Published in final edited form as:

Proteomics. 2012 September ; 12(18): 2767–2772. doi:10.1002/pmic.201270126.

Ten Years of Standardizing Proteomic Data: a report on the HUPO-PSI Spring Workshop 12–14th April 2012, San Diego, USA

Sandra Orchard¹, Pierre-Alain Binz², Christoph Borchers³, Michael K. Gilson⁴, Andrew R Jones⁵, George Nicola⁴, Juan Antonio Vizcaino¹, Eric W. Deutsch⁶, and Henning Hermjakob¹

¹EMBL Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK ²Swiss Institute of Bioinformatics, Geneva, Switzerland ³University of Victoria-Genome British Columbia Proteomics Centre, Victoria, British Columbia, Canada ⁴Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, CA, USA ⁵Institute of Integrative Biology, University of Liverpool, United Kingdom ⁶Institute for Systems Biology, Seattle, WA, United States

Abstract

The Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) was established in 2002 with the aim of defining community standards for data representation in proteomics and facilitating data comparison, exchange and verification. Over the last 10 years significant advances have been made, with common data standards now published and implemented in the field of both mass spectrometry and molecular interactions. The 2012 meeting further advanced this work, with the mass spectrometry groups finalising approaches to capturing the output from recent developments in the field, such as quantitative proteomics and SRM. The molecular interaction group focused on improving the integration of data from multiple resources. Both groups united with a guest work track, organized by the HUPO Technology/Standards Committee, to formulate proposals for data submissions from the HUPO Human Proteome Project and to start an initiative to collect standard experimental protocols.

1 Introduction

The HUPO Proteomics Standards Initiative (HUPO-PSI) have published data exchange formats, controlled vocabularies and reporting guidelines for many aspects of the identification and functional studies on proteins by gel electrophoresis and mass spectrometry [1]. Over the last 10 years these standards have been widely implemented and instrumental in promoting data sharing, with standards-compliant databases exchanging data and sharing curation effort. The 2012 meeting was hosted by Michael Gilson and George Nicolas at The Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego. Attendees were welcomed by the outgoing Chair, Henning Hermjakob (EMBL-EBI).

Michael Gilson (UCSD) gave an introduction to BindingDB (www.bindingdb.org), a database of measured binding affinities, focusing on the interactions of proteins considered to be drug-targets with small, drug-like molecules. Binding affinity data is collated, for both active and inactive compounds and made available to the research community to enable analysis of pharmacophores active against a particular target, to enable the validation of

computational models of ligand binding and to facilitate the prediction of potential drug side-effects. Curation is increasingly coordinated with ChEMBL (www.ebi.ac.uk/chembl) and the Psychoactive Drug Screening Program database (<http://pdsp.med.unc.edu>). Target links allow the user to access additional information in resources such as UniProtKB, pathway databases and PubMed. Efforts to encourage data deposition by authors are underway, as it is often difficult to curate the published articles – drawings are not machine readable and often do not show the complete scaffold and details of side chains may be in a separate table.

Parag Mallick (Stanford U.) described the ProteoWizard project, a set of modular and extensible open-source, cross-platform tools and software libraries that facilitate proteomics data analysis [2]. ProteoWizard provides a framework for rapid development of data analysis tools and provides reference implementation of the HUPO-PSI mzML mass spectrometry data format and mzIdentML peptide/protein identifications format. By bridging open-source and vendor specific formats via a unified data access interface, the resource allows cross-lab and cross-instrument comparisons to be undertaken. Controlled vocabulary terms play an essential role but need to be easy to modify and/or add, and ways need to be found to enforce semantic correctness and incorporate all nuances of a term when used. There are still issues with legacy files based on old software and new formats will need to be incorporated into the ProteoWizard framework when published. Large files currently slow down processing time and speed bottlenecks need to be analysed and overcome.

