A Collaborative Approach to Developing Proteomics Data Standards

Kevin Jones[1], James DeGreef[1], Kent Laursen[1], Rob Milne[1], Chris Taylor[2], Angel Pizarro[3], Ruedi Aebersold[4], Peter Hussey[5], Mark Igra[5], Andy Jones[6], Erik Nilsson[7]
[1]GenoLogics Life Sciences Software Inc., Victoria, BC, Canada; [2]HUPO Proteomics Standards Initiative (PSI);
[3]University of Pennsylvania – Institute for Translational Medicine and Therapeutics, PA, USA; [4]Institute for Systems Biology, Seattle, WA, USA;
[5]LabKey Software LLC, Seattle, WA, USA; [6]School of Computer Science, University of Manchester, UK; [7]Insilicos LLC, Seattle, WA, USA

Objective

Functional genomics and proteomics projects increasingly involve collaborators with diverse lab environments, experimental workflows, instrumentation, software and data types, necessitating frequent exchanges of data that may not be compatible within a set of standards. A challenge faced by researchers is that the exchange of incompatible data will lead to serious issues affecting results interpretation and reproducibility. Overcoming this challenge involves the creation of a common data standard for proteomics, with broad acceptance across "-omics" disciplines, through the input and support of a consortia of researchers, thought leaders, and commercial concerns.

This group aims to guide and influence the broad adoption of the data standards across academic and commercial spheres, through the creation of a real-life test case of a proteomic workflow, utilizing emerging FuGE-OM (Functional Genomics Experiment Object Model) and HUPO Proteomics Standards Initiative (PSI) data standards. The poster provides an overview of the data standards work initiated between several parties, our current work, and the plans towards showcasing an example of a set of proteomics data standards at work in a large scale clinical proteomics setting, in a biomarker discovery context.

Contributors and participants

in the development of the proteomics data standards include:

The HUPO PSI. **Chris Taylor** is coordinating the data standard initiative from the PSI perspective, and is working on the overlap and dependencies with the FuGE-OM. This standards initiative provides a supporting structure for PSI's proteomics-focused position.

Angel Pizarro (**UPenn**) is coordinating the development of FuGE-OM, a flexible data standard framework that has incorporated input from other bodies, to create a general standard in which to encode data that will enable a systems biology approach to data analysis. FuGE allows the extension of other technology datatypes - genomics, transcriptomics, and metabolomics - ensuring that emerging proteomics standards will integrate with these other "-omics" disciplines in the long term.

Angel Pizarro, **Andy Jones** (**University of Manchester**), and **Kent Laursen** and **James DeGreef** (**GenoLogics**) are leading the efforts to build the HUPO PSI-compliant extensions to the FuGE, interacting closely with Chris Taylor and other HUPO PSI members. This group, along other FuGE developers, and **LabKey Software**, are collaborating to spur on this effort.

The **Institute for Systems Biology** (**ISB**) contributes a set of freely-available open source tools for the bioinformatics pipeline, capable of the identification of peptides, the validation of peptide and protein identifications using robust statistical models, organizing data to a relational database, and the interpretation of proteomic data in the context of functional modules and biological pathways. These tools comprise a computational pipeline that migrates raw, common format data towards biological inference.

Insilicos, through the contribution of the Insilicos Viewer, provides a means to read mass spectrometer proteomics data in the mzXML and mzData formats. Insilicos has collaborated with the ISB to extend the set of standards-based supported mass spectrometers, search tools, and databases.

The **Fred Hutchinson Cancer Research Center** (**FHCRC**) and FHCRC Computational Proteomics Laboratory, working closely with LabKey Software, have developed the 100 terabyte Comparative Proteomics Analysis System (CPAS) data repository which will import a complete proteomics experiment archive (XAR) data file based on FuGE with proteomics extensions.

GenoLogics Life Sciences Software is supporting the standards initiative, recognizing that standardization is critically important for lab customers. Kent Laursen is focusing on standards implementation into current and future versions of ProteusLIMS™ .

