

Standardising Proteomics Data – the Work of the HUPO
Proteomics Standards Initiative

Sandra Orchard* and Henning Hermjakob

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

*Corresponding author: Tel: +44 (0)1223 494 675; Fax: +44 (0)1223 494 468; Email: orchard@ebi.ac.uk

Received April 03, 2008; Accepted April 09, 2008; Published April 20, 2008

Citation: Sandra O, Henning H (2008) Standardising Proteomics Data – the Work of the HUPO Proteomics Standards Initiative. *J Proteomics Bioinform* 1: 003-005.

Copyright: © 2008 Sandra O, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

The HUPO Proteomics Standards initiative has dedicated the last 6 years to the design and implementation of common data reporting and exchange standards enabling the transfer of proteomics data from originator to collaborator to a final public repository immediately prior to publication. The user community is now 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 now allow researchers to view and search published experimental data and 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

Five years ago, the nascent field of proteomics was in a state of relative chaos, with ever growing amounts of data being generated from increasingly high throughput machines but no downstream repositories in place to capture this information. The only option available to the researcher was to publish the cream of the results in a journal article, with protein lists largely available only in Supplemental Information and no access for the reader to the raw data from which these lists were generated. Any data not included in this article was then lost, and researchers were unable to track individual proteins across these datasets to obtain a picture of their expression profile. To add to this confusion, raw data generated by different mass spectrometers and by each peptide search engine was available only in the manufacturer's proprietary format and comparison of such data was impossible, even when generated within the same laboratory but by using diverse machines.

The amount of data which has effectively been lost during this period is incalculable. As the breadth and quality of protein sequence databases such as UniProtKB improves (The UniProt Consortium; 2008), previously unidentified spectra could now be assigned as novel proteins if only the original data were still available. Improved algorithms and the recognition of protein-specific signature peptides may now allow what was previously regarded as poor quality data to be seen as a true protein identification, if mass spectra produced some years ago were to be rerun in the current implementations of the search engines. The deposition of raw data into public domain repositories will not only have allowed such exercises to be undertaken but also allow comparison of datasets generated by different research groups – for example protein sets from diseased tissue against reference sets generated in normal, healthy tissue or the potential contaminating ef-

fects of the known plasma proteome accounted for when examining the proteome of tissues such as the heart or liver.

HUPO and the Proteomics Standards Initiative

The Human Proteome Organisation (HUPO) was formed in 2001 to consolidate national and regional proteome organisations into a single worldwide body. The Proteome Standards Initiative (PSI) was established by HUPO with the remit of standardising data representation within the field of proteomics to the end that public domain databases can be established where all such data can be deposited, exchanged between such databases or downloaded and utilised by laboratory workers. The Proteomics Standards Initiative (HUPO-PSI) have concentrated on bringing data standardization and common data reporting standards to a limited number of fields within the global umbrella of Proteomics; to date, protein/peptide separations, mass spectrometry and molecular interactions.

For each of these areas Minimum Information About a Proteomics Experiment (MIAPE) documents have been developed, analogous to the MIAME (Minimum Information About a Microarray Experiment) guidelines (Brazma; 2001) for DNA microarray experiments, to define those data items that should minimally be reported about a proteomics experiment to allow critical assessment of the experiment. This is a simple textual representation, independent of any formal data format. MIAPE guidelines consist of a general “parent document” (Taylor et al; 2007) and workgroup-specific modules, the first of which has recently been published (Orchard et al; 2007a). These guidelines summarise what could be considered “common sense” and what the community has agreed should be present in each and every paper, but is all too often not appropriately reported in publications – the precise identification of a protein entity and the species from which it originated being a

simple example of data often missing from articles. To facilitate data management and exchange, each domain area has also developed data exchange formats for which can minimally represent the data items specified in the MIAPE guidelines, but usually additionally allow a much more detailed representation. Normally, the data exchange format is specified as a fully annotated XML schema. HUPO-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. As XML is inherently verbose, standard compression algorithms are used to reduce the file size by 50-90% of the original, and such compression typically 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 semantics of data elements exchanged are described by a series of controlled vocabularies, either by referencing external resources such as the NCBI taxonomy or developed internally, for example to describe the details of mass spectrometry or molecular interactions. The combination of reasonably stable XML schemas and regularly maintained controlled vocabularies allows a quick adaptation to new terms and technologies, while providing the stability required for database and software development.

Progress in developing Data Interchange Standards

Molecular Interactions

The Molecular Interaction workgroup lead the field in publishing and implementing these standards, with Level 1.0 of the XML schema, and accompanying controlled vocabularies, published in 2004 (Hermjakob et al; 2004) and Level 2.5 currently the stable implementation, described in 2007 (Kerrien et al; 2007). All major interaction databases now make a data download available in this format and it has also been used by high throughput data generators for data deposition. In response to a request from the biological community, data is also made available in a simpler tab-delimited format (MITAB2.5), making the information suitable for immediate upload into an Excel worksheet. The adoption of this format, and the communication between several interaction databases during its development, resulted in the formation of the IMEx consortium, with several databases now agreeing to regularly exchange data and act as common portals for data deposition (Orchard et al; 2007b).

