Publication Committee Meeting

REPORT

HUPO 5th Annual World Congress
Long Beach, CA, USA
30 October 2006

Sandra Orchard[1], Albert Heck[2], Mathias Uhlen[3] and Peipei Ping[4]

[1] EMBL Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK
[2] Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
[3] Department of Biotechnology, AlbaNova University Center, Royal Institute of Technology, Stockholm, Sweden
[4] Department of Physiology and Medicine/Cardiology, UCLA, Los Angeles, CA, USA

This meeting brought together delegates from industry, academia and the publishing houses to facilitate discussions on the level of support from the journals for the use of standardised data formats and their interest in the creation of a network of proteomics repositories collaborating on a coordinated data curation effort. Discussions centred on how best to structure interactions between journals, databases and researchers to improve accessibility to data, and facilitate comparisons between datasets.

Received: January 5, 2007
Revised: January 19, 2007
Accepted: January 22, 2007

Keywords:
Data standardization / Human Proteome Organisation / Proteomics Standards Initiative

Welcome and introduction

Attendees were welcomed by Dr Rolf Apweiler (EMBL-EBI), President-elect of the Human Proteome Organisation (HUPO), who explained that this session was a follow-on to the highly successful meeting held at the HUPO Protein Standards Initiative (PSI) Spring meeting in San Francisco [1], at which the editors of a number of key journals, instrumentation manufacturers, software producers and also the developers and distributors of database repositories had discussed the role that HUPO can play in improving the standard of proteomics data both submitted to and published by the journals. At this meeting it was felt that there was a need for a single standard in each field to be clearly defined, and for a commitment from the increasing number of data repositories that these would be implemented, before journals would start to request their use by data submitters. The aim of the current meeting was to ascertain how much progress had been made over the summer in response to these comments, and to determine whether the journals felt that sufficient advance had been made towards full implementation.

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; **MIMIx**, Minimum information required for reporting a molecular interaction experiment

The HUPO-PSI

The work of the HUPO-PSI over the last four years, in developing data standards for a number of key areas in the field of proteomics, was summarised by Sandra Orchard (EMBL-EBI). Each standard consists of four documents – a formal requirements specification, a minimal reporting requirements (MIAPE) domain document, an XML data exchange format and a domain-specific controlled vocabulary. Extensive community input was sought at every stage in the development of these standards and the PSI have formulated an exhaustive publication and review process which exposes the documents to public comment for increasing periods of time, as they reach the publication stage. In addition to the MIAPE parent document, a further ten domain specific standards were at varying stages in this process: study design and sample generation, separations and sample

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

handling, column chromatography, mass spectrometry, mass spectrometry informatics, capillary electrophoresis, gel electrophoresis, gel electrophoresis informatics, molecular interactions and statistical analysis of data.

The most advanced of these is the molecular interaction standard, with the XML interchange format published originally published in 2004 [Kerrien, S. *et al.* in preparation] and the MIMIx minimal reporting requirements paper now accepted for publication [2]. All major public domain molecular interaction repositories already export data in this format and MIMIx-compliant data submission tools are publicly available (http://imex.sf.net). The publication of the mass spectrometry standard has proved less straightforward in that, concurrent with the development of the PSI-MS interchange standard mzData, a second standard was published by the Institute of Systems Biology, mzXML [3]. As it became obvious that the existence of dual standards was restricting the adoption of either one of them by both the user community and the manufacturers, an agreement has been reached to merge the two into a single, global standard and the work to achieve this is well advanced [4]. Data repositories are already in existence, which support one or both of the existing standards, and many of these have already committed to updating their schema as soon as the merged interchange format becomes available.

The proteomics standards have completed the initial round of development and are now entering the publication and implementation phase. Data repositories exist and are committed to collecting interaction, mass spectrometry and gel electrophoresis data, if the data producers can be brought to a point where deposition of such data becomes routine. It was seen that this was where the journals and funding agencies could play a major role, in encouraging such submission prior to publication, or as a requirement of grant applications, to ensure that this data were available in the public domain for accessing, downloading and reanalysing by the entire proteomics community.

