

# HUPO Proteomics Standards Initiative Spring Workshop

April 17<sup>th</sup>-20<sup>th</sup>, 2005 – Siena, Italy

**A report provided by: Khaled Mostaguir, Swiss Institute of Bioinformatics, Proteome Informatics Group, Geneva, Switzerland<sup>1</sup>**

The HUPO Proteomics Standards Initiative (PSI, <http://psidev.sourceforge.net/>) aims to define standards for data representation in proteomics to facilitate data comparison, exchange and verification. Founded in April 2002, it involves people from various areas, ranging from data producers and experimentalists, database and software providers, manufacturers and publishers. PSI has already proposed advanced standards for molecular (protein-protein) interactions (MI), mass spectrometry (MS) as well as a new emerging key area to cover the whole of proteomics (General Proteomics Standards: GPS).

In April 2005, about 100 people, representing both academic and industrial organisations, gathered to attend the HUPO PSI Spring Workshop in Siena, Italy (<http://psidev.sourceforge.net/meetings/2005-04/>). The objective of this meeting was to continue the advancement on the already well established MI and MS recommendations and developments, as well as to progress on the General Proteomics Standards. The latter involves a definition of minimum reporting requirements (MIAPE) and an XML format for data exchange (a markup language, PSI-ML) derived from a corresponding data object-model (PSI-OM). It also includes some more specific areas, like sample generation or Gel design and representation (Gel-ML). Ontologies were also one important area to be discussed during this meeting, as well as the need to represent PM (protein modifications) and PTM (post-translational modifications) in a standardized way. More general issues, including mainly interactions and collaborations with other standardization groups (e.g. the MGED: the international organisation for facilitating sharing of functional genomic and proteomic data) were to be presented. A potential combination with other workgroups (into functional genomics: FuGE) - while keeping PSI identity - was also on the agenda. Finally, there was a strong and announced desire to publish concrete results in 2005.

## Organisation of the meeting

The meeting was scheduled to take place between Sunday the 17<sup>th</sup> and Wednesday the 20<sup>th</sup> of April at the University of Siena, Italy. The organizing committee included the EBI (Hinxton, Cambridge, UK), Eli Lilly Corp., the Swiss Institute of Bioinformatics (Geneva, Switzerland), the Department of Molecular Biology (University of Siena,

---

<sup>1</sup> This report only reflects the understandings of its author. It may not necessarily represent the views of the different speakers. Inside this report, some of the author's personal comments are placed between brackets and are always in *italic*.

Italy), the Department of Mathematics and Computer Science (University of Siena, Italy) as well as a local organizing committee formed by the tutors of the master degree in Bioinformatics (University of Siena). Several commercial and academic sponsors also kindly contributed to cover the expenses of this event.

Common activities were to be shared between all attendees in plenary or concluding sessions, while more specific issues were to be conducted in three parallel tracks, namely, molecular interactions (PSI-MI), mass spectrometry (PSI-MS) and the general proteomics standards workgroup (PSI-GPS).

## **Presentations of the current state and objectives for this meeting**

### **The PSI molecular interaction group:**

Henning Hermjakob (EBI) presented to the audience the current state of the PSI-MI development:

The PSI MI XML (a data exchange format for protein-protein interactions, <http://psidev.sourceforge.net/mi/xml/doc/user/>) has a new beta version (2.0), which is already available and shall be discussed during this meeting. Several tools supporting this format are already available. There has been some intensive development on the ontology part (*what was presented was mainly controlled vocabulary rather than ontology in general*). There is also a need for continued exchange of user-submitted data in PSI-MI in order to provide a network of resources, to ensure consistency and avoid redundant curation. The IME consortium (international molecular interaction exchange consortium) is a consortium that was established in summer 2004. The workgroup will mainly focus on the finalisation of the MI XML 2.0 (integrate the IME requirements, improve syntax, extend vocabularies), as well on the review and the update of the PSI MI tools and XML validators.

Luisa Montecchi-Palazzi (EBI) then gave a presentation of the PSI-MI controlled vocabulary (CV) development. A set of different rules and associations, used to manage and control the terms, was shown.

