complementary: MIAPE Gel presents a lot of information according to protocols, methods and images and GelML could exchange this information with other bioinformatics tools. In spite of the fact that both present a great combination, some problems appeared during the transformation. For instance, MIAPE presents a lineal schema -well defined pipeline- whereas GelML is a FUGE based one, dividing the information along several linked sections. Furthermore, any assumption can be done about the cardinality of the relationships between both examples. Finally, these issues made us to define a set of configuration files to drive the transformation.

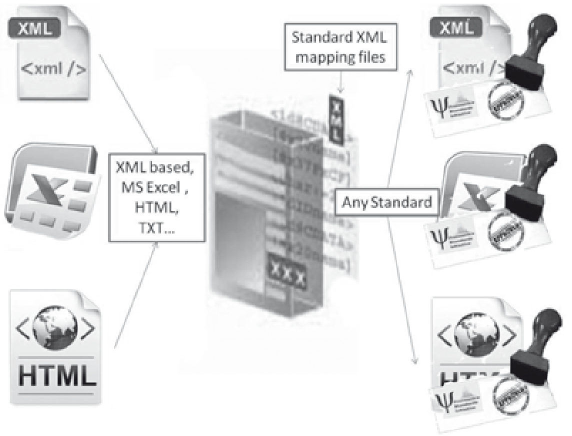
XML
<xml />
HTML
XML based, MS Excel , HTML, TXT...
Standard XML mapping files
XML
Any Standard
XML
<xml />
XXX

Figure 1. *Translation between proteomics standards will be easier.*

The described work results a proper correspondence between both formats, allowing a complete translation between MIAPE Gel and GelML schemas.

References

[1] Martens L, Orchard S, Apweiler R and Hermjakob H. Human proteome organization proteomics standards initiative: data standardization, a view on developments and policy. Mol Cell Proteomics 2007; 6: 1666-1667.

[2] Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H and Apweiler R. PRIDE: a public repository of protein and peptide identifications for the proteomics community; Nucl Acid Res. 2006; 34 (Database issue): D659-D663.

Open-source Bioinformatics Solutions for the Analysis of Mass Spectrometry-based Proteomics Data: Pipelines and Quantitation

Alexandre R. Campos

Proteomics Platform, Barcelona Science Park, Barcelona, Spain

The proteomics community has been generating openly available software framework for systematic proteomic data analyses and management. Processing and analysis of proteomics data involves a complex, multistage process including raw file pre-processing, peptide assignment to MSMS experimental spectra, protein identification and further validation, and in some cases, MS-based quantitation. Here, I list a number of freely available and open-source computational tools for the analysis of proteomics data. Particularly, I focus on available platforms and pipelines, and software for MS-based quantitation.

1. Platforms and Pipelines for LC-MS and LC-MS/MS Data Analysis

With the advent of new generation of mass spectrometers arose the necessity of developing new bioinformatics solutions for mining large data sets. In the past years, we have witnessed a dynamic software development process that embraces various functionalities of data analysis process such as the preprocessing of MS data, evaluation and assignment of MS/MS spectra to peptide sequences, comparison and quantitation of multiple LC-MS experiments. The burgeoning of proteomics data has fostered the development of a number of public domain proteomics pipelines that integrate a number of open-source, cross-platform tools providing a pluggable development framework for the proteomics scientific community. In the following sections, I briefly go over some of these projects.

2. The Trans-Proteomic Pipeline (TPP) (http://tools.proteomecenter.org/software.php)

The TPP is a collection of integrated tools for LC-MS/MS data analysis, developed at the Seattle Proteome Center (SPC). The suite includes tools for conversion of instrument vendor's raw data to mzXML or mzML formats; conversion of spectral search engine (Mascot, X!Tandem, Sequest, Phenyx, OMSSA) results to pepXML format; and statistical validation of search engine results at the peptide- and protein-level with PeptideProphet and ProteinProphet, respectively (and iProphet to combine multiple search results). TPP also supports quantitation analysis for MS1 and MS2 labeling techniques.

3. CPAS (https://www.labkey.org/project/home/begin.view)

The Computational Proteomics Analysis System (CPAS) is implemented as a Tomcat web-based application for mining LC-MS/MS proteomic experiments. CPAS is distributed with X!Tandem search engine and the PeptideProphet and ProteinProphet validation tools; in addition, it can be also used with other search engines, including Mascot and Sequest. The MS/MS analytic module enables users to filter, sort, customize, compare, and export experiment runs. CPAS can also be configured to perform Q3 or XPRESS quantitation analyses. In contrast to TPP, CPAS helps manage information about biological samples and preparation protocols.