Mass Spectrometry

The mass spectrometry workgroup published the mzData XML format in 2004, which 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 (Orchard et al; 2004). The format was rapidly implemented by several manufacturers of both mass spectrometers and search engines and files containing spectral data were soon being generated by several large groups, most notably the HUPO tissue initiatives.

The PRIDE 'PRoteomics IDentifications database' (<http://www.ebi.ac.uk/pride>) was the first PSI-standards compliant relational database to be implemented (Jones et al; 2008). PRIDE holds the protein and peptide identifications from an experiment along with accompanying details of the experimental protocol, mass spectrometry run conditions and search engine algorithm(s) used and the spectra from which these identifications were made. Data from experiments performed in a wide range of organisms,

tissues, cells and disease-states are already held in the database, including the reference datasets from the HUPO plasma, liver and brain tissue initiatives.

However, in 2004 a second, generic XML representation of MS data, was published by the Institute of Systems Biology, mzXML (Pedrioli et al; 2004). 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. To avoid user confusion, in 2006 the two groups decided to merge the two formats into a single, and much improved, XML schema and by 2007 this work had matured into the current mzML, currently available in beta format (www.psidev.info/index.php?q=node/80) and due to be formally released Spring 2008. Databases such as PRIDE already using the mzData or mzXML formats have committed to upgrading to the new format as soon as possible after this release.

Separation Techniques

Methods for the separation of complex protein mixtures prior to identification of the individual components by mass spectrometry are an integral part of any proteomics workflow, and will influence the spectra of proteins identified in any particular experiment. The activities of the HUPO-PSI include the field of 1- and 2-D gel electrophoresis, column chromatography and a generic model has been developed to describe other separation techniques. An XML interchange format (GelML 1.0) with accompanying controlled vocabularies has been developed to enable exchange of this data and was released as a stable version 1.0 the end of 2007 (www.psidev.info/index.php?q=node/83). Work is now progressing in the areas of column chromatography and capillary electrophoresis. The model for gas chromatography will be completed in conjunction with the Metabolomics Standards Initiative (<http://msi-workgroups.sourceforge.net/>) since this is a technique used by both communities.

The ultimate aim of the PSI is to make more proteomics data easily accessible in the public domain. Deposition of data in the public domain is now becoming more common practise, with an increasing number of journals now requesting this as part of the manuscript submission process. Previously deposited datasets, such as that published by the HUPO Plasma Proteome Project (Omenn et al; 2005), are now regularly cited in the literature as producers of new data compare their work to the growing body of previously published experiments, both to confirm their observations and to highlight novel observations and subsequent reanalyses of datasets are starting to appear in the literature increasing author visibility (Klie et al; 2008). Manufacturers are increasingly offering the ability to download data in HUPO-PSI XML formats from their instrumentation, and tools enabling data validation and aiding deposition are being released. The community is already benefiting from these common formats and this can only increase as their acceptance and implementation becomes more widespread.

References

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC P, Kim IF, Markowitz V, Matese JC, Parkinson H, Alan RA, 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.
2. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J,

- Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SGN, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vida M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol* 22: 177-183.
3. Jones P, Côté RG, Cho SY, Klie S, Martens L, Quinn AF, Thorneycroft D, Hermjakob H (2008) PRIDE: new developments and new datasets. *Nucleic Acids Res* 36: 878-83.
 4. Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-aryamontri A, Oesterheld M, Stümpflen V, Salwinski L, Nerothin J, Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H (2007) Broadening the Horizon - Level 2.5 of the HUPO-PSI Format for Molecular Interactions. *BMC Biol* 5: 44.
 5. Klie S, Martens L, Vizcaíno JA, Côté R, Jones P, Apweiler R, Hinneburg A, Hermjakob H (2008) Analyzing large-scale proteomics projects with latent semantic indexing. *J Proteome Res* 7: 182-191.
 6. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H, Paik YK, Yoo JS, Ping P, Pounds J, Adkins J, Qian X, Wang R, Wasinger V, Yue Wu C, Zhao X, Zeng R, Archakov A, Tsugita A, Beer I, Pandey A, Pisano M, Andrews P, Tammen H, Speicher DW, Hanash SM (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 5: 3226-3245.
 7. Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK Jr, Apweiler R (2005) Second proteomics standards initiative spring workshop. *Expert Rev Proteomics* 2: 287-289.
 8. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stümpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, 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 VM, Hogue C, Mewes HW, 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: 894-898.
 9. Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Nerothin J, Hermjakob H (2007) Submit Your Interaction Data the IMEx Way: a Step by Step Guide to Trouble-free Deposition. *Proteomics* 7 Suppl 1: 28-34.
 10. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R (2004) *Nat Biotechnol* 22: 1459-1466.
 11. Taylor CF, Paton NW, Lilley KS, Binz PA, Randall KJ, RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJR, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates III JR, Hermjakob H (2007) The Minimum Information About a Proteomics Experiment (MIAPE). *Nat Biotechnol* 25: 887-893.
 12. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 36: 190-195.