### **The PSI GPS and MS groups:**

Chris F. Taylor (EBI) presented an overview of the evolution of PSI, and agenda items for the upcoming PSI-GPS and PSI-MS sessions (<http://psidev.sourceforge.net/gps/index.html> and <http://psidev.sourceforge.net/ms/index.html>):

A rough summary of the evolution of PSI, since the adoption of the “Pedro” model as a rough draft to start with, is presented. The goal is to reach a *final* PSI-OM (PSI object model). Chronologically, there were the manufacturers’ output formats, which helped to establish the mzData model. Compatibility with search engines and development of tools were then emphasized. There have always been consultations with domain experts, as well as large contributions from the MGED community. At present, various specific MIAPE (minimum information about a proteomic experiment) documents are either in some mature state, or in their initial development

phases. PSI-Ontology has also been launched. On general agreement, we should now go towards the process of generating a more generic PSI-OM (a general PSI object model which should be proposed as a markup language PSI-ML, and optionally, as a PSI-DB for SQL implementation).

To summarize, main focuses will be:

- MIAPE: The minimum requirements to describe a proteomic experiment. To be reinforced by journals and publishers. Two general criteria are employed: sufficiency and practicability. There will be several technology-specific modules, all associated with a parent document. Those modules cover the study, design and sample generation (including administration, motivation,..), the pre-fractionation and sample treatment, liquid chromatography, gel electrophoresis, mass spectrometry, proteoinformatics (MS data analyse). Dealing with the 'legacy' problem, two principles should be applied: data set should be annotated to the fullest possible extent and flagged as legacy, and missing required data and metadata should never be created.
- XML formats for data exchange: The PSI-ML markup language for exchange and submission. Currently the mzData offers a mature version (1.05) with strong vendor support and implementation. There are plans to release a mature representation for mzIdent (conversion of peak lists, search engines, results and interpretations/identifications, unification, cross-comparison) at the next HUPO meeting. All those representations should be based on the general GPS workflow model (PSI-OM).
- Ontology: Ontology should avoid any ambiguity. GPS will contribute terms to the MGED ontology (under the PSI namespace). A lot of work is expected to make ontology accessible to users (e.g. develop tools, etc.)
- Miscellaneous: how to examine the comparability of search engines, how to decide on the reliability of an identification, etc.

## **The MIAPE session (MS and GPS tracks)**

The four documents presented (<http://psidev.sourceforge.net/gps/miape/>) were:

- MIAPE parent document, version 1.0, 6<sup>th</sup> January 2005.
- MIAPE Mass Spectrometry, version 1.0, 6<sup>th</sup> January 2005.
- MIAPE Mass Spectrometry Informatics, version 0.2, 16<sup>th</sup> April 2005.
- MIAPE Gel Electrophoresis, version 0.5, 16<sup>th</sup> April 2005.

Pierre-Alain Binz (SIB) raised the conceptual question if the MIAPE recommendations are intended for repositories or for publication? Angel Pizarro (University of Pennsylvania) commented that there should be no divergence between MIAPE and MIAME, and that they are built by non-computer people and intended for review (*MIAME is the minimum required information guidelines previously set up by the microarray community*). A discussion took place about whether data sets should be included within publications or restricted to data repositories. Randall Julian (Eli Lilly) thought that there is a difference between specific data presentations (e.g. with ontology) and the generic minimal description of an experiment. Pizarro added that

the report must provide ways to specify data types without judgment on data quality. It was concluded that MIAPE is a subset (an absolute minimum) of the GPS, which covers an experiment in a more complete manner, including all its data sets. Weimin Zhu (EBI) put forward that the PSI work is to discuss and submit proposals to experts (experimentalists, publishers,..) for review with relation to the kind of data they may require. Taylor also asked if we should make efforts to raise the level of both publications and repositories, as people are expecting this from PSI. Next, some points were raised on both the MIAPE gel electrophoresis module and the MIAPE MS module. There was a short debate if gel separation techniques are to be bound or separated from acquisition techniques and 2D analysis. After a brief vote on the subject, it was decided to separate them (*the vote was roughly 50%-50%*). Several points specific to MIAPE MS were then invoked, e.g. the description of SELDI techniques as static-fed or LC-fed, as well as some advanced details on properties ('location of parameter file',..).