The goal is to showcase a working example of proteomics data standards in use, with an integrated comprehensive suite of software tools, a lab information management system (ProteusLIMS™), and a proteomics data repository (CPAS), highlighting data integrity and efficiency of analysis. Showcase will take place in October at the Fred Hutchinson Cancer Research Center. Throughout the summer of 2005, weekly conferences and discussions are being held between the FuGE group and HUPO PSI representatives, to reach a common ground on standards implementation.

For more information, see:
http://fuge.sourceforge.net/
http://psidev.sourceforge.net/
http://www.systemsbiology.org/
http://tools.proteomecenter.org/software.php
http://www.genologics.com/

Some of the relevant data standards are expressed in the following formats:

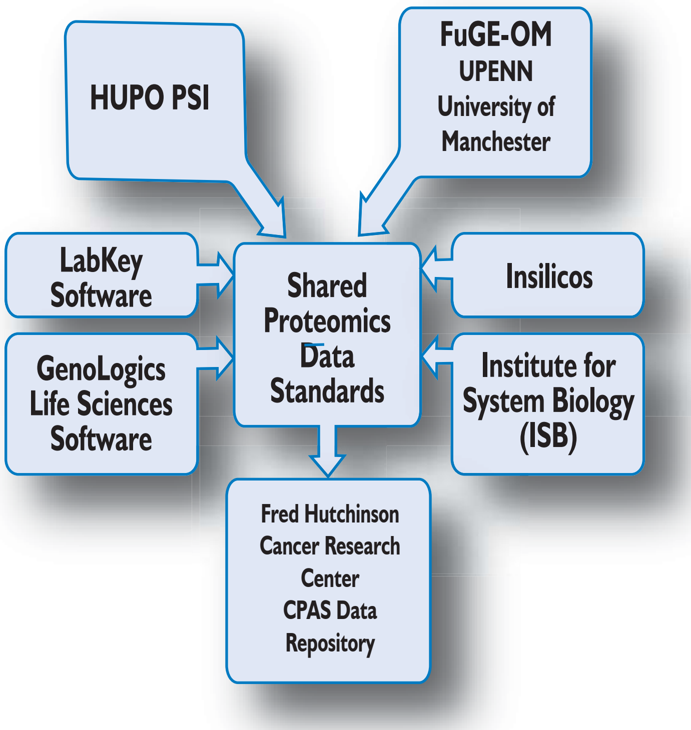
- **MIAPE** (the Minimum Information About a Proteomics Experiment) – This is a HUPO PSI set of reporting criteria containing just enough information to assess the relevance of a set of methods, results and conclusions, for the purpose of reporting consistency. The MIAPE is a subset of the total information available from a proteomics experiment.
- **Experiment XML** (expXML, or FuGE XML),
- **XAR** (eXtensible ARchiver file format). The XAR file format provides an interchange format for exchanging life sciences experiment data and protocols, and is based on concepts derived in the FuGE model.
- **mzXML**, **pepXML**, and **protXML**: early open data formats, broadly supported by ISB, FHCRC and others
- **mzData**, **AnalysisXML** (formerly mzIdent) are HUPO PSI standards; early support is forming behind mass spec vendors, and within FuGE.
- **LSID** (Life Science Identifier): are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources, including (but not limited to) individual genes or proteins, and the data objects that encode information about them.

Conclusion

The benefits associated with the widespread adoption of proteomics data standards are significant, and include:

- Better integration between disparate equipment within individual laboratories, and research and development methods that encourage sharing of results across disperse geographies and organizations, breaking down organizational silos,
- Better data quality, accuracy and speed associated with the transfer of data between laboratories,
- Faster and more accurate biomarker identification,
- Greater responsiveness in adhering to safety standards in bringing new drugs to market, and
- Increased interaction between the discovery, development and clinical trials phases in drug development, moving from sequential to parallel processes.

Figure: Proteomic Data Standards Initiative Participants


HUPO PSI
FuGE-OM
UPENN
University of Manchester
LabKey Software
Shared Proteomics Data Standards
Insilicos
GenoLogics Life Sciences Software
Institute for System Biology (ISB)
Fred Hutchinson Cancer Research Center
CPAS Data Repository