI mzData format which then allowed the iteration of the data through multiple analyses. Storing raw data allowed multiple injections of a sample to be compared and peptides present in each to be unambiguously identified and quantified via integration of extracted ion chromatograms. Since no mass spectrometer can be run continuously without a change in sensitivity, the format allowed $\log_2$ peptide peak areas to be quantile normalized and exponentiated to obtain a normalized value on the original peak area. Variation in column retention time can be compensated for by matching charge states across runs and fitting a shift function regression plot to the results. Similarly, analysis of metabolites by GC/MS was subject to the same processes. Standards-compliant repositories will allow access to the raw data required to allow such iterative processing of results.

Jayson Falkner (U. Michigan) described ProteomeCommons.org, a public repository for proteomics data with easy access for multiple users and data integrity checks. By storing the data on a public network, 14 terabytes of on-line storage has been achieved at relatively low cost. There are many ways to both upload and download the data, including a JAVA-based graphical tool, web tools and command line access. Peer-to-peer (P2P) protocols allow multiple computers to share the load and data integrity is assured by the use of a signed hash using SHA and DSA encryption. It is possible to search for spectra or for protein matches and a GUI tool allows conversion between several file formats, including mzData.

Jeff Kowalak (NIH) detailed the work of the Association of Biomolecular Resource Facilities Proteomics Standards Research Group (sPRG) in promoting the development and use of standards in proteomics. A mixture of 49 known proteins from a single species were distributed across a number of laboratories with participants being asked to supply details of methodology with their putative peptide IDs. A total of 500 identifications were made, although 58 extra proteins were identified by two or more peptides and may have been genuine contaminants. No single method was notably better or worse at identifying the protein content and the mixture is about to be made commercially available by Sigma.
Finally, Adam Clark (NCI) presented the Clinical Proteomic Technologies Initiative for Cancer, a multi-disciplinary approach to refining and standardising technologies and analytical methods.

There was then a panel discussion in which representatives from industry, scientific journals and academia led a lively debate on the role of proteomics standards in the publication process. There was a strong feeling from the floor that it was the role of reviewers, rather than the journals, to impose quality standards on data submitted for publication but that the journals could assist in this process by issuing clear and consistent guidelines. The MIAPE documents provide a checklist to ensure that the required experimental data is included in the article. There is a clear need for accompanying mass spectrometry data to be made available for reviewers but managing the repositories for such data is a role for database providers, not the journals.

The second day started with a discussion about the formal framework of the HUPO-PSI, which needs to provide the organisation with both stability and accountability but also provide a framework for the publication of documentation and for outreach activities. To this end a steering committee has been appointed and a defined number of working groups established each of which must have a working charter in place with a defined set of deliverables and timelines. New working groups must submit such a charter to the Steering Committee for approval. An editorial process has been defined through which documentation moving towards publication must pass. It was suggested that an additional member of the steering committee could be appointed, with responsibility to connect with Government bodies and professional organisations, thus engendering the support of communities whose interests extend beyond proteomics.

Angel Pizarro then spoke about FuGE (http://fuge.sourceforge.net), a framework for developing standards for functional genomics. The FuGE project was developed in coordination with members of both the MGED consortium and HUPO-PSI, but the project retains independent status from both organisations. FuGE provides a model of common components in functional genomics experiments, for example sample processing and protocols, and also allows the integration of pre-existing data formats. A UML model has been produced which uses the AndroMDA UML template engine (http://www.andromda.org) to produce platform-specific models. Existing formats can be tied together using a protocol application class. It is planned to release version 1.0 in Autumn 2006, which will serve as the basis for the second iteration of the MAGE microarray standard and many of the PSI GPS modules, for example GelML and spML. Other PSI groups are evaluating the methodology endorsed by the FuGE project, namely PSI-GE and PSI-PI (analysisXML).

Closely aligned to these efforts is the Functional Genomics Ontology (FuGO) which was described by Susanna-Assunta Sansone (EMBL-EBI). FuGO (http://fugo.sourceforge.net) seeks to develop an integrated ontology that provides both a set of 'universal' terms, i.e. terms applicable across various biological and technological domains, and domain-specific extensions for terms relevant only to a given domain. A top level structural relationship is currently being developed and contributors are currently sorting their terms according to these high level categories in a process described by Trish

Whetzel (U.Penn) on behalf of the transcriptomics community. Daniel Schober (EMBL-EBI) talked about the work of the MSI in developing CVs for instrument-dependent domains and finally Chris Taylor summarised the status of the PSI-MS CVs which are now available in OBO format and are also now being re-evaluated prior to being added to the FuGO efforts.