Binz then presented the MIAPE Mass Spectrometry Informatics in its current state. Reference will be used to point to peak lists (from mzData) rather than raw data. Identified proteins or peptides should be reported in the same way. The problem of artefacts vs. natural PTM also has to be considered. We should further describe the scoring schema (which is sometimes parametrable) within the tool description part. For identified 'molecules': is a description needed (e.g. for isoforms)? are the searched database identifiers sufficient? Some special use-cases where then invoked by Binz. A final issue on the validation and confidence representation was then raised (for confidence, it seems that free text will be, at least initially, adopted).

Taylor then presented the MIAPE Gel Electrophoresis. The document mainly covers the gel preparation and running details, staining, image capture and the bioinformatic analysis (*the new draft - version 0.5 - seemed to miss some important elements*). As time was almost over for this session, he proposed to summarize suggestions and to keep the discussion open by e-mails in the future. He proposed to concentrate the next day on the model/XML proposal.

## **The Gel-ML: Design, critiques and evolution (GPS track)**

Chairman: Chris Taylor

Attendance: approximately 20 people, one third of which were experimentalists.

Should the modelling start from scratch, or should existing models be considered? A brief reminder of mzData format is presented. We should ensure the Gel model to be flexible and generic. Would it be better to break some relevant parts into files in order to avoid redundancy and duplication (e.g. administrative part)? We would then use some unique identifiers for various types of data (cf. RDF).

There was a small presentation of an UML diagram for 2-DE that was supplied by Andrew Jones (University of Manchester) - *see later the FuGE model*. It was agreed that it is not PSI's role to give software vendors guidelines for algorithms requirements, but to only recommend export formats.

The ontology part was then discussed. Jones pointed out that there are three parts to deal with: the ontology schema, the mapping between existing controlled vocabularies and the data schema itself. It was agreed to use namespaces to distinguish between the

ontology and the data schema. He also pointed out that it might be better to concentrate on the data schema, as the ontology might be much more complicated to start with (in the case of gels, ontology is related to human manipulations, rather than to machines properties like in the case of MS). The model should allow data exchange and support MIAPE.

A first plan to put the model on a general context was derived (without any workflow):

- external part (ontology, administration, contact,..)
- pre-treatment (sample origin, including references to external files)
- main (equipments, 2D separation methods, data, analysis)
- post-treatment (identification)

The discussion then focused on the distinction between two main components: the Gel-ML and the ImageAnalyze-ML (not agreed yet on the name to be adopted).

Everybody seemed to agree on the following:

The Gel-ML would include the origin of material, its structure, administration, equipment, separation methods and data (data is then the image itself). Evidence on why a gel is chosen must be given.

The ImageAnalyze-ML should refer to the origin of images, the analysis tool and the data generated by those analyses. There should be evidence on how spots were chosen.

The Gel-ML should offer a way to link between gels, while images may or may not be analysed by bioinformatics. Spots should be referred to by their coordinates. It may be possible to use several tools for analysis and to combine their interpretations.

Finally, we should think of a method to refer to all linked gels within the ImageAnalyze-ML.

This conclusion was then presented to the plenary audience, which was followed by a discussion of some of these items (e.g. reflection on the approach to link several gels to one single image).

Further discussions on the Gel-OM were resumed later on during the next PSI-OM talks. Questions regarding the quality criteria for data to be released are raised (e.g. nice gels but poor annotations or poor gels but high quality annotations), as well as questions on how to deal with collections of gels to generate a single image (see later). In case of combined samples, it should also be possible to trace back the origin of all the combined samples used to generate a gel.

## **The PSI-OM and PSI-ML (GPS track)**

Chairman: Chris Taylor

Attendance: approximately 20 people, one third of which were experimentalists.

