Joint Meeting of the HUPO "Proteomics Standards Initiative" (PSI) and the ASMS "Computer Applications Interest Group" (CAIG) On A Data Interchange Standard for Proteomics Data

Meeting Facilitators:
Randy Julian (Eli Lilly, ASTM E13.15, ASMS CAIG) and
Weimin Zhu, (EBI, HUPO-PSI)

June 11, 2003, Conference of the ASMS, Montreal, Canada

Introduction

The purpose of this meeting was to establish the scope and essential components for a draft standard for exchanging experimental data used to exchange, store and process data from proteomics experiments involving mass spectrometry. The objective was to focus on representing low-level proteomics data (peak lists and data processing meta data). A second objective was to leverage work being done on a new general analytical information standard by the ASTM E13.15 subcommittee. Some proposed elements of the new ASTM standard are very similar to proposals for proteomics data representations and representations developed by several research groups for internal use. Perhaps the most common feature is the selection of XML as the basic representation mode. Several XML formats for analytical data exist and it is the goal of the ASTM to create a general purpose Analytical Information Markup Language (AniML). The ASTM standard is intended to be comprehensive and thus will require time to fully define.

Since some key decisions about the new ASTM standard have been made, it is possible to begin work on a new proteomics standard which will be aligned with the ASTM standard. However, due to the narrow scope of a proteomics standard, a draft version should be complete much earlier than the complete ASTM standard.

Overview of the ASTM AnIML E13.15 Standard Design

To achieve a scope which covers all analytical information, the ASTM team has developed a layered approach. At the center of the model is a "core" data set which represents the multidimensional numerical data associated with an experiment. Decisions have been made on the nature of the core, including the representation of data as a collection of arrays of W3C XML-Schema data types. This allows multidimensional data to be represented either before or after signal processing. The goal of the core is to define the minimum common elements between all analytical measurements. Here, controlled vocabularies for units and variable types will be defined.

The next layer in the model has been named the "application" layer. Here, information generic to a technique or application will be stored. In the proteomics context, this layer would be populated by data elements from a controlled vocabulary established by experts in the field. It is proposed that the ASMS CAIG in collaboration with the HUPO-PSI could establish the proteomics application-specific layer.

Other layers beyond the application-specific information would include a vendor layer, where vendor-specific information could be stored for a given experiment. A user-layer allows individual users to add information specific to their experiment and needs. A final layer consisting of enterprise-level information could be specified by an organization to facilitate data management. More information on the ASTM E13.15 AnIML standard can be found at the project web site (animl.sourceforge.net).

The Goals of the HUPO-PSI Data Standard

The volume of data created in a typical proteomics experiment now easy exceeds what can be included in a publication or easily shared between laboratories. Differences in instrumentation, in procedures and methods and degree of data pre-processing all lead to difficulty in creating a global initiative to map the human proteome. A standard for representing proteome data is essential for success, the realization of which led to the creation of the Proteomics Standards Initiative (PSI).

The PSI has set an objective of creating a draft standard that can be presented at the HUPO congress in October 2003. To be successful in such a short time frame it was deemed essential to obtain broad participation from vendors, software developers, and standards organizations. Existing efforts to standardize analytical data formats were included in the HUPO-PSI plan from the start.

The goal is for researchers to be able to generate data in a standard format directly from their instrumentation. This data, written by the vendor data system should be usable by third-party software tools such as spectral databases, search engines and other computational tools. Since most of the major instrumentation vendors are involved in creating the ASTM standard, alignment with this effort should minimize the work needed by vendors to meet the standard. The HUPO-PSI standard will also need to include other information far beyond the processed data and includes information for many other experiments that do not involve mass spectrometry. It is possible that the generic nature of the ASTM standard design could be used to meet the other needs of the PSI. However, it is also possible that other standards might already exist which are supported by the other instrumentation and software vendors (i.e. image data, sequence data, etc.) allowing these parts of the PSI standard to be developed independently from the ASTM effort. More information on the HUPO-PSI effort can be found at the project web site, (psidev.sourceforge.net).

Approach to Developing a Draft Standard

To deliver a standard that has both broad support and is available quickly, the PSI proposed using as much of the ASTM model as is currently available and focus on the largest volume component of proteomics data first. A step-wise approach to standardization would include first the peak-list type data followed by the search engine result, or spectral annotation information. These two steps could be combined with other meta-data to be used as a repository or database submission format.

The largest volume portion of data in proteomics experiments is the peak-list information for MS and tandem MS experiments. This portion of the data represents the input into protein identification and quantitation software applications and is presumably one of the key minimum data components that would need to be shared between laboratories.

Several groups engaged in proteomics research have adopted similar models (see for example the Institute for Systems Biology project "SASHIMI" at sashimi.sourceforge.net). It was proposed that a narrowly scoped standard covering peak-list data for proteomics could be developed in a short time frame. Since there is already fairly good agreement on what data is needed at this step, a standard approach and a controlled vocabulary was considered an easy first step but also a very useful step since there are several file formats currently being used for this across multiple vendors.