The session on Protein Modifications commenced with a summary by Sean Seymour of a small working group of experts with representation of at least 6 search engines which agreed to establish a simple CV of common names for modifications, with particular emphasis on artefactual modifications as having the greatest variance in current nomenclature. This CV is ready for final public comments, and is urgently needed, as most users and vendors now use more than 1 search engine. The mass spectometry community often need moiety-centric names, which usually requires a different granularity of names from that required by the biologist. The standard names list will be accompanied by a set of rules to guide the addition of new content, and is accompanied by a list of synonyms. The list is available at the UniMod website (http://www.unimod.org/cross-ref-20060414.xls) for community input until the end of June. It will eventually be aligned with the developing PSI-MOD OBO-based CV. John Garavelli then went on to describe PSI-MOD, which now contains 842 terms arranged in 26 major categories. Each term consists of an ID, a unique short label, a full name, definition, synonyms and additional information such as the elemental formula allowing calculation of mass and mass difference. The ontology will soon be available on the OBO website for user-input (http://obo.sourceforge.net/). Luisa Montecchi-Palaza (EMBL-EBI) then spoke of the general format to be used by all PSI CVs, which are all to be available in OBO format, registered on the OBO site with the PSI-XXX namespace and naming conventions are to be harmonized across all projects.

Finally in this session, Chris Taylor gave an update on the progress of the MIAPE documents. The parent document is now available for publication and several of the domain-specific documents, such as the Mass Spectrometry and Mass Spectrometry Informatics documents are in the PSI editorial process prior to publication whilst the molecular interaction paper has already been submitted to a journal. Several new documents are in preparation and it is hoped that these will begin the editorial process by autumn of this year.

The sessions then separated for their parallel sessions

Track 1 Molecular Interactions

The PSI-MI session started by discussing a number of minor changes to the schema, which had been logged with the SourceForge tracker by various users throughout the year. A maintenance release is planned for early June. About 50 new or updated CV terms were discussed and agreed upon – the editorial board that keeps the CVs up to date between meetings had lost a number of personnel due to job changes so a new board was appointed. The PTM CV within PSI-MI will be maintained in line with and mapped to PSI-MOD.

The Minimum Information about a Molecular Interaction eperiment (MIMIx) paper has been submitted and is now under review. Since this was not pre-released on the web prior to submission, it was agreed that the journal in which it is published will be asked to make the paper open access so that it could be displayed on the PSI-MI web site. A paper describing the PSI-MI 2.5 XML format is currently in preparation. An updated Cytoscape plug-in, that can read both PSI-MI1.0 and 2.5, is in development and will be added to the Cytoscape website when complete. The XSLT script required to transform between the compact and expanded form of the schema has also been updated and made publicly available.

Following user requests, the format of a minimal tabular representation of interaction data was agreed. It was decided to extend the existing Grid format, to allow users to take advantage of tools already written to accept such files. The previously reported plans to produce a common query interface across PSI-MI compatible databases has been dormant for the last few months, since most of the participating databases have been actively preparing for the IMEx collaboration. A need to centralise the documentation undertaken to date was recognised. Finally, Samuel Kerrien (EMBL-EBI) demonstrated a semantic validator which is able to run pre-written rules over the XML schema, for example checking that specific CVs have been used in the correct positions in a file and give a variety of error messages or warnings. The validator is generic enough to adapt to many other PSI XML schemata with each workgroup writing their own rules.

Future plans for this group include the maintenance of both CVs and the XML schema, with requirement being gathered for version 3.0 although there are no immediate plans to develop this. Outreach activities will increase, particularly once the MIMIx paper has been published and the International Molecular Exchange Consortium (IMEx) collaboration has achieved regular data exchange. Increasing the range of tools which support this format will be a major focus of the next few months.

Track 2 Mass Spectrometry

The mass spectrometry session opened with a discussion as to the status of the current mzData 1.05 schema which has now been stable for 2 years. A number of issues have been identified, for example difficulties in annotating only selected peaks within an entry with details of charge and also in describing complex hierarchical additional data. These shortcomings may be best resolved in the next major version of mzData which may also merge with mzXML and American Society for Testing and Materials Standards (ASTM) formats. MS vendors would favour a roadmap to such a major update and mzData remaining largely stable in the interim, and some considerable time was spent analysing the two formats and determining how this may be achieved. It is intended that by September, a provisional plan will be in place detailing how the separate schemas will work together. The development of converters will be supported by the open source community, vendor technical support and should use currently supported vendor API interfaces.

Short term efforts on mzData should concentrate on the production of improved user guides, example documentation and a schema validator tool. The development of a distributed referencing mapping system for spectra, such as LSID would give the ability to refer to individual spectra or other pieces of evidence.

