thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer *msf* files.

Niklaas Colaert[1,2], Harald Barsnes[3,4], Marc Vaudel[5], Kenny Helsens[1,2], Evy Timmerman[1,2], Albert Sickmann[5], Kris Gevaert[1,2], Lennart Martens[1,2*]

1. Department of Medical Protein Research, VIB, Ghent, Belgium
2. Department of Biochemistry, Ghent University, Ghent, Belgium
3. Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway
4. Computational Biology Unit, Uni Computing, University of Bergen, Bergen, Norway
5. Leibniz-Institut für Analytische Wissenschaften - ISAS. Dortmund, Germany

*Corresponding author: Professor Lennart Martens, Department of Medical Protein Research and Biochemistry, VIB and Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. Email: lennart.martens@ugent.be. Tel: 32-92649458 Fax: 32-92649484.

Keywords:

proteomics, bioinformatics, mass spectrometry

Abbreviations:

LC: liquid chromatography, UML: Unified Modelling Language

Abstract

The Thermo Proteome Discoverer program integrates both peptide identification and quantification into a single workflow for peptide-centric proteomics. Furthermore, its close integration with Thermo mass spectrometers has made it increasingly popular in the field. Here, we present a Java library to parse the *msf* files that constitute the output of Proteome Discoverer. The parser is also implemented as a graphical user interface allowing convenient access to the information found in the *msf* files, and in Rover, a program to analyse and validate quantitative proteomics information. All code, binaries and documentation is freely available at http://thermo-msf-parser.googlecode.com.

Introduction

MS/MS based peptide identification is a key element in peptide centric proteomics. The identified peptides are used to infer proteins, ultimately providing proteomics information about the studied biological system. Additionally, the development of both labelled and label-free quantitative techniques allows peptides and therefore proteins to be quantified, adding further detail to the study of biological systems [1]. Proteome Discoverer, a tool developed by the mass spectrometer vendor Thermo Scientific, lets users perform both peptide identification and quantification starting from the instrument raw data. Here we present a free and open source Java library to access and parse the output files (.msf files) of Proteome Discoverer. An open-source Thermo MSF Viewer has also been developed as both an end-user tool as well as a demonstration of the library's capabilities.

Peptide identifications are based on MS/MS spectra generated by mass spectrometers. The first step in a proteomics workflow is the extraction of these MS/MS spectra from the vendor-specific raw data files [2]. This extraction can be performed by an array of free (e.g. MaxQuant [3]), open source (e.g. ProteoWizard [4]) or commercial (e.g. Mascot Distiller http://www.matrixscience.com/distiller.html) tools. The second step is the identification of the extracted MS/MS spectra, where the most popular method is to match the spectra to peptides via protein database searching. Here, a protein database is scanned for peptides with a correct precursor mass dependent on the protease used and the

instrument precision. I*n silico* MS/MS spectra are generated for the selected peptides and these spectra are compared to the experimental spectra. Different algorithms (e.g., Mascot [5], SEQUEST [6], OMSSA [7] and X!Tandem [8]) have been developed to perform these database searches and to score the peptide-to-spectrum matches. The last step in a quantitative proteomics workflow is the quantification of peptides and proteins. The peptides identified in the previous step can be tagged with an MS or MS/MS level label, allowing quantification of the relative abundance changes between two samples [1]. Peptide quantification can also be performed by making use of label-free techniques [9]. Different algorithms were developed to perform this label and label-free based quantification of peptide and protein ratios [1,10]. A typical proteomics peptide identification and quantification workflow is thus built from many different algorithms and tools.

Proteome Discoverer (Thermo Scientific) is a program that integrates all the different steps in a quantitative proteomics experiment (MS/MS spectrum extraction, peptide identification and quantification) into user-configurable, automated workflows. Currently, Mascot [5], SEQUEST [6] and ZCore (http://www.thermo.com.cn/Resources/200807/1112724272.pdf) are supported as fully integrated peptide identification tools. Peptide quantification on the other hand is performed by an integral algorithm in Proteome Discoverer itself. The results of such a workflow are stored in so-called .msf files. These files hold information on peptide identification, quantification, the recorded MS, MS/MS and quantification spectra, and the chromatograms. This wealth of included information makes it possible to analyse all relevant data from this format alone without the need for the raw data or the peptide identification algorithm output. We here therefore present an open source Java library that parses *msf* files created by Thermo's Proteome Discoverer (version 1.2 and 1.3).

