Published in final edited form as:
Proteomics. 2011 November ; 11(22): 4284–4290. doi:10.1002/pmic.201190117.

Enabling BioSharing – a report on the Annual Spring Workshop of the HUPO-PSI EMBL-Heidelberg, Germany, 11–13th April 2011

Sandra Orchard[1], **Juan-Pablo Albar**[2], **Eric W. Deutsch**[3], **Martin Eisenacher**[4], **Juan-Antonio Vizcaino**[1], and **Henning Hermjakob**[1]

Sandra Orchard: orchard@ebi.ac.uk

[1]EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[2]ProteoRed, National Center for Biotechnology, CSIC, Madrid, Spain

[3]Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109-5234, USA

[4]Medizinisches Proteom-Center, Ruhr-Universitaet Bochum, Bochum, Germany

Keywords

Data standardization; Human Proteome Organisation; Proteomics Standards Initiative

Introduction

The meeting was opened by the Chair of the HUPO-PSI, Henning Hermjakob (EMBL-EBI) who thanked delegates for attending and acknowledged the funding for this meeting which came from two major EC FP7 grants, PSIMEx and ProteomeXchange.

The first speaker was Anne-Claude Gringras (Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto) who described the incorporation of the HUPO-PSI data interchange formats for both mass spectrometry and molecular interactions into ProHits [1], an open source software platform for MS-based interaction proteomics that manages the entire pipeline from raw MS data files to fully annotated protein-protein interaction data sets. The group combines affinity proteomics data with that obtained using orthogonal approaches such as siRNA screens and LUMIER, and are using such approaches to explore specific interactomes and the dynamics of interactions across the cell cycle.

The complete ProHits software package consists of two main components: a 'Data Management' module, and an 'Analyst' module supported by an 'Admin Office' module, in which projects, instruments, user permissions and protein databases are managed. A simplified version ('ProHits Lite'), consisting only of the Analyst module and Admin Office, is also available for users with preexisting data management solutions or who receive precomputed search results from analyses performed in a core MS facility. Compliance with the MIAPE (Minimum Information about a Proteomics Experiment) [2] guidelines is fully

Correspondence to Sandra Orchard, EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK **Fax:** (0)1223-494-468.

integrated into ProHits via linked search engine servers. The Analyst module organizes data by project, bait, experiment and/or sample. An annotation page tracks experimental details including descriptions of the Sample, Affinity Purification protocol, Peptide Preparation methodology and liquid chromatography-tandem MS (LC-MS/MS) procedures. The mzML format [3] is incorporated into this module. Controlled vocabulary lists for experimental descriptions can be added by drop-down menus to facilitate compliance with annotation guidelines, such as MIAPE and MIMIx (Minimum Information about a Molecular Interaction Experiment) [4]. For interaction data, ProHits generates MIMIx compliant reports, in both MITAB2.5 or in PSI-MI XML2.5 formats appropriate for submission to molecular interaction databases. MS raw files associated with a given project can also be easily retrieved and grouped for submission to data repositories, such as Tranche or PRIDE. The major lesson learned from this exercise is the importance of usability – users only want to see a user-friendly graphical interface, not the formatted files.

Two major collaborations have grown out of the work of the HUPO-PSI. Sandra Orchard (EMBL-EBI) spoke about the work of the International Molecular Exchange Consortium (IMEx) which is an increasing network of molecular interaction databases which have agreed to share curation effort, annotating selected publications to a common set of curation rules and making a non-redundant set of protein interaction data available on a common website (www.imexconsortium.org) via a specific PSICQUIC service. Eight databases currently contribute to this effort. Curation is centrally managed with a database (IMExCentral) both ensuring that each publication is non-redundantly curated and assigning IMEx accession numbers to both the paper and the data is describes. Juan-Antonio Vizcaíno (EMBL-EBI) then gave an update on the ProteomeXchange consortium [5] (www.proteomexchange.org), more recently established to provide a single point of submission to proteomics repositories, and encourage the data exchange and sharing of identifiers between the repositories so that the community may easily find datasets in the participating repositories. Both these efforts are currently funded by EC FP7 grants.

