Beilstein-Institut ESCEC, September 23rd – 26th, 2007, Rüdesheim/Rhein, Germany

How to Develop a Standard – the HUPO-PSI Experience

Sandra Orchard

EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

E-Mail: orchard@ebi.ac.uk

Received: 2nd June 2008 / Published: 20th August 2008

Abstract

The HUPO Proteomics Standards initiative has designed and implemented common data reporting and exchange standards to enable the transfer of proteomics data from originator to collaborator to a final public repository immediately prior to publication. This work has been undertaken with extensive community involvement at every stage of the process to ensure that the end product fulfils the users' needs. The scientific community is already benefiting from this work, with XML formats to exchange and import data into databases, allowing direct access and comparability irrespective of the originating instrumentation. Public repositories allow researchers to access and search published experimental data with the result that reference datasets are becoming available for benchmarking purposes. Collaborations between databases are exposing these datasets to an ever increasing audience and enabling exciting new science to be derived from existing data.

Introduction

Since 2002, the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) have worked towards producing standards formats by which proteomics data can be collected, transferred from application to application, and finally submitted to a repository where it becomes available to the user community [1]. Within the field of proteomics, the ever-increasing amounts of data generated by mass spectrometers with ever-faster cycle

times, high sensitivity and high quality MS spectra required a move beyond the traditional route of printed publication. Whilst journal articles may still be the most practical method for disseminating the conclusions drawn from such experiments, they display only a small proportion of the actual results and almost none of the underlying data, such as peptides or spectra, from which this data was generated. When mass spectrometry data has been made available, for example via an author's website, it is generally only provided in vendor-specific formats. In most cases, the sheer size of the data files generated by a typical mass spectrometry experiment, result in raw data being erased once published, with only the final processed protein lists being retained and published.

To enable the community to effectively mine an increasingly rich data source, the experimental data needs to be collected in central repositories, in a single common format, from where it can be searched, and researched. As protein sequence databases, such as UniProtKB [2], improve in both coverage and sequence quality, high quality spectra produced at an earlier point in time can be rerun, and protein assignments made to previously orphan spectra, provided these spectra have been retained. To maximize the value which can be extracted from each experiment, the metadata needs to be both well annotated, and consistently annotated across different datasets. The aim of the work of the HUPO-PSI is to make it possible for labs across the world to run experiments and combine results, to distribute workloads, to enable the user to select the best tools for specific tasks rather than be forced to use only those which use a very specific format, and to derive new methodologies by comparing results. To this end a four-fold approach was taken, for each aspect of the potential workflow within a proteomics experiment:

1. Formal requirements specification with use cases submitted as many users as possible.
2. A checklist of minimum information with which each experiment should be annotated.
3. The development of an XML interchange format to enable data transfer which can be used by instrumentation, tools and repositories.
4. Controlled vocabularies to enable a rich annotation of the data within the schema.

Each of these have now been developed and released for gel electrophoresis, chromatographic techniques, mass spectrometry, protein identifications and molecular interactions.

Formal Requirements Specification – Standards as a Community effort

The best standard in the world will be of limited use if it is not adopted and employed by the majority of the community, it is not designed with ease of use and accessibility in mind and is not accompanied by a suite of tools which maximize its use. With that in mind, all

workgroups within the PSI have actively sought the input of as broad a group as possible, including hardware manufacturers, software developers, data producers, data repositories and bioinformaticians. Standards developments have centered around (bi-)annual workshops, which have been widely advertised and made freely available to all to attend. Development continued between workshops using all possible means of communications, including discussion groups, mailing lists and telephone conferences. All discussion documentation was made publicly available on the HUPO-PSI website (http://www.psidev.info/). Completed work enters a formal documentation process designed to ensure a good balance between expert design and public scrutiny [3]. Documents are first subjected to anonymous review followed by a set period of time in which the document is put forward for public comment. At all points, any feedback is responded to and, if appropriate, standards are updated and the document resubmitted into the review process. Any documents which are formally published are then additionally subject to standard journal review processes – *Nature Biotechnology* also established a community review page which provided further opportunity for potential users to comment on the standards prior to final publication. The PSI review process has also been used for the validation of documents not directly produced by the HUPO-PSI, for example the FuGE data model was reviewed through this proves prior to journal publication [4].