The discussion started on the PSI object model in its current state. There was a need to clearly define the effective use of the object model and how it is intended to be implemented. Everybody agreed on the fact that the model is the fundamental representation that subsequent implementations (principally PSI-ML) will need to respect. The only purpose of such a model is to help data exchange and data

submission to repositories, hence, implementations and useful tools are expected to be derived from it.

The discussion went on about the process of comparison. Data needs to be normalized for comparability and for conclusions to be significant.

Regarding a question about the involvement of PSI on evaluating the quality of data, it was clearly stated that this is definitely not a PSI mission. Rules for defining the minimum number of peptides for identification, the quality of an image, and so on, are not PSI tasks.

## **The PSI Ontology and Collaborations (GPS track)**

**The Ontology Working Group** (<http://mged.sourceforge.net/ontologies/>)

The Ontology Working Group, in relation to the MGED project, is in charge of developing ontology to describe samples used in microarray experiments. Trish Whetzel (University of Pennsylvania) gave a presentation with regard to the ongoing collaboration with PSI.

She presented the different editors which had been examined for use with the project: Protégé, DAGedit and OilEd. It was the latter that had been chosen. The model itself is built according to MAGE, though it still needs further organisation and adaptation to expand. Although it is considered to be “good”, rich and updateable, it is mainly MAGE-OM (MAGE object model) centric view. Plans exist to extend the model (ontology) to cover in a more general way functional genomics (inclusion of proteomics, metabolomics,..). To achieve this task, class hierarchy will be rearranged and shared classes (common to the various –omics) will be under the “Common” class. Some work on building middle layers to map between the model and other ontology models should get underway. Finally, to acknowledge people out of the microarray community of its use, the ontology model will be renamed into FuGO (functional genomics ontology).

The future use of Protégé may also be reconsidered, because of its OWL Web Ontology Language plug-ins.

To answer a question regarding the possibility to map between existent ontologies, the speaker said that mapping is possible to some extent, but it would not cover everything.

A mailing list is available at: <http://lists.sourceforge.net/lists/listinfo/mged-ontologies>. MGED mailing lists archives may be found at [http://sourceforge.net/mail/?group\\_id=16076](http://sourceforge.net/mail/?group_id=16076).

Taylor then commented on the collaboration part and the aspiration to converge towards a common ontology for functional genomics. He emphasized the need to provide people (and computers) with clearly designed terms. The fact that this is a politic issue, his final comment was “We do our bit, they do their bits”.

*(Note: it is important to mention that there was an optimistic feeling among the audience about these ongoing advances on the ontology part)*

## Reporting Structure of Biological Investigations Working groups (RSBI WGs)

RSBI is another MGED activity. Susanna-Asunta Sanson (EBI) presented this working group (<http://www.mged.org/Workgroups/rsbi/rsbi.html>).

Following plans of the MGED Society to extend its mission to other functional genomics, proteomics and metabolomics / metabonomics, the MGED Toxicogenomics Working Group is revisiting its mission statement. The Working Group is enlarging objectives and restructuring activities to include other communities, where efforts are already underway to promote standardization and develop databases to facilitate data exchange. The group recognizes the need for a 'single point of focus' for these domains and recommends the formation of a new working group to include toxicogenomics, nutrigenomics, environmental genomics and many other domains of application. The proposed name is the Reporting Structure for Biological Investigations Working Groups (RSBI WGs). Beside the domains' endorsing standards, the group benefits from a large user support and has three main objectives: collecting use cases, optimising interactions and facilitating description (common reporting structure). Concerning the collaboration part, the group is building a "core group" of representatives from various inter-related initiatives, which includes representatives from PSI. Currently, efforts are being made to map to PSI.

This was followed by a discussion about how to represent an investigation. An investigation comprises studies related to assays. What would be exactly the definition of an "experiment" within this representation? Sanson responded that there is no specific definition of an "experiment", that it depends of the context. In RSBI, they use a mixture of study/assay.

The next steps for RSBI would be to:

- Finalise the baseline ontology
- Deliver high level classes ontology
- Shape FuGO with the other ontologies
- Provide use cases for FuGE

The Standard Metabolic Reporting Structure (SMRS) is an open standard for reporting metabolic data. It can be visited at <http://www.smrsgroup.org/>.

