MSnbase—an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation

Laurent Gatto∗ and Kathryn S. Lilley

Cambridge Centre for Proteomics, Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Tennis Court Road, CB2 1QR, Cambridge, UK

Associate Editor: John Quackenbush

ABSTRACT

Summary: MSnbase is an R/Bioconductor package for the analysis of quantitative proteomics experiments that use isobaric tagging. It provides an exploratory data analysis framework for reproducible research, allowing raw data import, quality control, visualization, data processing and quantitation. MSnbase allows direct integration of quantitative proteomics data with additional facilities for statistical analysis provided by the Bioconductor project.

Availability: MSnbase is implemented in R (version ≥2.13.0) and available at the Bioconductor web site (http://www.bioconductor.org). Vignettes outlining typical workflows, input/output capabilities and detailing underlying infrastructure are included in the package.

Contact: lg390@cam.ac.uk

Supplementary information: Supplementary data are available from Bioinformatics online.

Received on June 8, 2011; revised on September 29, 2011; accepted on November 16, 2011

1 INTRODUCTION

Proteomics, the analysis of the entire protein complement expressed by a genome, has recently benefited from substantial technological advances (Nilsson et al. 2010). Among the numerous approaches to interrogate proteomes, a popular technique is the quantitation of samples using isobaric tags. Nevertheless, high-throughput proteomics lags behind genomics, genetics or transcriptomics in terms of reproducible data analysis and development of analytical strategies. One of the challenges within proteomics is the availability of open, extensible and efficient data mining and statistical assessment environments (Lillie et al. 2013). Here, we introduce the MSnbase package, part of the Bioconductor project (Gentleman et al. 2004). MSnbase extends Bioconductor with tools and data structures to import raw tandem mass spectrometry (MS2) data, perform exploratory data analysis and quantification of isobaric reporter tags. The infrastructure is compatible with existing Bioconductor core packages, incorporating available features and allowing direct utilization of additional functionality.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

2 AVAILABLE FUNCTIONALITY

2.1 Data structures

MSnbase provides computational data structures to allow representation of mass spectrometry data types including, individual mass spectra, full experiments, quantification data and information pertaining to isobaric tags. Data structures for commercially available reagents like iTRAQ (Ross et al. 2004) or TMT (Thompson et al. 2006) tags are included in the package, although any arbitrary set of peaks can be defined.

MSnbase also supports storage and direct access to the description of individual features (mass spectra, peptides or proteins), samples and analysis workflows as well as MIAPE (Taylor et al. 2007) compliant information. Metadata associated with an experiment and the transformations it has undergone are accounted for and inherited along the analysis pipeline.

2.2 Data processing and visualization

A typical workflow starts by loading profile-mode or centroided data in an open XML-based standard format or mgf peak lists. Alternatively, quantitative data stored in spreadsheet files can also be imported into the MSnbase framework for post-processing and statistical analysis. Several methods geared towards quality control are available within MSnbase. The availability of easily accessible structured data combined with R data analysis and plotting facilities allows extensive exploratory data analysis to obtain a precise overview of the data and to perform experiment-specific quality control.

MSnbase enables visualization of individual or sets of spectra of particular interest, optionally highlighting the reporter ions (see left panel on Fig. 1 for an example) to facilitate assessment by proteomics scientists. Pre-processing of individual spectra, including removal of low intensity peaks and experiment-wide filtering to remove low-quality spectra or extract specific spectra of interest into new sub-experiments, is easily accomplished. Specific spectra regions can be retrieved to focus on mass to charge (m/z) regions, e.g. the reporter ions m/z range, while preserving both data structure and metadata.

2.3 Quantitation

MSnbase focuses on MS2 level quantitation using isobaric reporter tags. Whole experiments or individual spectra can be quantified using the quantify method upon specifying the isobaric reporter ion tags. Quantitation is performed, among others, by calculating...
MSnbase

This note introduces the Bioconductor package for convenient data manipulation, processing, visualization and quantification, from which users and programmers can build and develop coherent and reproducible analysis pipelines. Compatibility with pre-existing infrastructure takes advantage of available normalization algorithms and facilitates application of statistical analysis methods provided by other Bioconductor packages. In future, MSnbase will be further developed to define standard data processing pipelines, facilitate the inclusion of peptide identification data and allow label-free quantitation. Support is provided on the Bioconductor help mailing list.

ACKNOWLEDGEMENTS

The authors thank Dr Matthew Trotter, University of Cambridge, for helpful comments on previous versions of the manuscript.

Funding: BBSRC Tools and Resources Development Fund (Award BB/G0124618/1); European Union 7th Framework Program (PRIME-XS project, grant agreement number 262067).

Conflict of Interest: none declared.

REFERENCES


Huber, W. et al. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics, 18 (Suppl. 1), S96–S104.