María Martín (EMBL-EBI) described planned updates to the UniProt databases, which will improve their services for the Proteomics community. The International Protein Index (IPI) was created at the EMBL-EBI in 2001 to reconcile differences in gene predictions between different databases for key eukaryotic organisms, by clustering protein sequences from different databases (e.g. UniProt, Ensembl and RefSeq) to provide non-redundant complete data. Ongoing collaborative efforts between Ensembl, Refseq and UniProt to improve gene prediction quality coverage for many of the most-studied genomes have now removed a need to maintain this resource and it has now been decided to replace IPI with Complete Proteome sets available from UniProtKB, which can be accessed from the website or downloaded as fasta files, with isoform sequences included as separate entries, available on the ftp site for upload into search engines. All full-length predicted protein sequences from Ensembl that were absent from the UniProtKB/Swiss-Prot complete human and mouse proteomes have been imported into the unreviewed component of UniProtKB, UniProtKB/TrEMBL and tagged with the keyword ‘Complete proteome’. This will ensure that all alternative protein variants and isoforms are presented in the set and also enable the synchronization of the UniProtKB set with the CCDS project. Over time, these TrEMBL entries will be merged with the parent entry, as is UniProtKB/Swiss-Prot curation policy.

Complete proteome files will be available for human and mouse from UniProt release 2011_05 (3rd May) and for rat, zebrafish, chicken, cow, dog, *Arabidopsis thaliana* and pig along with *Caenorhabditis elegans, Drosophila melanogaster* and *Saccharomyces cerevisiae* from 2011_07 (28th June). Arabidopsis will be included following the release of TAIR10. In addition to this, UniProtKB/Swiss-Prot is currently performing a demerge of identical proteins produced by multiple genes within a single organism, currently grouped into a single entry, enabling the database to display a single unique entry per gene per species.

Toby Gibson (EMBL-Heidelberg) then challenged the current textbook view of cellular signalling as a linear series of events. Rather, the cell is now considered to be regulated by highly discrete yet dynamic protein complexes in which individual proteins can make a large number of often low-affinity interactions to assemble discrete signalling platforms [6]. These then integrate the multitude of incoming cell state signals. Kinases, which in isolation may be apparently promiscuous with regard to substrate, are regulated by intricate networks and through binding to scaffolding proteins gain substrate specificity. There is causality in the assembly of such signaling complexes with pre-complex assembly often needing to occur, driving a series of post-translational modification or conformational events. Work is currently ongoing to determine whether the current PSI-MI XML format is flexible enough to capture such a sequential series of biological events.

Chris Taylor (EMBL-EBI) described the work of MIBBI : Minimum Information for Biological and Biomedical Investigations which enables access to Minimum Information guidelines for diverse bioscience domains (www.mibbi.org). All MIAPE modules are registered with the portal, which is a simple hyper-linked list of all available MI guidelines. Work is now ongoing to develop the MIBBI foundry, which organizes the guidelines into an Omics workflow and allows user to pick those relevant for their particular set of experiments. The HUPO-PSI was encouraged to pro-actively contribute to this effort, such that their documents can be kept in line with related efforts by the MIBBI project.

The meeting then separated into the various workgroups:

Molecular Interactions (PSI-MI) Henning Hermjakob

The existing XML format (PSI-MI XML2.5) has been relatively stable since publication in 2007, with only minor point releases since that date. A review of its current limitations was undertaken during this meeting with proposed workarounds and solutions discussed. The ability of the format to describe cooperative interactions was also considered. A number of work items were added to the list of requirements that v3.0 of the format will be expected to address. Extensions to the MITAB format to form MITAB2.6 had been agreed in the previous year’s meeting, new columns enabling the description of molecule stoichiometry and annotated features such as binding domains were added this year to create MITAB2.7.

The launch of PSICQUIC 1.0 (http://code.google.com/p/psicquic/) has been an over-whelming success, with 16 databases making over 16 million interaction evidences available for search (April 2011). A review of existing services lead to the production of a preliminary set of ‘Data Distribution Best Practises’ guidelines, with further work in this area planned for beyond the workshop. Once agreed and implemented, these recommendations will make searching and clustering of data an easier experience for the user. It is planned to write a PSICQUIC Enricher tool which will perform much of the implementation work for the participating services. This will include recommended display names for each molecule type so that the user can see a consistent view of the data. Discussion then moved to the development of PSICQUIC 2, which will run from both MITAB 2.6/2.7 and PSI-ML XML. A number of ideas, which had been drafted during a mini-Hackathon at the EBI were presented – a full Hackathon is planned for later in 2011 to push the development of v2.0 forward. An update on the status of PSISCORE (http://code.google.com/p/psiscore/) was then given. The input to PSISCORE is a set of molecular interactions in a HUPO-PSI-defined file format (i.e. MITAB or PSI-MI XML 2.5). A PSISCORE client, PSISCOREweb, sends this file to multiple scoring servers. All the calculated scores are then added to the input file, which the user can then download again. A major branch was added to the PSI-MI controlled vocabularies to enable description of the various scoring systems, and a branch enabling tagging of the content of PSICQUIC records was also agreed upon.

