MSnbase input/output capabilities

Laurent Gatto
lg390@cam.ac.uk
Cambridge Center for Proteomics
Kathryn S. Lilley Group
University of Cambridge

December 5, 2012

Abstract

This vignette describes MSnbase's input and output capabilites.

Keywords: Mass Spectrometry (MS), proteomics, infrastructure, IO.

1 Overview

MSnbase's aims are to facilitate the reproducible analysis of mass spectrometry data within the `R` environment, from raw data import and processing, feature quantification, quantification and statistical analysis of the results. Data import functions for several formats are provided and intermediate or final results can also be saved or exported. These capabilites are presented below.

2 Data input

Raw data Data stored in one of the published `XML`-based formats. i.e. `mzXML` (Pedrioli et al., 2004), `mzData` (Orchard et al., 2007) or `mzML` (Martens et al., 2010), can be imported with the `readMSData` method, which makes use of the mzR package to create `MSnExp` objects. The files can be in profile or centroided mode. See `?readMSData` for details.

Peak lists Peak lists in the `mgf` format[1] can be imported using the `readMgfData`. In this case, the peak data has generally been pre-processed by other software. See `?readMgfData` for details.

Quantitation data Third party software can be used to generate quantitative data and exported as a spreadsheet (generally comma or tab separated format). This data as well as any additional metadata can be imported with the `readMSnSet` function. See `?readMSnSet` for details.

MSnbase also supports the `mzTab` format[2], a light-weight, tab-delimited file format for proteomics data developed within the Proteomics Standards Initiative (PSI). `mzTab` files can be read into `R` with `readMzTabData` to create and `MSnSet` instance.

[1] http://www.matrixscience.com/help/data_file_help.html#GEN
[2] http://code.google.com/p/mztab/

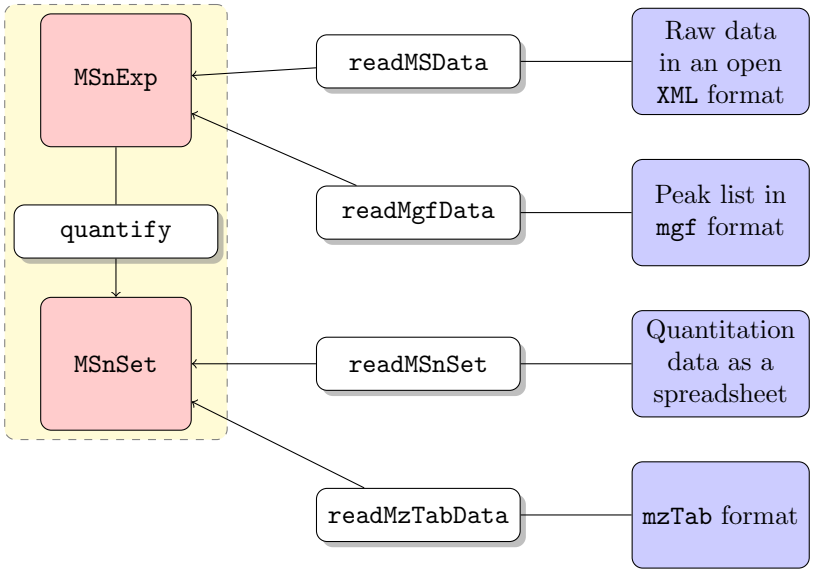

Figure 1: Illustration of `MSnbase` input capabilities. The white and red boxes represent `R` functions/methods and objects respectively. The blue boxes represent different disk storage formats.

3 Data output

RData files `R` objects can most easily be stored on disk with the `save` function. It creates compressed binary images of the data representation that can later be read back from the file with the `load` function.

Peak lists `MSnExp` instances as well as individual spectra can be written as `mgf` files with the `writeMgfData` method. Note that the metadata in the original `R` object can not be included in the file. See `?writeMgfData` for details.

Quantitation data Quantitation data can be exported to spreadsheet files with the `write.exprs` method. Feature metadata can be appended to the feature intenstiy values. See `?writeMgfData` for details.

`MSnSet` instances can also be exported to `mzTab` files using the `writeMzTabData` function.

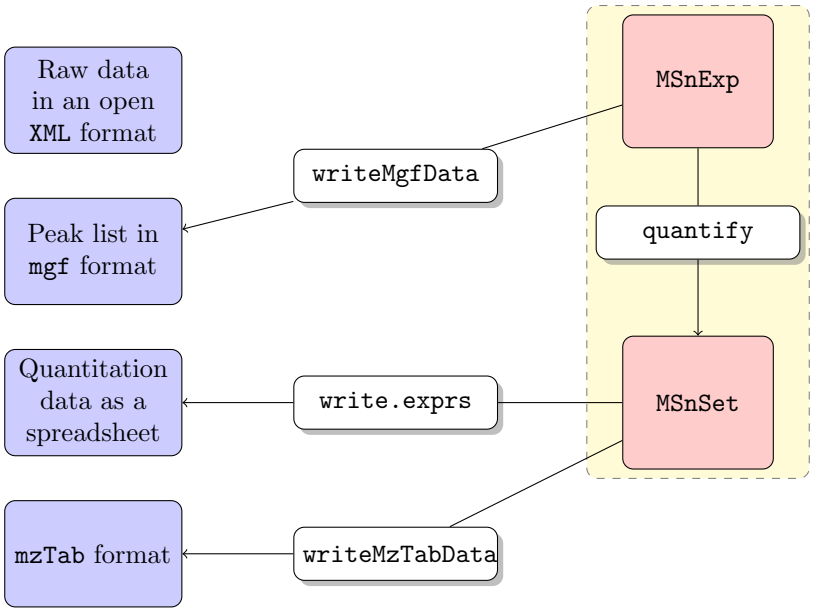

Figure 2: Illustration of `MSnbase` output capabilities. The white and red boxes represent `R` functions/methods and objects respectively. The blue boxes represent different disk storage formats.

References

Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kes sner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Ro mpp, Steffen Neumann, Angel D Pizarro, Lu isa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Pune et Souda, Henning Hermjakob, Pierre-Alain Binz, and Eric W Deutsch. mzml - a community standard for mass spectrometry data. *Molecular & Cellular Proteomics : MCP*, 2010. doi: 10.1074/mcp.R110.000133.

Sandra Orchard, Luisa Montechi-Palazzi, Eric W Deutsch, Pierre-Alain Binz, Andrew R Jones, Norman Paton, Angel Pizarro, David M Creasy, Jérôme Wojcik, and Henning Hermjakob. Five years of progress in the standardization of proteomics data 4th annual spring workshop of the hupo-proteomics standards initiative april 23-25, 2007 ecole nationale supérieure (ens), lyon, france. *Proteomics*, 7(19):3436–40, 2007. doi: 10.1002/pmic.200700658.

Patrick G A Pedrioli, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, Brian Raught, Brian Pratt, Erik Nilsson, Ruth H Angeletti, Rolf Apweiler, Kei Cheung, Catherine E Costello, Henning Hermjakob, Sequin Huang, Randall K Julian, Eugene Kapp, Mark E McComb, Stephen G Oliver, Gilbert Omenn, Norman W Paton, Richard Simpson, Richard Smith, Chris F Taylor, Weimin Zhu, and Ruedi Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, 22(11):1459–66, 2004. doi: 10.1038/nbt1031.