Sandra Orchard (EMBL-EBI) gave an update on the progress of the IMEx Consortium, a network of currently 9 data resources which collectively provides users with a non-redundant set of protein interactions. The consortium has been in production mode since February 2010 and a publication describing its work has just been accepted [3]. Juan Antonio Vizcaíno (EMBL-EBI) described parallel efforts to share mass spectrometry proteomics data across multiple resources: the ProteomeXchange Consortium. Tandem MS data will be submitted to PRIDE, and SRM data to the PeptideAtlas SRM Experiment Library (PASSEL), as the initial submission points. A supporting raw file archive is being developed at the EBI. The integrated pipeline should be publicly available by June 2012. A submission tool for MS/MS data has been developed and a dataset tracking tool, ProteomeCentral, is being established. Digital Object Identifier (DOIs) will be assigned to datasets to enable their citation when reanalysed.

The meeting then separated into separate work tracks.

2 Molecular Interactions (Henning Hermjakob and Sandra Orchard, EMBL-EBI)

The PSICQUIC web service has been a marked success story with 25 services currently displaying evidence for over 152 million interactions [4]. Work is now needed to enable the management of such a volume of data and enhance the user experience. The discussion on data usage was triggered by a talk given by Gary Bader (U. Toronto). The Bader group is working to predict pathways and interactions of entire genomes. Pathway Commons integrates pathway and interaction data from major public domain resources, and provides a download facility and API with additional web services planned (<http://www.pathwaycommons.org>). Cytoscape 3.0 is due for release this year and is available in beta format for testing (www.cytoscape.org). This version of the service is planned to be more modular with stable APIs, making it easier to write plugins, scripting and macros and to operate in command line mode. GeneMANIA finds genes that are related to a set of input genes, using functional association data (<http://www.genemania.org/>). Association data

include protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. An automatic updating pipeline has been built to bring in the data from external resources. GeneMANIA was released as a PSICQUIC delivering 120M interactions just prior to this meeting.

Users need data summaries, the clustering of all evidences of Protein A interacting with Protein B across all these resources. This is not trivial, because different databases have chosen different ways of filling the format fields. A Data Distribution Best Practice document was agreed upon, with a joint decision taken as to the information which should be included in each field. A MITAB validator is now required to ensure that the file is correctly formatted and a data enricher is also planned. This will take the minimal information in the file and use web services (e.g. PICR, UniProtKB) to add more information to fields such as the molecule details, CV terms and organism details in a consistent manner across multiple files. Users also need to be able to differentiate experimental data from predicted/text-mined and also primary data sources from imported information. PSICQUIC is currently based on the very limited MITAB2.5 format and fails to capitalise on the detail supplied by most manually curated databases. There is a recognised need to develop PSICQUIC MITAB 2.7 and eventually PSICQUIC XML. A Hackathon is planned in the UK in May and potential aims, such as a reference PSICQUIC implementation in SOLR, XML indexing and adding more download options such as BioPax, RDF, XGMML, JSON to enable more use of PSICQUIC, for example in the semantic web.

During the controlled vocabulary session a nomenclature system for molecular interaction confidence scoring systems was discussed but not resolved upon. Kim van Roey (EMBL) described the use of the XML schema to model cooperative interactions such as allosteric binding or cases where a required complex pre-assembly is necessary for binding. The proposed model was reviewed, and necessary CV terms approved. The possibility of replacing the database CV terms in the PSI-MI CV by using the MIRIAM registry was discussed (<http://www.ebi.ac.uk/miriam>). This would avoid a duplication of effort, and allow the use of the identifiers.org service to maintain links and direct users to the most stable version of any service. However, this would not be a trivial change as all PSI-MI tools are currently written to look for PSI-MI terms, and the CV hierarchy is used in applications such as the XML validator. A compromise of cross-referencing MIRIAM within the PSI-MI CV and indirectly using these resources was agreed upon. Finally, a request from the Semantic Systems Biology group, Norwegian University of Science and Technology for more terms relating to transcription factor/gene binding was reviewed and generally agreed.

The need for a common API to work with molecular interactions was discussed. This API could be an interface that PSI-XML 2.5 and MITAB would implement. It would simplify the work of developers in that tools such as the PSI-MI validator, data enricher, protein update module written by IntAct would use the common interface to read/write files and could work with both XML and MITAB without having to deal with two different APIs. It was agreed that this would be a long term goal of this work track to produce and implement such an API.