Finally, Pascal Braun (Center for Cancer Systems Biology, Boston) lead a discussion on the identification of sets of known positive and randomly generated interaction pairs to act as positive and negative standard sets respectively, for the verification of binary experimental interaction data. These were originally identified in the public domain molecular interaction databases, and the positive set subsequently re-examined in a number of experimental systems. In collaboration with Javier de Las Rivas (Cancer Research Center, Salamanca) it is planned to continue this exercise to substantially increase the size of such sets. It was agreed that the HUPO-PSI would actively support this effort by placing links on the PSI-MI homepage, and potentially on other related websites, and making relevant experimental data available through the IMEx Consortium.

Mass Spectrometry (Eric Deutsch, Institute Systems Biology, Seattle)

The PSI Extended Fasta Format (PEFF) has been designed to provide a more richly annotated fasta file to search engines, in a standardised format, irrespective of sequence database from which is was generated. The format had previously been submitted to the PSI document process. Comments which had been received were reviewed during the meeting and the format will be resubmitted to the PSI Document Process once updates are complete. A Java PEFF viewer is being produced by Harald Barsnes (University of Bergen) (http://code.google.com/p/peff-viewer) which is capable of parsing a PEFF file and displaying the sequence entries annotated with the information parsed from the entry headers.

MIAPE-MS was reviewed and all components relating to quantitation were removed, these will constitute part of the MIAPE-QUANT document. Three example documents for the revised MIAPE-MS are currently being worked on and the document will be submitted to the Documentat Process on completion. In collaboration with the PSI-PI workgroup, version 0.1 of MIAPE-QUANT was completed during the meeting. This will outline the minimum requirements for all quantitation workflows, including selected reaction monitoring (SRM), within a single document. The ProteoRed group will produce appropriate example documents and MIAPE-QUANT will then be reviewed by the mzQuantML developers, and then, submission to the PSI document process will be considered. A new set of minimum requirements relating to sample preparation is also being prepared. “Protein extracts preparation from cell culture, tissues and fluids, Protein digestion (in gel and liquid) and phosphopeptide enrichment”.

The PSI-MS controlled vocabulary is now serving mzML, mzIdentML, mzQuantML, TraML, PRIDE XML and PEFF. This has resulted in a need to reorganise the entire CV to serve all these formats. The PRIDE group is leading a reorganisation effort during 2011.

mzML, the standard output format for mass spectrometer output data, has been stable for two years, and no significant changes are planned at this time, aside from general expansion of the controlled vocabulary to accommodate new instruments and techniques. A proposal to adopt mzML by the metabolomics community was discussed. Further alignment of controlled vocabulary terms between mzML and the non-PSI imzML format was discussed. Minor additions such as the addition of digital signatures to enable data authentication and integrity were discussed. An SQLite version of mzML was also proposed. A light-weight, tab-delimited file format for proteomics data, mzTab, has also been developed. In contrast to the PSI-standard formats mzIdentML and (in the future) mzQuantML, mzTab will only hold the summary results of proteomics experiments without any detailed evidence. As mzTab files are simple tab delimited text files they can be opened in MS Excel and enable users to easily share proteomics results with researchers outside the field.

During the workshop, some progress was made in finalizing TraML, a standardized format for the exchange and transmission of transition lists for SRM experiments, with issues relating to ion series and neutral loss combination being resolved. A Java API is currently under development (http://code.google.com/p/jtraml/) to provide a parser/writer library for working with TraML files.

Protein Separations (PSI-Sep) Juan-Pablo Albar (ProteoRed, National Center for Biotechnology-CSIC, Madrid)

The Proteins Separation group reviewed the work of Kenyani et al. [7] who described how gel-based protein identification data sets can now be deposited in the PRIDE database (www.ebi.ac.uk/pride), using a new software tool, the PRIDESpotMapper, developed to work in conjunction with the PRIDE Converter (ref?) application. The ProteoRed MIAPE generator tool [8] can be used to create and share a complete and compliant set of MIAPE reports for such experiments. For the remainder of the meeting, this group merged with other groups to work on mzQuantML and MIAPE-QUANT.