Materials and methods

The thermo-msf-parser library was created in the Java programming language. The parser makes use of the Xerial SQLite JDBC library (http://www.xerial.org/trac/Xerial/wiki/SQLiteJDBC) to make a connection to the *msf* files. Since each SQLite file is effectively a file-based miniature relational

database, the data contained in these files is readily accessible, even as random-access. The thermo-msf-parser can therefore extract all desired information from these *msf* files and is able to transform it into simple, interconnected Java objects (see Figure 1 for a simplified UML diagram).

Results

The ease of using the thermo-msf-parser Java library is demonstrated by two different examples. The first is a freely available end-user oriented graphical user interface, called Thermo MSF Viewer, that visualises the information found in *msf* files (see Figure 2). Because the parser is very memory efficient, it can load multiple files simultaneously, making the integration, comparison and analysis of identified and quantified peptides and proteins across a complete study a lot easier, since all data found in different LC-runs and *msf* files can be displayed at once. In addition to having a low memory footprint, the parser is very fast. Loading several *msf* files in the Thermo MSF Viewer presented here is approximately four times faster than opening the same files on the same pc with Proteome Discoverer itself. Besides displaying lists of identified peptides and proteins in interactive tables (making use of the JSparklines library, http://jsparklines.googlecode.com), the Thermo MSF Viewer also visualizes MS spectra, fragment ion annotated MS/MS spectra, quantification spectra, chromatograms and protein sequence coverage, making detailed data analysis straightforward. The displays are derived from the compomics-utilities Java library for proteomics [11]. Due to the open nature of the Java programming language, the thermo-msf-parser is fully platform independent. Analysing and viewing *msf* files can thus not only be done on a computer running Microsoft Windows but also on Apple OS-X and Linux systems. In contrast, Proteome Discoverer itself can only run on certain versions of the Windows operating system. Our viewer thus allows *msf* files to be used as inherently portable and easily visualized data files for detailed mass spectrometry information.

The second example of the use of the thermo-msf-parser library is found in the Rover program (http://compomics-rover.googlecode.com) [12]. Developed to visualize quantitative proteomics data in the context of a complete experiment, with special emphasis on the protein inference problem,

Rover has now extended its supported data import formats from Census [13], MsQuant [14], MaxQuant [3] and Mascot Distiller to also include Thermo *msf* files. Furthermore, we recently implemented an algorithm in Rover merging the output from different quantification workflows obtaining an increase in the number of quantified proteins [15]. With the *msf* support added to Rover, this powerful algorithm can now also be used with quantitative data from Proteome Discoverer.

Discussion

By creating a free and open source Java library for Thermo's Proteome Discoverer *msf* files it becomes straightforward to incorporate the results stored in these files in downstream analysis applications. Here, the use of the library was demonstrated with two end-user oriented examples. First, a graphical user interface for displaying all information found in one or more *msf* files was presented. Second, the implementation of the library in the Rover tool was shown. Binaries, Java source and additional documentation is freely available at http://thermo-msf-parser.googlecode.com. Both the *msf* parser and the viewer are available under the permissive Apache2 open source license.

References

(1) Vaudel, M.; Sickmann, A.; Martens, L. Peptide and protein quantification: a map of the minefield. *Proteomics* **2010**, 10 (4), 650-670.

(2) Martens, L.; Nesvizhskii, A. I.; Hermjakob, H.; Adamski, M.; Omenn, G. S.; Vandekerckhove, J.; Gevaert, K. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* **2005**, 5 (13), 3501-3505.

(3) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, 26 (12), 1367-1372.

(4) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, 24 (21), 2534-2536.

(5) Pappin, D. J.; Perkins, D. N.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, 20 (18), 3551-3567.

(6) Eng, J. K.; McCormack, A. L.; Yates III, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, 5 (4), 976-989.

(7) Geer, L. Y.; Markey, S. P.; Kowalak, J. a; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, 3 (5), 958-964.

(8) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, 20 (9), 1466-1467.

(9) Colaert, N.; Vandekerckhove, J.; Gevaert, K.; Martens, L. A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision. *Proteomics* **2011**, 11 (6), 1110-1113.

(10) Colaert, N.; Gevaert, K.; Martens, L. RIBAR and xRIBAR: Methods for Reproducible Relative MS/MS-based Label-Free Protein Quantification. J. Proteome Res. **2011**, e-pub.

(11) Barsnes, H.; Vaudel, M.; Colaert, N.; Helsens, K.; Sickmann, A.; Berven, F. S.; Martens, L. compomics-utilities: an open-source Java library for computational proteomics. *BMC bioinformatics* **2011**, 12, 70.