3 Mass Spectrometry (Eric Deutsch, ISB, USA)

TraML, the standardized format for encoding transition lists and associated metadata, has recently been published [5]. The use of TraML will facilitate the exchange of transition lists between groups, reduce time spent handling incompatible list formats and increase the reusability of previously optimized transitions. A format for SRM data analysis result format is required but it is hoped that mzQuantML will fulfil that need. Outstanding issues

discussed during the meeting were the enhancement of small molecule representation in the PSI-MS CV, with external identifiers such as the Human Metabolome Database (HMDB), KEGG, ChEBI, PubChem, Inchi, SMILES, etc. being added. Additional desired implementations of the standard by resources such as Skyline, SRMATlas and OpenMS were identified and will be contacted. The question of whether iSRM (Thermo) and TMRM transition list modes were fully supported was also discussed and will need to be resolved. Finally, it was decided not to produce a separate MIAPE-TraML, but rather to incorporate the needs of SRM into the MIAPE-Quant document and produce specific example documents. It was proposed that this should be presented to the HUPO New technologies committee, as well as being reviewed through the HUPO-PSI documentation process.

mzML continues to be stable, with ongoing maintenance of the CV the major effort currently required. An update on cooperative development with imzML was mentioned. There are currently discussions with the metabolomics community as to the adoption of mzML by this group.

M. Walzer (U. Tuebingen, OpenMS group) gave a presentation about a new XML format called qcML, developed in the context of the PRIME-XS EU grant (<http://www.primexs.eu/>). The idea is to develop a file format that can provide a way of reporting any quality control (QC) measurements that have been performed. For instance, QC measurements for MS data would be encoded in a qcML file and there would be references in that file to the corresponding mzML file.

Finally, the MS group also discussed recommendations to be made to the HUPO Human Proteome Project (HPP) teams as to data deposition requirements for recognised HPP datasets. There are two approaches to be taken by this major initiative, chromosome-centric and a biology/disease driven, but both will be using the same data submission routes. The recommendation will be that all contributions should be backed by well-annotated raw data and these data should be submitted to an approved data repository and given a ProteomeXchange accession number. Protein interaction data should be submitted to an IMEx database, checked by an IMEx curator and issued with an IMEx accession number. A repository for gel images was not finalised, although ProteoRed was one possibility, nor for immunohistochemistry. Poly-nucleotide sequence data should be submitted to the INSDC (www.insdc.org/).

4 Proteomics Informatics (Andy Jones, U. Liverpool)

The manuscript describing mzIdentML v1.1, the data standard for peptide/protein identification data from MS, has been recently published [6]. This format is now regarded as stable although work is still on-going to increase the coverage of terms in the PSI-MS CV. In particular, discussions centred around the description of ambiguity in protein inference i.e. when a peptide could belong to one or more proteins. A number of proposals have been made to address this issue and a consolidating document will be circulated by Sean Seymour and Terry Farrah for further discussion, attempting to define CV terms for protein group membership and the approach used by software for resolving ambiguity. It was also acknowledged that search engine scores associated with the positions of modifications to peptides is not handled natively in the mzIdentML schema; there is no simple way to express that the location of a modification is ambiguous or to associate a score with this ambiguity. A proposed solution to this issue was discussed and is to be tested against examples collected from the most popular search engines. There is also a known issue with the speed of file parsing, resulting from large searches, and the possibility of an indexed version was discussed. The main focus for the next 12 months for this group, however, is to encourage uptake and implementation of this format.

Version 1.0 of mzQuantML, the format for capturing the output of quantitative software used in MS-based proteomics, was submitted to the PSI document process in October 2011. Reviewers commented on the potential for heterogeneity in how different implementations would use the format. They also identified cases where the example files were not consistent with the accompanying rules. This points to a clear need for semantic validation rules to check that the recommendations are being followed for each type of quantitation technique. At the meeting, a draft set of validation rules were defined and it was agreed that the University of Liverpool group would work on the validator software. A revised version of the mzQuantML specifications will be re-submitted to the PSI process in the coming months.