Minimum Information About a Proteomics Experiment (MIAPE)

Proteomics data should therefore ideally be accompanied by contextualizing metadata, making explicit both where samples came from and how analyses were performed. To that end, the Proteomics Standards Initiative develops guidance documents specifying the data and metadata that should be collected from various proteomics workflows, known collectively as the ”minimum information about a proteomics experiment” (MIAPE) guidelines [5]. These consist of a series of documents, linked by a single parent which makes explicit the scope, purpose and manner of use of the modular MIAPE guidelines that accompany it, and lay out the principles underlying module production. The first of these modules has already been published, the Molecular Interaction guidelines MIMIx [6], with those for gel electrophoreses and column chromatography to follow. Broadly speaking, MIAPE documents require that any description of a proteomics experiment should allow the user to understand, qualify and reproduce the work described in that paper. Each document provides a checklist of information and data to provide when an experiment is reported and acts as an aid to assessing quality control but does not attempt to tell a scientist how to run an experiment, or how to represent the final data and does not state how any quality judgment should be made. It is envisaged that these requirements will eventually be adopted and enforced by journals, repositories and funding agencies. Finally, these documents will ensure that the presentation of the information published will be in a style compatible with the HUPO-PSI data formats.

An increasing number of minimum requirements documents have now been produced, since the original, MIAME [7], produced by the micro-array community covering a broad range of biological disciplines. Such 'minimum information' checklists are usually developed independently, from within particular biologically- or technologically-delineated domains. Consequently, the full range of checklists can be difficult to find without intensive searching; they are also inevitably partially-redundant one against another, and where they overlap arbitrary decisions on wording and sub-structuring make integration difficult. This presents significant difficulties for the users of checklists; for example, in the area of systems biology, where data from multiple biological domains and technology platforms are routinely combined. A common portal to such MI checklists; to act as a 'one-stop shop' for those exploring the range of extant projects, foster collaborative development and ultimately promote gradual integration, MIBBI (http://mibbi.sourceforge.net/, http://www.mibbi.org/) has now been developed. All MIAPE modules are submitted to MIBBI, on publication.

XML Interchange Standards

To facilitate data management and exchange, the HUPO-PSI has developed data exchange formats for proteomics. For each work group/domain, these can minimally represent the data items specified in the MIAPE guidelines, but usually allow a much more detailed representation. Normally, the data exchange format is specified as a fully annotated XML schema. PSI schemas are developed to facilitate data exchange between databases as well as databases and end users. They explicitly do not propose any internal data representation for databases or tools. While XML is inherently verbose, standard compression algorithms typically reduce the file size by 50 – 90% of the original, and such compression normally is not the limiting factor on modern computer systems. On the plus side, XML is well supported by standard mechanisms for querying, native XML databases, and automated mappings to both relational databases and object models.

The molecular interaction interchange format, PSI-MI XML 2.5, is now used by all interaction databases and an increasing number of graphical visualisation and analysis tools. The mass spectrometry standard, mzData, was released in 2004 and was well accepted by many users. The format allowed the storage of proteomic-related mass spectral data, ranging from basic details about the sample, instrument details and data processing steps, through to the actual spectral lists of mass-to-charge values and intensities, using base64 encoding to represent the floating point mass-to-charge (m/z) and ion intensity. Following some refinement in response to their feedback, the format was rapidly implemented by several of these manufacturers and files containing spectral data were soon being generated by several large groups, most notably the HUPO tissue initiatives. A standards compliant repository, PRIDE (http://www.ebi.ac.uk/pride) was also established. However, in 2004 a second open, generic XML representation of MS data, was published by the Institute of Systems Biology, mzXML [8]. Whilst this was originally designed to be work-flow specific, other workers began to find wider uses for the format with the result that manufacturers were faced with

the prospect of having to implement two separate open-source formats. Rather than loose the initial good-will with which the manufacturers had entered this project, and to avoid user confusion, in 2006 the two groups decided to merge the two formats into a single, and much improved, XML schema was released in June 2008, mzML with full vendor support was a critical part of the design process and many open-source implementations.

Accompanying an XML representation of MS data, there is a need for a corresponding representation of the peptide and protein identifications made in any experiment to capture results from MS search engines and represent the input parameters for analysis algorithms, thus unifying results from different search engines. The development of AnalysisXML has proven far from straightforward, partly because the scope of the project has changed often in a fast moving field however Version 1.0 will be submitted to the PSI documentation process summer 2008. Quantitation will not be addressed until version 2.0, however version 1.0 documents will be backwards compatible with the 2.0 schema.

GelML, designed for the interchange of protocols and image data from 1D- and 2D-gel electrophoresis, completed the PSI document process and was released as a stable version 1.0 late in 2007. The interchange format for non-gel based separations, spML, is in development.