Results of the Discussion

Since the HUPO-PSI workshop in January 2003, many vendors and research groups have contributed XML file formats. These contributions were extremely important in developing an understanding of the scope of the application-specific, and vendor-specific needs. Following the January meeting, formats were contributed by the following companies (in alphabetic order):

Bruker-Daltonics, Inc.
Ciphergen Biosystems, Inc.
Micromass-Waters
Protogen, Inc.
Thermo LabSystems

Without the kind cooperation of each of these companies, it would have been impossible to establish what was a basic 'core' data item versus a application-specific or vendor-specific data item.

The contributions of each vendor helps ensure a reasonable standard is established quickly. The HUPO-PSI effort is greatly enhanced by the open spirit of cooperation shown by all the vendors involved. Thank you. – Randy Julian and Weimin Zhu.

Other formats used for reference included the NIST SpectroML format, the PedroML format from Univ. Manchester, and the mzML format from the Institute for Systems Biology. These formats are also very informative in helping establish minimum requirements for data exchange.

Most vendor formats intermingled data elements which could be represented by the ASTM 'core' with application- and vendor-specific information. Each has been valuable in ensuring that vendor-specific information can be adequately captured and in establishing the minimum meta-data for the proteomics application layer. The ISB format most closely followed the proposed ASTM core by using arrays of base64 encoded binary to represent the actual data arrays. This was originally intended to allow raw data to be represented in a compact fashion. Due to the similarity with the ASTM core, the data component of the ISB's mzML format was used as a starting point for the discussion of what meta data should be included in the application-specific layer.

Conclusions

From this discussion, the following conclusions were reached:

1. The data elements will be represented as arrays of IEEE floating point numbers. The minimum is a single array of mass values. A second array could contain intensity values. Additional arrays with identical index values could be added to allow annotation of each peak with either numbers (charge state, S/N, etc.) or text (sequence, fragment annotation, etc.).
1.a. The data type for each array must be stored (IEEE-32, IEEE-64, string, int, etc.)
2. The number of elements in the array is an optional value
3. A “sequence” number representing an order of collection is an optional value
4. A “retention time” value is an optional value
5. The type of mass spectrum is a required value (A vocabulary is needed for this)
6. The number of mass analysis steps conducted is a required value
7. For each stage of mass spectrometry, the mass of the isolated ion must be stored.
8. For each stage of mass spectrometry other parameters such as intensity, S/N, isolation width, activation energy value, etc. should be allowed as optional values.
9. Charge state of the isolated ion is an optional value for each stage of MS conducted
10. The name and version information for any processing program used to create the peak list are optional values.
11. The name of the original raw data file is an optional value
12. The name/version of the instrument is an optional value

With the agreement to store the peak list in a format compatible with the ASTM core and to generate an XML-Schema which modeled the concepts above, the meeting was adjourned. The minutes are to be distributed to the attendees along with a prototype Schema (.xsd) file for review. This will be done via e-mail with the documents posted on the PSI project web site.

Attendees

Alex Polley (apolley@siaman.co.uk)
Mathias Dreger (siahtam@chemie.fu-berlin.de)
Alexey Nevizhskii (nesvi@systemsbiology.org)
Frederic Schutz (schutz@wehi.edu.au)
Kai Maass (kai.maass@chemie.uni-giessen.de)
Patrick Pedrioli (ppatrick@systemsbiology.org)
David Creasy (dcreasy@matrixscience.com)
John Cottrell (jcottrell@matrixscience.com)
Ian Brookhouse (ian.brookhouse@kratos.co.uk)
Dana Robinson (derobins@uiuc.edu)
Chris Wheeler (chris.wheeler@sciex.com)
Ed Hui (huien@sciex.com)
Nedim Mujezinovic (mujezinovic@imp.univie.ac.at)
Eugene Kapp (eugene.kapp@ludwig.edu.au)
Phillip Young (phillip.young@waters.com)
Nathan Pedrick (npedrick@iupui.edu)
Veronica Pedrick (vbannon@iupui.edu)

Steve Shrader (steve@shraderlabs.com)
John Chik (jchik@ucalgary.ca)
Chris Taylor (chris@bioinfo.man.ac.uk)
Simon Hubbard (simon.hubbard@umist.ac.uk)
Mark Manning-Lee (markhl@prodigy.net)
Philip Doggett (philip.doggett@proteomesystems.com)
Martin Bleuggel (martin.bleuggel@protogen.de)
Johanathan Katz (johnathan@ucla.edu)

Randy Julian (rkj@lilly.com)
Karen Gooding (Gooding_Karen@lilly.com)
Weimin Zhu (weimin@ebi.ac.uk)