Proteomics Informatics (PSI-PI) Martin Eisenacher (Medizinisches Proteom-Center, Ruhr-Universitaet Bochum)

The PSI-PI group first selected a new co-chair (Martin Eisenacher), a new ontology coordinator (someone from the EBI, contact: Juan-Antonio Vizcaino) and a new editor (Gerhard Mayer).

The protein/peptide identifications interchange format, mzIdentML was discussed. Minor updates mzIdentML have resulted in a new version 1.1. This version now consists of only one .xsd schema file (merged mzIdentML and FUGElight schemas), and the group implemented minor changes due to upper/lower case writing of attributes and elements, and merged or moved some elements for simplicity or readability reasons. With appropriate CV terms, it now enables the annotation of types of protein identification ambiguity (conclusive/inconclusive/redundant). Gorka Pietro from the ProteoRed consortium presented a tool for annotating this information to mzIdentML files. The revised schema was reviewed and most open issues in the google code project (http://code.google.com/p/psi-pi/) were closed. The specification and example documents need to be updated before the format is submitted to the Document Process for 30 days public comments and review, as the update consists of minor changes, which are not backwards compatible to version 1.0

An alternative schema for mzQuantML with different and flexible layers for example for features, peptides and proteins (but also for ratios and statistical values) was discussed and has been deposited in a google code project (http://code.google.com/p/mzquantml/). Corresponding example documents for SILAC, iTRAQ, and LC/MS label-free quantitation have been assembled and discussed. The schema will presumably also be able to describe spectral counting, metabolite quantities and also SRM data in the future. It could also act as an intermediate format in quantification workflows. There was a recognized need to standardize data processing section in all PSI formats. Work was also begun on updating the PSI-MS CV with appropriate terms for quantitation, especially for the description of quantitation methods and format layers.

Immediately following the PSI meeting, on April 14–16, the kickoff meeting of the ‘ProteomeXchange’ consortium [5] took place in the same location. This is a coordination action project in the 7th Framework Programme of the European Commission, which started on January 2011 and will last for 3 years. The partners of the grant (including representatives of these databases) and some other representatives from scientific journals and the proteomics community as a whole, discussed about the best way to promote data sharing in proteomics and the practicalities to have the system implemented. The related IMEx consortium, which has been formed to share the curation and release of molecular interaction data, and is supported by the EC FP7 PSIMEx grant held its annual meeting in parallel.

Acknowledgements

Much of this work has been funded by the European Commission under PSIMEx, contract number FP7-HEALTH-2007-223411 and ProteomeXchange, grant number 260558. ME and CS are funded from P.U.R.E. (Protein Unit for Research in Europe), a project of Nordrhein-Westfalen, a federal state of Germany. EWD is also funded by NIH/NIGMS grant number R01 GM087221, NIH/NIGMS grant number PM50 GMO76547/Center for Systems Biology, and the Systems Biology initiative from the Grand Duchy of Luxembourg.

References

1. Liu G, Zhang J, Larsen B, Stark C, et al. ProHits: integrated software for mass spectrometry-based interaction proteomics. Nat Biotechnol. 2010; 28:1015–1017. [PubMed: 20944583]

2. Taylor CF, Paton NW, Lilley KS, Binz PA, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25:887–893. [PubMed: 17687369]

3. Martens L, Chambers M, Sturm M, Kessner D, et al. mzML--a community standard for mass spectrometry data. Mol Cell Proteomics. 2011; 10:r110.000133. [PubMed: 20716697]

4. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol. 2007; 25:894–898. [PubMed: 17687370]

5. Hermjakob H, Apweiler R. The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. Expert Rev Proteomics. 2006; 3:1–3. [PubMed: 16445344]

6. Gibson TJ. Cell regulation: determined to signal discrete cooperation. Trends Biochem Sci. 2009; 34:471–482. [PubMed: 19744855]

7. Kenyani J, Medina-Aunon JA, Martinez-Bartolomé S, Albar JP, Wastling JM, Jones AR. A DIGE study on the effects of salbutamol on the rat muscle proteome - an exemplar of best practice for data sharing in proteomics. BMC Res. Notes. 2011; 4:86. [PubMed: 21443781]

8. Martinez-Bartolome S, Medina-Aunon JA, Jones AR, Albar JP. Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports. Proteomics. 2010; 10:1256–1260. [PubMed: 20077409]