(12) Colaert, N.; Helsens, K.; Impens, F.; Vandekerckhove, J.; Gevaert, K. Rover: a tool to visualize and validate quantitative proteomics data from different sources. *Proteomics* **2010**, 10 (6), 1226-1229.

(13) Park, S. K.; Venable, J. D.; Xu, T.; Yates, J. R. A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **2008**, 5 (4), 319-322.

(14) Mortensen, P.; Gouw, J. W.; Olsen, J. V.; Ong, S.-E.; Rigbolt, K. T. G.; Bunkenborg, J.; Cox, J.; Foster, L. J.; Heck, A. J. R.; Blagoev, B.; Andersen, J. S.; Mann, M. *J.* MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res*. **2010**, 9 (1), 393-403.

(15) Colaert, N.; Van Huele, C.; Degroeve, S.; Staes, A.; Vandekerckhove, J.; Gevaert, K.; Martens, L. Combining quantitative proteomics data processing workflows for greater sensitivity. *Nat. Methods* **2011**, 8 (6), 481-483.

Acknowledgments

H.B. is supported by the Norwegian Research Council. K.H. is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO-Vlaanderen). L.M. and K.G. acknowledge the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”), the PRIME-XS project, grant agreement number 262067, and the 'ProteomeXchange' project, grant agreement number 260558, both funded by the European Union 7th Framework Program.

