Biochimica et Biophysica Acta xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Review

A tutorial for software development in quantitative proteomics using PSI standard formats ☆

Faviel F. Gonzalez-Galarza [a,1], Da Qi [a,1], Jun Fan [b,2], Conrad Bessant [b,2], Andrew R. Jones [a,*]

[a] Institute of Integrative Biology, University of Liverpool, Liverpool, UK
[b] Bioinformatics Group, Cranfield Health, Cranfield University, Cranfield, UK

ARTICLE INFO

Article history:
Received 7 November 2012
Received in revised form 22 February 2013
Accepted 5 April 2013
Available online xxxx

Keywords:
Quantitative proteomics
Software
Standard formats
APIs

ABSTRACT

The Human Proteome Organisation — Proteomics Standards Initiative (HUPO-PSI) has been working for ten years on the development of standardised formats that facilitate data sharing and public database deposition. In this article, we review three HUPO-PSI data standards — mzML, mzIdentML and mzQuantML, which can be used to design a complete quantitative analysis pipeline in mass spectrometry (MS)-based proteomics. In this tutorial, we briefly describe the content of each data model, sufficient for bioinformaticians to devise proteomics software. We also provide guidance on the use of recently released application programming interfaces (APIs) developed in Java for each of these standards, which makes it straightforward to read and write files of any size. We have produced a set of example Java classes and a basic graphical user interface to demonstrate how to use the most important parts of the PSI standards, available from http://code.google.com/p/psi-standard-formats-tutorial. This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era.

1. Introduction

The large scale identification and quantification of proteins in proteomics studies have always relied upon a close association with computational developments termed 'proteome bioinformatics' [1]. This field, also known as 'computational proteomics', involves the development of methods, algorithms, databases, visualisation techniques and high-throughput analysis to interpret large scale experimental studies [2–4]. This has been necessitated as mass spectrometry (MS) data, used in the identification and/or quantification of proteins, which are complex to interpret [5,6]. At present, there is no all-in-one software solution in quantitative proteomics, with huge variability in the protocols employed in different labs related to protein or peptide separation, labelling protocols and MS instruments. A full description of methods and software for protein identification/quantification is out of scope for this article, but for more details see [7–10]. The complexity in data analysis may vary considerably amongst techniques [9], for example, the following list describes the different stages that may be performed in a 'label-free' peptide feature-based quantification pipeline: (i) raw signal processing to locate (and potentially quantify) peptides in a two-dimensional data set, i.e. retention time (RT) versus mass/charge (m/z); (ii) alignment of parallel runs in the RT dimension; (iii) identification of a peptide sequence via a peptide-spectrum match (PSM), by querying a peptide's fragmentation pattern (MS^2 spectrum) against a sequence database; (iv) processing a number of PSMs to infer the presence of proteins; (v) statistical analysis of differential expression and so on (Fig. 1). This example serves to illustrate the diverse types of analysis that may be performed to obtain protein expression values (i.e. quantified proteins in one or more samples) from raw MS data. Current research is focussed on the optimisation of all of these stages by the vendors of instruments, vendors of commercial analysis software or by proteome bioinformatics research groups [11–14].

The Human Proteome Organisation — Proteomics Standards Initiative (HUPO-PSI, or simply PSI) is a consortium of academic and industrial research groups, instrument manufacturers, commercial software vendors and other stakeholders aiming to standardise how proteomics data sets are reported and shared [15]. The need for the PSI's efforts came about since, historically, proteome bioinformatics has been impeded by the diverse range of data formats used for representing raw data (e.g. instrument vendors' proprietary formats or open-source formats), partially processed data for peptide identification (i.e. peak lists in text-based formats — e.g. MGF, dta, pkl, etc.), results of search engines and results of quantification software (see review in [12]). The most well-known open formats developed in the past include mzXML (for raw spectral data), pepXML (peptide identifications) and protXML (for protein identifications), developed as part of the Trans-Proteomic

☆ This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era.

* Corresponding author at: Institute of Integrative Biology, University of Liverpool, The Biosciences Building, Crown Street, L69 7ZB, Liverpool, UK. Tel.: +44 151 795 4514; fax: +44 151 795 4410.
E-mail addresses: F.Galarza@liv.ac.uk (F.F. Gonzalez-Galarza), D.Qi@liv.ac.uk (D. Qi), j.fan@cranfield.ac.uk (J. Fan), c.bessant@cranfield.ac.uk (C. Bessant), Andrew.Jones@liv.ac.uk (A.R. Jones).
[1] Tel.: +44 151 795 4555; fax: +44 151 795 4410.
[2] Tel.: +44 123 475 8512; fax: +44 123 475 8517.

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.04.004

Fig. 1. Prototypical workflow for a label free quantitative analysis, showing which stages are covered by different PSI formats.

Pipeline (TPP) [16]. The developers of these formats have joined the PSI to develop a suite of international standards, developed in a collaborative manner.