4. TOPP - OpenMS Proteomics Pipeline (http://open-ms.sourceforge.net/TOPP.php)

TOPP is a customizable collection of several small applications, based on OpenMS, an open-source C++ library for LC-MS data management and analysis. The individual applications of TOPP can be grouped into several distinct packages: import/export, signal processing, identification, quantitation and analysis. Importing data into TOPP is handled with the FileConverter, which converts several commonly used MS formats into each other. Supported formats are mzML, mzData, mzXML, among other formats. TOPP shares most of the features provided by other pipelines; in addition, TOPP provides a set of computational tools for signal preprocessing (e.g., denoising, baseline correction, and smoothing), and processing (e.g., peak picking, deisotoping, centroiding, and feature extraction).

5. Proteios Software Environment (ProSE) (http://www.proteios.org/)

ProSE was initially created as a repository for proteomics data, and just recently has been extended to automate some data assembly and reporting. ProSE runs on Tomcat server and provides a Web interface for data access and analysis. What makes ProSE different from other platforms is the availability of a programming interface for method developers to write extensions and plug-ins. Extensions and plug-ins make ProSE a highly customizable platform. One can for instance carry out batch searches in OMSSA and X!Tandem via the Web interface, and then automatically upload the results to ProSE. In addition, data can be extracted into dedicated database tables for further work with the results, such as filtering, protein inference, and combination of search results.

6. Software for Analysis of MS-based Quantitative Proteomics Data

In the past years, MS-based quantitation has emerged as a promising core technology for proteomics profiling. Jaffe et al. [1] have classified the available MS platforms for quantitative proteomics into three categories: 1) *identity-based methods* that rely on peptide analysis by data-dependent LC-MS/MS and protein identification by database searching, 2) *pattern-only methods* that focus on production of MS-derived protein patterns, and 3) *hybrid identity/pattern-based methods* use pattern recognition algorithms to *ex post facto* assign the identity of LC-MS peaks against databases of peptide sequence, mass, and retention time built from multiple

Table 1. *Summary of tools for MS-based quantitative proteomics data analysis.*

Software	Input spectra format	Input peptide identification format	Type of quantitation	Language	Graphical User Interface?	Operating System
Census	MS1/MS2 or mzXML	DTASelect or pepXML	MS1, MS2 and Label-free (Id)	Java	Yes	L.O.W[a]
MSQuant	Instrument vendor raw[e]	Mascot results	MS1, MS2 and Label-free (Id)	.NET	Yes	Win
MaxQuant	XCalibur .raw[e] (high resolution)	Mascot results	SILAC and label-free (Id)	.NET	Yes	Win
XPRESS	mzXML	pepXML	MS1	C++	Yes[b]	L.O.W[a]
ASAPRatio	mzXML	pepXML	MS1	C++	Yes[b]	L.O.W[a]
Multi-Q	mzXML	pepXML	MS2	.NET	Yes	Win
iTracker	.mgf or .dta	none	MS2	Perl	No	L.O.W[a]
Quant	.mgf	Mascot .dat	MS2	Matlab or .NET	Yes	L.O.W[a] (Matlab)
Libra	mzXML	pepXML	MS2	C++	Yes[b]	L.O.W[a]
APEX	none	pepXML	Label-free (SC)	Java	Yes	L.O.W[a]
SuperHirn	mzXML	pepXML	Label-free (HIP)	C++	No	L.O.W[a]
PEPPeR	Feature list (e.g., msInspect)	PEPPeR .txt format	Label-free (HIP)	Perl	Yes[c]	L.O.W[a]
msInspect	mzXML	pepXML	Label-free (Id, HIP, AMT)	Java	Yes	L.O.W[a]
IDEAL-Q	mzXML	pepXML	Label-free (HIP)	.NET	Yes	Win
msBID	mzXML	pepXML	Label-free (Id)	Java,Perl	No	L.O.W[a]
TOPP	mzML	TOPP adapters	MS1, MS2 and Label-free (Id)	C++	Yes	L.O.W[a]
PNNL pipeline	Decon2LS files[d] or mzXML	X!Tandem or Sequest	Label-free (AMT)	.NET	Yes	L.O.W[a]

a *The acronym L.O.W. refers to Linux, OSX and Windows*

b *Using TPP graphical user interface*

c *Using GenePattern graphical user interface*

d *Decon2LS files can be extracted from mzXML, or a number of instrument vendor's raw format*

e *To convert instrument vendor's raw format, the vendor's MS analysis software must be installed.*

Label-free methods: SC (Spectral Counting); Id (Identity-based); HIP (hybrid identity/pattern-based); and AMT (accurate mass and time)

MS1: *methods based on precursor intensity or area (e.g., 15N, 18O, SILAC)*

MS2: *methods based on reporter ions in MS/MS (e.g., iTRAQ, TMT)*

experiments. As MS-based quantitation remains a rapidly developing field with many different experimental approaches, a large number of open-source (or freely available (academic)) software has been developed and made available for the scientific community (Table 1).

Reference

[1] Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA y Carr SA. PEPPeR, a platform for experimental proteomic pattern recognition. Mol Cell Proteomics 2006;5:1927-41.