Alex Garcia is the RSBI person to contact for collaborative building.

Taylor then talked about the "Metabo meeting" that was held at the EBI on the 7<sup>th</sup> of March 2005. The meeting had representatives from both PSI and RSBI (Sanson and Taylor). The meeting was aimed at setting up a process standard, bringing together existing groups addressing the standardization of resources and exploiting poster interaction with those groups. PSI is actively trying to get this collaboration's input in the PSI process.

*(Note: again, there was an optimistic feeling among the audience about those collaborations)*

Some discussion went then into how to represent algorithms and data structure.

## **Presenting the Functional Genomics Experiment (FuGE) project**

The Functional Genomics Experiment (FuGE) project is defined as follows: “Standards efforts related to functional genomics investigations. Consists of a data model (UML), the Use Cases that drove development, platform specific implementations (Java, XML, Perl, etc.), documentation, and some sample applications” (<http://sourceforge.net/projects/fuge/>).

Angel Pizarro and Andrew Jones gave some presentations outlining what FuGE is and how to fit some proteomics use cases within FuGE – cf. <http://fuge.sourceforge.net/>.

“The MGED society is currently developing a proposal for the second version of MAGE to solve problems identified with MAGE-OM. MAGE-OM could be used to describe any type of functional genomics experiment as long as ontologies exist to describe technology-specific details. A new proposal, called FuGE-OM (Functional Genomics Experiment Object Model) is organised into several parts (namespaces) to make explicit which parts describe microarray experiments and which parts describe a generic functional genomics experiment”.

FuGE-OM is far simpler than MAGE-OM. It covers a wide range of use cases due to its generic structure. There are two namespaces in core: Bio and Common. Currently, verifications are undertaken to see if proteome experiments could be modelled within FuGE-OM. Several use cases were presented, including description of protocols, studies, sample preparations, and materials. Proteome workflows for 2D LC-MS use cases were assimilated. The issues are now focused on how FuGE can help on PSI model development, on how the “Experiment” package may capture all proteome / RSBI use cases and on how to integrate specific models (e.g. mzData) into FuGE.

The FuGE developers aim to extend collaboration with the PSI and the proteomics community by:

- XML schema generation for creating test data sets.
- Defining available use cases to demonstrate correct use of model.
- Arranging weekly conference calls to solve modelling issues.
- Encouraging PSI involvement in further development of the Bio and Common models.

*(Note: It is to be mentioned that the audience seemed to be highly impressed by the FuGE presentations. Many thought that we could have started the discussion on the GPS track based on what was already available in FuGE).*

## **Lightning talks (during the GPS sessions)**

During the GPS sessions, some lightning talks were planned. Talks were initially intended to be short, of about 5-10 minutes (*many of them largely exceeded this amount of time*). Chronologically, they were given in the following order:

- Proteios, an initiative for the development of an open source system for storage, organisation, analysis, and annotation of proteomics experiments (<http://www.proteios.org/>) - by Per Gärdén (Lund University).
- CEBS (Chemical Effects in Biological Systems, <http://cebs.niehs.nih.gov/>), a knowledge base to house data from multiple complex data streams in a systems friendly manner that will accommodate extensive querying from users - by Sandhya Xirasagar (Science Applications International Corporation).
- A Repository Development Strategy (using XML instances of RDF) to build a repository for *mouse*, including MS/MS storage and analysis – by Mark Igra (Fred Hutchinson Cancer Research Centre).
- The International Protein Index, a database that provides a top level guide to collection of databases that describe the proteomes of higher eukaryotic organisms (<http://www.ebi.ac.uk/IPI/>) – by Paul Kersey (EBI).
- ProteusLims, a LIMS specifically designed for proteomics (<http://www.genologics.com/products/proteuslims/>) - by James DeGreef (GenoLogics Life Sciences Software).
- The Make2D-DB II, a tool to build and interconnect 2-DE databases (<http://www.expasy.org/ch2d/make2ddb.html>) - by Khaled Mostaguir (Swiss Institute of Bioinformatics).

