

The HUPO Pre-Congress
Proteomics Standards Initiative Workshop
HUPO 5th Annual World Congress
Long Beach, CA, USA
28 October–1 November 2006

REPORT

Sandra Orchard¹, Andrew R. Jones², Christian Stephan³ and Pierre-Alain Binz^{4, 5}

¹ EMBL Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK

² School of Computer Science, University of Manchester, UK

³ Medizinisches Proteom-Center, Ruhr-Universität Bochum, Germany

⁴ Swiss Institute of Bioinformatics, Proteome Informatics Group, Geneva, Switzerland

⁵ Geneva Bioinformatics (GeneBio) SA, Geneva, Switzerland

The plenary session of the Proteomics Standards Initiative of the Human Proteome Organisation provided an opportunity to update delegates on the progress of the work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO-PSI) to develop and implement standards in the field of proteomics. Significant advances have been made since the previous congress, with several of the interchange standards and minimal requirements documents being submitted for publication in the literature and being more widely adopted by both manufacturers and data repositories. An exciting development over the interim twelve months is the ongoing merger of the two existing mass spectrometry standards, the PSI mzData and Institute for Systems Biology mzXML, into a single product. This should be achieved by early in 2007.

Received: January 5, 2007

Revised: January 19, 2007

Accepted: January 22, 2007

Keywords:

Data standardization / Human Proteome Organisation / Proteomics Standards Initiative

Introduction

Attendees to the Human Proteome Organisation's Proteomics Standards Initiative (HUPO-PSI) were welcomed by the chair, Pierre-Alain Binz (Swiss Institute of Bioinformatics) who summarised the aims of the group, namely to produce, in collaboration with the scientific community, data exchange formats and minimal reporting requirements for the key processes in a typical proteomics workflow. These

deliverables enable results from proteomics experiments to be accurately reported, the data to be exchanged between collaborators, deposited in public domain repositories and be available for use by other workers in the field. He then briefly outlined the program for the morning's session before handing over to the first speaker.

Molecular Interactions

Sandra Orchard (EMBL-EBI, UK) summarised the work of the Molecular Interaction (MI) workgroup of HUPO-PSI. The MI group was the first PSI group to publish an XML interchange standard in 2004 with accompanying controlled vocabularies [1]. The PSI-MI XML1.0 was widely adopted by molecular interaction databases, many of whom made data available in this format, but it soon became apparent that the simple representation of protein:protein interaction data made possible by this format was too restrictive for many users. To this end, a more flexible version 2.0 was developed and quickly

Correspondence: Sandra Orchard, EBI, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SD, UK

E-mail: orchard@ebi.ac.uk

Fax: +44-1223-494-468

Abbreviations: **HUPO-PSI**, Human Proteome Organisation's Proteomics Standards Initiative; **MI**, molecular interaction; **MIMIX**, Minimum Information Required for reporting a Molecular Interaction Experiment

advanced to a stable release version 2.5. PSI-MI XML2.5 allows the user to describe interactions between any type of molecule and is available in two forms – the compact, in which repetitive elements such as experiment descriptions and interactors are described only once and referenced throughout which is suitable for the description of large datasets, and the expanded form, in which all elements are described in full and related data is grouped together (Kerrien, S. *et al.*, *in preparation*). Additional features include the ability to describe the hierarchical build-up of complexes, the flagging of both negative and modelled data and an increased ability to handle quantitative data such as kinetic measurements.

The accompanying CVs have also been expanded and updated to enable an increased depth of annotation of an interaction. The CVs are now indirectly mapped to the XML schema allowing the use of vocabularies developed by other workgroups in the PSI, when appropriate. In response to user-demand, a simple tabular format of undirected binary data has also been developed (MITAB2.5). Tool development has continued in parallel to support both formats. Finally, the Minimum Information Required for reporting a Molecular Interaction Experiment (MIMIX) guidelines have been written and accepted for publication and may currently be viewed on the Nature Biotechnology website (<http://www.nature.com/nbt/consult/index.html>) where it is being displayed prior to final publication for further community input [2].