In the context of the PSI, several high-profile outputs have been made. In the experimental proteomics domain, these include guidelines describing the minimum information that should be reported about a proteomics experiment — the parent 'MIAPE' document [17] and a set of technology specific modules [18–23]. The PSI has also produced standard data formats — mzML for raw or processed MS data [24], TraML for input transitions in selected reaction monitoring (SRM) approaches [25], mzIdentML for peptide and protein identification data [26] and two new efforts for capturing quantitation data — mzQuantML capturing a detailed trace of each stage of quantitative analysis [27] and mzTab capturing a simple summary of final results designed for viewing in spreadsheets or statistical processing software [28]. All of the standard formats have been designed to allow the capture of MIAPE-compliant details according to the corresponding module, but typically the formats can also be valid in different contexts if they contain more or less detail than stipulated by MIAPE documents.

With the exception of mzTab, the PSI standards are represented in Extensible Markup Language (XML) — an industry standard specification in which data are enclosed in opening and closing brackets (called elements or tags) that describe the type of data stored e.g. <spectrum>*the spectral data*</spectrum>. In XML documents, nesting of elements is allowed to build up a hierarchical tree structure so that complex concepts can be represented in a manner that can be interpreted by other developers and by software. XML files have several advantages over other formats such as platform independency, no limit on the number/type of tags defined, and that they can be easily manipulated with a large number of free software tools helping in the design, processing and visualisation of data. XML files are also known to have some disadvantages including files being relatively verbose compared with other encodings, and they tend to require specialist software to be developed for data manipulation and visualisation. Despite these constraints, XML is generally preferred by the PSI because software can be developed using industry standard tools and the format can be formally defined via an XML Schema.

All of the standards described here make use of a common controlled vocabulary (CV), called the PSI-MS CV [29], containing more than 2000 well-defined terms describing all aspects of proteomics analysis — instruments and their parts, software and parameters, etc. CV terms are used within the format to ensure that concepts can be described using standardised terminology, comprehensible to both people and software, and additional validation software has been implemented by the PSI to verify that CV terms are used correctly within formats [30].

The requirement to work with large files (>10 GB) and fast parsing has lead bioinformatics groups to work on applications to handle these tasks. For each PSI format, various implementations have been developed for import/export from commercial and open-source software, plus software interfaces (Application Programming Interfaces or APIs) to assist developers to implement standards in their own analysis software. In this domain, the ProteoWizard project [31] has produced several software utilities for converting most proprietary vendor formats into mzML and some search engine output formats into mzIdentML. ProteoWizard also contains an internal data model (in C++) for working with MS or identification data, allowing developers to create tools without requiring an underlying knowledge of the source data format. Various groups have also collaborated to build APIs in Java for processing each of the standards, called jmzML [32], jTraML [33], jmzIdentML [34], jmzQuantML [35] and jmzTab [36]. The Java APIs have read/write capabilities and implement random-access strategies allowing files of any size to be processed without requiring the whole file to be loaded into memory.

In this article, we briefly review the model behind three of the core (XML-based) formats — mzML, mzIdentML and mzQuantML, focusing on the most important features that developers should be aware of when implementing support for them in software. The mzTab standard is not covered in this article, since, due to its limited content and simpler tab-separated structure, it is more straightforward for developers to work with. We then illustrate how the corresponding Java APIs can be used to develop support for these formats rapidly, for a range of common tasks that may be employed in a quantitative proteomics pipeline. We anticipate that this article will serve as a useful guide to proteome bioinformatics developers as to how the formats can be easily supported and integrated into existing or new software workflows.

2. Data standards and programming interfaces

A variety of quantitative proteomics approaches have become increasingly popular over the last few years, with a wide range of software supporting some or all approaches (reviewed in [37]). In this section, we discuss the basic usage of three main PSI standards in a quantitative analysis pipeline and how potential users of the standards can convert their own files into the standards (Table 1). The PSI standards are all maintained in regularly updated subversion repositories on Google Code. In addition to the published articles [24,26], basic tutorial documents and full specifications are available. For more details, consult the PSI website and follow appropriate links at http://www.psidev.info/.

2.1. Converting MS data into mzML and accessing spectral data

Each instrument vendor exports raw data into their own file format. In some cases, these companies have developed their own tools for converting these proprietary files into readable text or XML formats; however, full export to the most recent PSI standards is not yet available for several vendors. To assist bench scientists in the conversion of different vendor formats, a number of software packages have been developed by different groups. For example ProteoWizard [31] can convert most of the common vendor files such as .RAW (Thermo Scientific), .raw (Waters), .wiff (Applied Biosystems), .d (Agilent), and others into

Please cite this article as: F.F. Gonzalez-Galarza, et al., A tutorial for software development in quantitative proteomics using PSI standard formats, Biochim. Biophys. Acta (2013), http://dx.doi.org/10.1016/j.bbapap.2013.04.004