The PSI-Proteomics Informatics (PSI-PI) working group now have responsibility for the production of the mass spectrometry informatics standards, such as analysisXML, which will cover among other things protein identification reporting. The remit of the groups is to produce a UML data model with an XML implementation, example instance documentation, a validation tool, and an accompanying ontology. The use cases were reviewed and expanded upon and the existing version analysisXML reviewed in the light of these use cases. Migration to a UML model should be achieved in time for ASMS in order to generate an XML schema for public viewing.

Track 3 General Proteomics Standard

The group modelling separation techniques (seML) met to discuss use cases and requirements for both a generic sample separation/treatment model and specific models dedicated to explicit techniques such as column separation. A draft UML model has been produced for the latter which has been validated by several users but the accompanying CV still needs to be worked upon.

The MIAPE-Gel documentation has received input from 23 independent reviewers and is now ready for publication. The document encompasses the sample, gel matrix and electrophoresis, inter-dimension process, detection, and image acquisition. A number of minor additions were made to the document.

The contents of the Gel electrophoresis GelML and Gel image informatics GIfML model were reviewed. The development of a common gel CV is now seen as a priority, particularly where searching over terms is required. The CV must be in place for both models to significantly advance. Units of measure should be unambiguously specified wherever possible, and not extensible to support arbitrary units provided through an ontology term. It is intended to develop both models and accompanying documentation into publication by the end of the year, with the GelML interchange format being expected to be the first to be released.

Details of the MIAPE-GI (gel informatics) draft reporting guidelines are at: www.proteomeinformatics.net, where there are facilities for accessing, editing and commenting on the document. The scope of the document covers both raw and scanned images, spots, spot matching, quantities, expression profiles and annotated images. The group is now looking to expand into checking requirements for 1D Gels, improving the object model and data format, developing tools such as converters for validation, and to recruit wider participation into developing use cases and the testing phase.

The MIAPE sample processing group are at the requirements gathering stage and have produced draft documentation to describe both column chromatography and capillary electrophoresis. A number of issues were discussed, including the request from some users for the ability to use the model in a peak-based rather than a fraction-based manner and also the need to address tracking of the life history of a column.

A common session then addressed the status of all the CVs required by the models being prepared as proteomics domains with the FuGE model. An OWL ontology has been developed by Frank Gibson that describes various concepts in proteomics; the names of instances of certain of the concepts in the ontology could be used to populate the OBO format which will be common to all the HUPO-PSI CVs. Domain-specific efforts such as the Gel and Separations working groups can add terms required to populate their data models.

Track 4 – Satellites

The MIACA group discussed the dependence of reporting requirements on the eventual destination of the data, in that LIMS systems, repositories and publications often require differing levels of granularity. The MIACA group have already produced a draft XML schema which has been distributed to a number of groups for testing with experimental data. Future plans include working with the FuGE data model and FuGO to produce an implementation which adds to the current community efforts, enabling experimentalists to accurately represent the perturbation of any cell type with which they are working. Discussions are also underway with the MIARE initiative at the U. Edinburgh with a view to merging these two standards.

The FuGO group identified several areas of commonality across the many groups represented at the meeting, for example the use of units, protocols, assay design and implementation and areas which the groups were aware were not yet being tackled such as the ability to describe the change in state of materials over time. The group is now working actively with the individuals within the PSI workgroups responsible for the development of CVs to ensure that these become part of the FuGO efforts and a non-redundant set of terms is ultimately derived.

The members of the MIxxx group gave presentations on a variety of non-proteomic checklist projects (MIGS, MSI, MIARE, MIACA, and MIFACE) and discussed the need for better integration between checklist initiatives. The result of this meeting was a proposal that the current proliferation of minimum reporting guidelines would need active management to ensure that the users can find the appropriate guidelines for their workflow, that any potential overlaps or duplicated work can be identified and resolved and conversely that gaps in a workflow not covered by documentation can be highlighted. Chris Taylor described how this process should start with a central registry point, such as OBO provides to the ontology community, and Chris Taylor and Dawn Field have agreed to initiate this process and communicate with the wider MIxxx community.

Summary
Much of this meeting was necessarily about process, as the increasing maturity of the HUPO-PSI brings a corresponding need for stability, accountability and credibility and the steering group responded to this need by putting the appropriate mechanisms in place. The explicit aim of this meeting, the outreach to other communities beyond proteomics, has been achieved with great success, also thanks to the organisation of this meeting together with the HUPO Publications Committee meeting. A much more active dialog between PSI, journals, and other standardisation efforts has been initiated. The already existing PSI standards are becoming more widely accepted and are reacting to increased usage by updating to become need novel needs. Several new models will enter testing and documentation phases over the course of this summer prior to publication at the end of the year. A number of the minimum reporting requirements documents are nearing completion, and it is hoped to see several of these in publication this summer. Significant advances should be made in all fields by the next congress in Washington DC at the American Chemical Society headquarters, Sept 25-27, 2006.