## **The Protein Modifications (PM) presentations (plenary session)**

John S. Garavelli (EBI) presented the “RESID Database of Protein Modifications” which is a collection of annotations and structures for protein modifications (<http://www.ebi.ac.uk/RESID/>). He insisted on the necessity for a controlled vocabulary for protein modifications (pre-, co- and post-translations). There is a need for standards on describing the process of modification as well as on naming the modified proteins. Garavelli also insisted on the fact that no real single experiment gives all the info to identify an unknown modification. The PM have to be captured as accurately as reported. The mission of the PM workgroup would be to propose a CV (controlled vocabulary) and the appropriate standards for reporting.

David Creasy (Matrix Science) then gave a presentation on UNIMOD database (<http://www.unimod.org/>). UNIMOD aims to create a community supported, comprehensive database of protein modifications for mass spectrometry applications. This presentation was followed by a practical demonstration of RESID (Garavelli). The demonstration was accompanied by a long discussion on PM and their curation on RESID implying all the audience. Regarding the controlled vocabulary part, Montecchi-Palazzi pointed out the problem of reported aliases for some modifications. Hermjakob concluded that the group would have to come up with a common hierarchical ontology (top-down, GO-like).

## **General assembly and closing remarks (plenary)**

### **The PSI molecular interaction group:**

Henning Hermjakob gave a summary of the MI activities during the meeting:

The PSI-MI new schema has been discussed and updated. There has been some simplification on its structure: addition of controlled vocabularies for kinetics, sample preparation, typing of XRefs typing of aliases as well as integration of terms related to BioPAX (currently there should be a nice direct mapping between PSI and BioPAX). Complex assemblies are now well presented. The updated schema should be available upon request from the sourceforge site.

Regarding the tools, there is a PSI-MI XML validator: Syntactic validation (standard XML) and semantic validation of controlled vocabulary (based on XPATH).

Concerning the common interface, the group defined the specification of PSICQIC (the PSI-MI Common Query Interface). The interface will be based on web service simple queries and will always reply in a valid PSI-MI format.

Plans are announced as follows:

- 20.04.2005: change requests integrated.
- 01.05.2005: new development schema.
- 15.07.2005: freeze of development and announcement.
- 15.09.2005: release of MIF 2.5 and submission to BMC Bioinformatics.

### **The PSI-MS group**

Randall Julian gave a summary of both the mzData and mzIdent discussions:

#### mzData:

- mzData vendor release underway.
- Annual schema major release (next release spring 2006).
- Continuous controlled vocabulary releases (current CV release on sourceforge is 1.0).
- Next minor release will have: XML namespace version, merged CV with mzXML and improved object identification.

#### mzIdent:

The schema conception was started. It is based on pepXML (ISB), which is a format for storing the results of database search at the peptide level. There is a generalisation on what is identified (a 'molecule' type, rather than a peptide or a protein entity). The mzIdent controlled vocabulary (CV) is a generalized version of the pepXML attributes. Protein modification CV is under development (Garavelli). Vendors are also expected to supply some CVs. A detailed CV covering MS, molecules, PTMs and analysis should be released on sourceforge, until a more stabilized mechanism takes place for the GPS ontology.

There should be a single schema for MS analysis. The schema will be in pure XML with its naming and structure consistent with mzData. There should also be some optimisation on the file size, and use case development is on its way.

The model itself is composed of four components: data, parameters (software, algorithms,..), results and molecules (identified entities). The use cases will determine the various relationship between those components.

Quantification has not been approached during the sessions. A workshop is announced to take place in autumn 2005. It should include quantification refinement and integration with GPS. The goal is to release a mzIdent schema version 1.0 by the end of the year.

### **The PSI-GPS workgroup**

Chris Taylor summarized what was described earlier for the various GPS sessions.

### **Final remarks**

*(As a whole, the participants seemed to be fairly satisfied with the outcome of this meeting).*

The next PSI meeting will be take place in autumn 2005. Its location is still to be announced. To stay updated, visit the PSI site at: <http://psidev.sourceforge.net/>.