Controlled Vocabularies

While XML schemas provide a syntax for data exchange, they do not specify the semantics of data elements exchanged. As an example, the yeast two-hybrid technology might be designated by many different terms, most of which are sufficiently distinct to make automatic recognition impossible. Thus, the PSI either references external controlled vocabularies (CVs) or ontologies such as the NCBI taxonomy where possible, or develops its own controlled vocabulary, for example for protein interaction detection technologies, where necessary. The combination of reasonably stable XML schemas and regularly maintained controlled vocabularies has proven to allow quick adaptation to new terms and technologies, while providing the stability required for database and software development. All PSI CVs are written in OBO format and maintained at the Open Biomedical Ontologies website (http://www.obofoundry.org) and can be viewed using the Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup/"\t"_parent) [9].

Interaction with the Publishing Community

The need for underlying experimental data to be available to the reader has long been recognised by the scientific publishing community, and this has been addressed in many ways – from the mandated submission of nucleotide and protein sequences to public domain repositories, with accession numbers then being given in the article, to the provision of Supplementary Material held by the journal itself. As the volume of data increases, the

journals will become more reliant on external repositories to collect and manage this data and, in recognition of this, are actively supporting the implementation of the required standards needed to enable deposition. Several journals have already published their own guidelines as to how proteomics data should be represented and these requirements have been incorporated in the MIAPE guidelines [5]. Increasingly, journals are starting to request data deposition: *Proteomics*, and *Nature Biotechnology* [10] and *Nature Methods* [11] have recently started to request that authors deposit proteomics and interaction data in HUPO-PSI standards-compliant databases prior to publication and it is anticipated that this trend will increase as both standards and databases mature.

Summary

The HUPO-PSI have produced a range of standards allowing the user to describe all aspects of a proteomics experiment, to exchange and compare data across collaborating groups, or use established datasets as standards, and to deposit the final results into standards-compliant repositories. All of these have been written in full consultation with an extensive user community to ensure as much by-in as possible. Where an existing standard, be it an interchange format or CV the PSI have strived to work with that facility rather than replace it or produce a redundant application. Cross-community efforts are actively supported to ensure that users working in multiple ''omics'', for example analysing the same sample by both transcriptomic and proteomic techniques can find the tools to assist in the data handling and resorting processes. The emphasis now is on developing the tools to make these standards usable and accessible to the bench scientist and to encourage the direct deposition of data into public domain repositories.

References

[1] Orchard, S., Hermjakob, H. (2008) The HUPO proteomics standards initiative–easing communication and minimizing data loss in a changing world. *Brief. Bioinformatics* **9**:166 – 73.

[2] The UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **36**:D 190 – 195..

[3] Vizcaíno, J.A., Martens, L., Hermjakob, H., Julian, R.K., Paton, N.W. (2008) The PSI formal document process and its implementation on the PSI website. *Proteomics* **7**:2355 – 2357.

[4] Jones, A.R., Pizarro, A, Spellman, P., Miller, M., FuGE Working Group (2006) http://www.ebi.ac.uk/citexplore/citationDetails.do?externalId = 16901224&dataSource = MED FuGE: Functional Genomics Experiment Object Model. *OMICS* **10**:179 – 184.

[5] Taylor, C.F., Paton, N.W., Lilley,K.S., Binz, P.-A., Randall, J.R.K., Jr, Jones A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., Dunn, M.J., Heck, A.J.R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T.A., Patterson S.D., Ping, P., Seymour, S.L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T.M., Whitelegge, J.P., Wilkins, M.R., Xenarios, I., Yates III J.R., Hermjakob, H. (2007) The Minimum Information About a Proteomics Experiment (MIAPE). *Nat. Biotechnol*. **25**(8):887 – 893.

[6] Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A, Chatraryamontri, A, Armstrong, J., Woollard, P., Salama, J.J., Moore, S., Wojcik, J., Bader, G.D., Vidal, M., Cusick, M.E., Gerstein, M, Gavin, A.-C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V.M., Hogue, C, Mewes, H.-W., Apweiler, R., Xenarios, I., Eisenberg, D. Cesareni, G., Hermjakob, H. (2007) The Minimum Information required for reporting a Molecular Interaction Experiment (MIMIx). *Nat. Biotechnol*. **25**(8):894 – 898.

[7] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C.P., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. **29**:365 – 371.

[8] Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W., Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol*. **22**(11):1459 – 1466.

[9] Côté, R.G., Jones, P., Martens, L., Apweiler, R., Hermjakob H (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res*. In press.

[10] Nature Biotechnology (2007) Editorial Time for leadership. *Nat. Biotechnol* **25**(8):821.

[11] Doerr, A. (2007) Standardizing proteomics. *Nat. Methods* **4**:774.