Gel Electrophoresis

Andrew Jones (University of Manchester, UK) reviewed the progress of the Gel Electrophoresis workgroup. The MIAPE-Gel module has been submitted for publication and is currently under review. It is anticipated that, as for the MIMIX paper, the document will go through a further round of community review and consequent update prior to final publication. The GelML XML interchange format for representing gel electrophoresis experiments is now close to a stable version. Recent alterations to the model include improving the representation of constituents of the gel matrix and the existence of gradients of the constituents. Further changes have been made to facilitate the recording of scanner calibration information required for image analysis. It is planned that the GelML model will enter the formal PSI review process by the end of 2006.

The controlled vocabularies are now a focus of current efforts and guidelines will be added to the specifications as to their usage. The GelCV is to be merged with the sample processing and general separations CV to produce a single, unified resource known as sepCV.

The Gel group is currently developing a MIAPE document for informatics performed on gel images and a corresponding data interchange format called GelInfoML. Gel InfoML will be the main focus in 2007, and additional developers with an interest in gel informatics are encouraged to join the effort.

Mass Spectrometry

The major activity of the Mass Spectrometry workgroup over the last few months, as summarised by Pierre-Alain Binz, has been to push forward the merger of the HUPO-PSI mzData XML interchange format for MS traces and peak list data with that of the ISB mzXML [3]. This effort was significantly advanced at the Autumn PSI workshop in Washington [4], and a further meeting is planned for November in Seattle to finalise version 1.0 of the unified standard. An instance document has been prepared to both elucidate and demonstrate the merged format and it is intended to make better use of the PSI-MS CV with less reliance on named XML elements/types within the schema. Refined representation of instruments and data processing more realistically representing the multiplicity of hardware and software components will also be achieved.

Work on analysisXML has also advanced, with the information required for the identification workflow and spectral library search modelled, although this has yet to be completed for quantitation parameters. The CV terms for protein and peptide identification have been developed, though again this effort is still ongoing for quantification and for input parameters.

ProDaC

Christian Stephan (Medizinisches Proteom-Center, Germany) gave the delegates a brief overview of a European-funded effort (ProDaC) to assist in the coordination of the development of the HUPO-PSI data interchange standards for high-throughput proteomics data, the implementation of these standards in data submission pipelines and systematic data collection *via* a network of public domain standards-compliant repositories. Many of the instrumentation manufacturers, of the proteomics software vendors and of various labs are involved in this process. Likewise, journals, such as PROTEOMICS, are playing an active role in the group. Associated partners from around the world will join the core group by contributing data into the pipeline, once established.

MIAPE

Pierre-Alain Binz updated the attendees on the progress of the MIAPE documentation – the minimum reporting requirements for domain-specific technologies, linked by a single over-arching document, the MIAPE parent. There are a total of ten domain specific modules in preparation, the first of these are currently under journal review alongside the parent document. These documents are aimed at assisting the scientist in writing papers which both meet the editorial requirements of the individual journals and contain all the information required for a reader to understand, and if appropriate reproduce, the experiment. Each of these docu-

ments has been prepared following extensive consultation with domain-specific experts and has been published on the HUPO-PSI website for community input. Each will be registered with MIBBI (Minimum Information for Biological and Biomedical Investigations). This will be a minimum information checklist resource, which will provide a central point for the community to access all these and related standards which have been developed over the last few years. This is being jointly developed by three organisations, the PSI, the Genomic Standards Consortium (GSC) and the MGED Reporting Structure for Biological Investigations (RSBI). In addition to providing a 'one-stop shop' for researchers, journal editors and reviewers, and funders to these documents, the management of the MICheck Foundry, will ensure that each domain remains unique, whilst ensuring that complementary reporting requirements can easily be integrated by the user.

All such efforts require support from the user community and the PSI is actively seeking input and advice from all

quarters. Anyone wishing to become involved is invited to visit <http://www.psidev.info>, to participate in the discussion groups listed, and to contribute to the further development of community standards for proteomics data. Details of the Spring Workshop April 23-25, 2007 in Lyon, France will be made available soon; participation at this meeting is welcomed.

References

- [1] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J. *et al.*, *Nat. Biotechnol.* 2004, 22, 177–183.
- [2] Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L. *et al.*, *Nat. Biotech.* 2007, in press.
- [3] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, *Nat. Biotech.* 2004, 22, 1459–1466.
- [4] Orchard, S., Taylor, C., Jones, P., Montecchi-Palazzo, L. *et al.*, *Proteomics* 2007, 7, 337–339.