MIAPE-MS and MIAPE-MSI [1] are currently being updated to remove all requirements for information of quantitative procedures. This will now be only outlined in the new MIAPE-Quant document. All three documents will be submitted to the document review process once this is complete.

mzTab is a light-weight, tab-delimited file format for proteomics and metabolomics data coming from MS-based approaches, primarily designed as a way to communicate the *final* experimental results, to enable biologists to access and process such data in a standard and straightforward way, but also to present the data using protocols similar to DAS (<http://code.google.com/p/mztab/>). It will soon be submitted to the document process, and substantial refinements are being made as a result of that process. There are currently two main focuses to the further development of mzTab – the provision of parser libraries to allow developers to integrate mzTab into their applications and the provision of conversion tools from the most commonly used file formats in MS proteomics to mzTab.

The PSI Extended FASTA Format (PEFF) was described by Pierre-Alain Binz (SIB). This aims to incorporate detail on known protein sequences, including isoforms, maturation events, sequence variants and post-translational modifications into a FASTA header, which can then be read by any search engine. Proposals were made that allow the format to include sequence and PTM ambiguities where the existence, but not the positions are known. Qualifiers such 'by similarity' can also be added. Non-XML based separators will have to be carefully chosen as these might currently cause parsing failures. Again the need for improvements to the PSI-MS CV was noted. Viewer and converter tools have been written but need to be updated, and a range of implementation options should be explored but the format is essentially ready for publication. neXtprot have already released human sequence data in this format.

5 HUPO Technology/Standards Committee Bruno Domon (Luxembourg Clinical Proteomics Centre) and Christoph Borchers (University of Victoria-Genome BC Proteomics Centre, Canada)

A guest track with the subject "Harmonization of proteomics methods / Quantification methods" was invited to participate in this meeting. As HUPO begins to work on the Human Proteome Project [7], the collection of standardized experimental protocols and centralization such that they can be easily accessed is essential. Methods should first be submitted to a review committee then sent out to approved laboratories for testing with standardized reagents, before the protocol is accepted as fully validated. Data analysis protocols should also be included, with reference datasets and results of the analysis of these reference sets, although it was noted that this would require a certain amount of common infrastructure in place, such as the decoy database. The protocol will be published for the community to both use and comment upon. The protocols could potentially then be

published using the same mechanism as is currently in place for the HUPO tutorials, with both the original submitter and validating laboratories sharing credit.

This highly successful meeting was then brought to a conclusion by the incoming Chair of HUPO-PSI, Eric Deutsch (ISB). A number of changes in the composition of the steering committee and of the working groups were agreed that can be found in the redesigned HUPO PSI website (www.psidev.info). Delegates were thanked for attending, and the location of the 2013 meeting discussed.

Acknowledgments

This study was made possible in part by grant GM70064 from the NIGMS. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

7 References

1. Orchard S, Hermjakob H. Data standardization by the HUPO-PSI: how has the community benefitted? *Methods Mol Biol.* 2011; 696:149–160. [PubMed: 21063946]
2. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics.* 2008; 24:2534–2536. [PubMed: 18606607]
3. Orchard S, Kerrien S, Abbani S, Aranda B, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods.* 2012; 9:345–350. [PubMed: 22453911]
4. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods.* 2011; 8:528–529. [PubMed: 21716279]
5. Deutsch EW, Chambers M, Neumann S, Levander F, et al. TraML--A Standard Format for Exchange of Selected Reaction Monitoring Transition Lists. *Mol Cell Proteomics.* 2012; 11:r111.015040. [PubMed: 22159873]
6. Jones AR, Eisenacher M, Mayer G, Kohlbacher O, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics.* 2012:M111.014381. [PubMed: 22375074]
7. Legrain P, Aebersold R, Archakov A, Bairoch A, et al. The human proteome project: Current state and future direction. *Mol Cell Proteomics.* 2012:O111.009993.