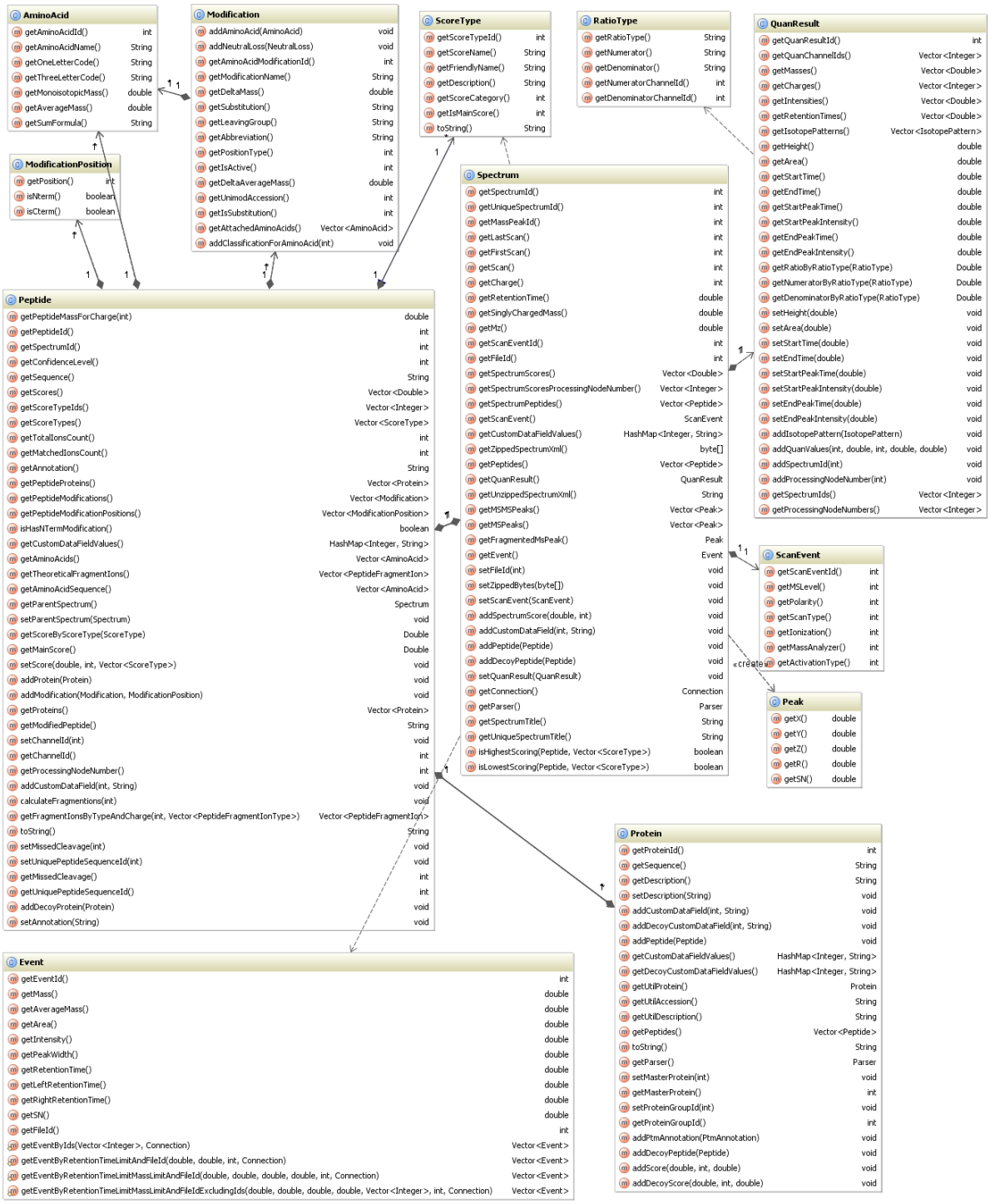

Figure 1. Simplified UML class diagram of the components in the thermo-msf-parser Java library showing the data contents of, and relationships between the different Java objects.

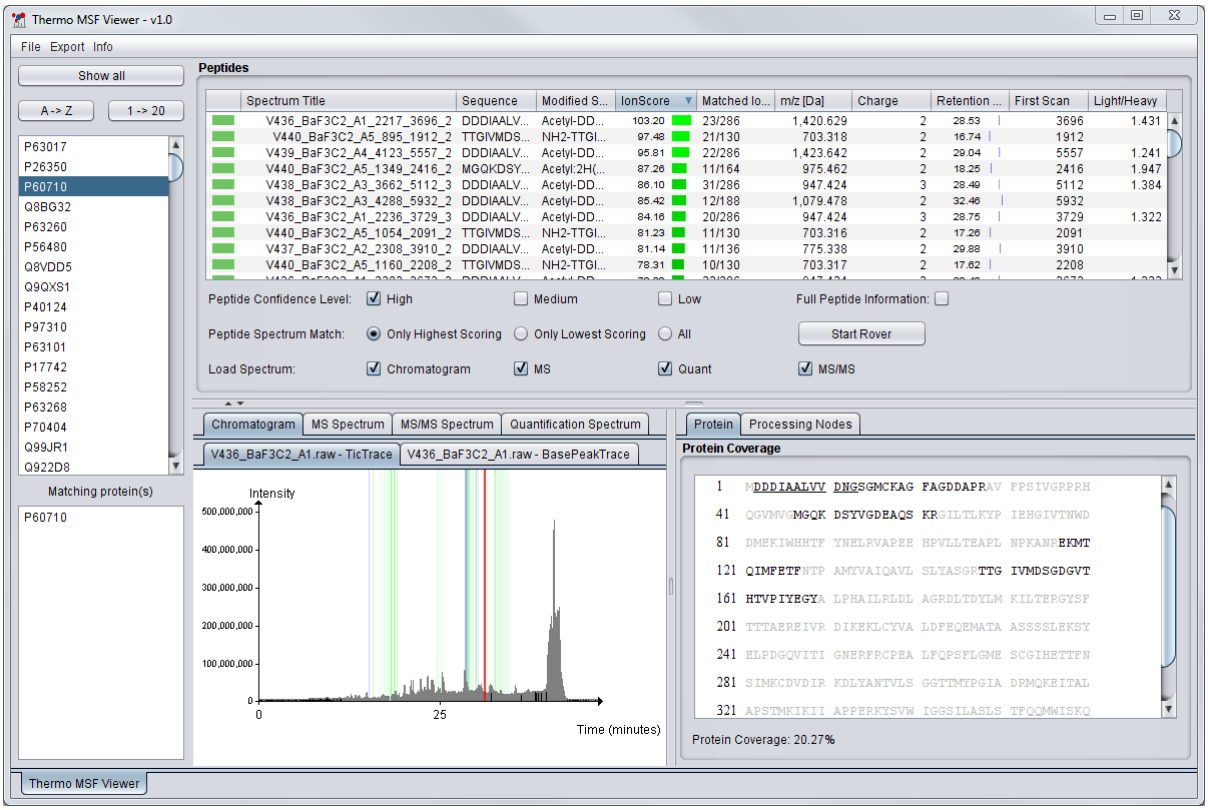

Figure 2. Screenshot of the main interface of the Thermo MSF Viewer, visualizing multiple *msf* files at once.

TOC

The Thermo Proteome Discoverer program integrates both peptide identification and quantification into a single workflow for peptide-centric proteomics. Here, we present a Java library to parse the *msf* files that constitute the output of Proteome Discoverer. The parser is implemented in a graphical user interface allowing convenient access to the information found in the *msf* files.