BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 2 2012, pages 288–289
doi:10.1093/bioinformatics/btr645

Gene expression

Advance Access publication November 22, 2011

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 stategies. One of the challenges within proteomics is the availability of open, extensible and efficient data mining and statistical assessment environments (Lilley *et al.*, 2011).

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.

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.*, 2003) 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 repoter 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

*To whom correspondence should be addressed.

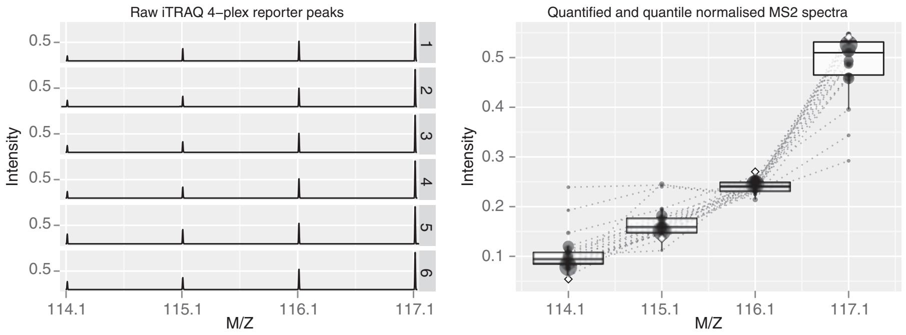

Fig. 1. Illustration of some of `MSnbase` data processing and visualization capabilities. On the left, reporter ion peaks (re-scaled between 0 and 1) for 6 MS2 spectra of different peptide ions derived from a spiked-in protein have been extracted from a large experiment. On the right, relative intensities distribution of all spectra of the same protein are plotted. Point size is proportional to the sum of reporter intensities, and expected values are marked as diamonds.

the area under the curve (for profile mode data) or the maximum peak intensity (for centroided data).

2.4 Data post-processing

Due to isotopic contamination in tags, reporter ion peaks will contribute to those of neighbouring reporter ions. Purity correction of such overlap can be performed with the `purityCorrect` method, providing relative reporter percentages available from the tag manufacturers.

After quantitation, individual spectra can be amalgamated using the `combineFeatures` method. Structures defining which features to combine (for example based on precursor *m/z*, peptides or proteins) can be extracted directly from the feature metadata. Lastly, a summarization function that describes how to combine feature intensities may be selected from a list of possible options (mean, median, weighted mean, sum or median polish) or, a user defined alternative.

Normalization of expression data is performed using the `normalize` method, and accommodates several well-known normalization algorithms such as quantile normalization (Bolstad *et al.*, 2003) (applied in the right panel of Fig. 1) or variance stabilization normalization (Huber *et al.*, 2002; Karp *et al.*, 2010).

We illustrate the above functionality on a real, publicly available dataset as a Supplementary File. Data processing, from import to post-processing took about 20 min on a standard laptop.

3 CONCLUSION

This note introduces the Bioconductor `MSnbase` package for the analysis of quantitative proteomics experiments that use isobaric tagging. The package provides the necessary environment 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/GO24618/1); European Union 7th Framework Program (PRIME-XS project, grant agreement number 262067).

Conflict of Interest: none declared.

REFERENCES

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

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.

Karp,N.A. *et al*. (2010) Addressing accuracy and precision issues in itraq quantitation. *Mol. Cell Proteomics*, **9**, 1885–1897.

Lilley,K.S. *et al*. (2011) Challenges for proteomics core facilities. *Proteomics*, **11**, 1017–1025.

Nilsson,T. *et al*. (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods*, **7**, 681–685.

Ross,P.L. *et al*. (2004) Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics*, **3**, 1154–1169.

Taylor,C.F. *et al*. (2007) The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.*, **25**, 887–893.

Thompson,A. *et al*. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, **75**, 1895–1904.