Mike Dunn (Proteomics, Wiley-VCH) then spoke as the representative of the publishing houses. He first summarised the support already given by PROTEOMICS to the work of the HUPO-PSI, in publishing meeting reports, editorials and tutorials and pledged that this would continue in the future. In addition to this, the journal has dedicated entire issues to the work of the HUPO tissue initiatives, all of which are depositing their data in public repositories, such as PRIDE [5]. He supported the concept of data deposition, although it was beyond the remit of the publishing houses to supply such repositories which would have to be housed by third party organisations such as the European Bioinformatics Institute. PROTEOMICS is an active partner in the ProDaC project, a European initiative aimed at the standardised data collection and analysis of protein identification by mass spectrometry and the development of an international data exchange network of proteomics data repositories.

Peter Hare (Nature Publishing Group) then followed, describing the work undertaken by the Nature journals in support of data standardisation efforts. A number of proteomics standards papers have been submitted for publication in Nature Biotechnology and are at various stages of the journal review process. Once through this process, NBT intend to publicise each paper for community consultation on its website – any input from this process will then be considered for inclusion in the final document by the authors and editors prior to final publication of the completed article. The publication which is the most advanced through this process is the molecular interaction standard, MIMIx, which is already available for comment on the NBT website (http://www.nature.com/nbt/consult/index.html). He stated that it is the Nature policy to support such standardisation efforts once the tools and repositories are in place to make deposition prior to publication fully practicable.

Finally Robert Barkovich (Thermo Electron) spoke on behalf of the software and hardware developers who will need to implement these formats into their products for their adoption to become widespread. He talked briefly about experiences so far in the use of HUPO-PSI mzData implementation and praised both the HUPO-PSI and ISB for their determination to merge the current two standards, which will encourage the widespread adoption of the final product.

There then followed a short panel discussion, chaired by Rolf Apweiler and Mathias Uhlen (KTH, Sweden). A number of people, including Ralph Bradshaw (J. Proteomic Research) expressed concern, not about the standards themselves, but about the burden on small laboratories trying to adopt their use without the bioinformatics support enjoyed by the larger institutions. It was agreed that tool development must now be a priority, and indeed a number of these already exist, with both a MIMIx-compliant Excel spreadsheet and web based deposition tools already available, and a standards-compliant Excel submission tool under test for the PRIDE database. It was also pointed out that a list of manufacturers and instruments or software which have already implemented the formats is not easy to access – this information is obtainable *via* the HUPO–PSI website but needs to be made more readily available.

Gil Omenn (U. Michigan, USA) spoke about the positive benefits experienced by the HUPO plasma proteomics project both in utilising these standards as they were being developed and in having the data subsequently available for use by the community. However, he admitted that persuading people to submit results in an XML format had proved difficult. He suggested that education and training were the way forward, and that HUPO was an appropriate body to encourage this through the pre-congress educational sessions organised at both a global and more local level.

Rolf Apweiler summed up the spirit of the meeting. The proteomics standards have been four years in development with prolonged consultation and testing periods. The initial versions of these standards have now reached the publication stage, with further community consultation being organised by the journals. The next phase is hoped to be their wide-

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

spread implementation and this will require investment in tool development, both by the open-source user groups and by the manufacturers. Data repositories must also adopt these standards and work together to share and exchange data, thus making it as easy as possible for the user to access and download the maximum amount of data from a single source. The journals and funding agencies have a key role to play, in that without their persuasion only a small percentage of the data producers will deposit their experimental results in the public domain. Their positive cooperation in this process, to date, gave every encouragement for future success.

References

[1] Orchard, S. and Ping, P., *Proteomics* 2006, *6*, 4436–4438.

[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. Biotechnol.* 2004, *22*, 1459–1466.

[4] Orchard, S., Taylor, C., Jones, P., Montechi-Palazzo, L. *et al.*, *Proteomics* 2007, *7*, 337–339

[5] Jones, P., Cote, R. G., Martens, L., Quinn, A. F. *et al.*, *Nucleic Acids Res.* 2006, *34*, D659–D663.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim