MALDIquant: a versatile R package for the analysis of mass spectrometry data

Sebastian Gibb[*] and Korbinian Strimmer

Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Associate Editor: Alex Bateman

ABSTRACT

Summary: MALDIquant is an R package providing a complete and modular analysis pipeline for quantitative analysis of mass spectrometry data. MALDIquant is specifically designed with application in clinical diagnostics in mind and implements sophisticated routines for importing raw data, preprocessing, non-linear peak alignment and calibration. It also handles technical replicates as well as spectra with unequal resolution.

Availability: MALDIquant and its associated R packages readBrukerFlexData and readMzXmlData are freely available from the R archive CRAN (http://cran.r-project.org). The software is distributed under the GNU General Public License (version 3 or later) and is accompanied by example files and data. Additional documentation is available from http://strimmerlab.org/software/maldiquant/.

Contact: mail@sebastiangibb.de

1 INTRODUCTION

Mass spectrometry profiling is increasingly becoming an important tool in clinical diagnostics, for example to identify biomarkers for cancer (e.g. Fiedler *et al.*, 2009). Similarly as with other high-throughput technologies, sophisticated statistical algorithms are essential in the analysis of spectrometry data (Morris *et al.*, 2010).

We have developed MALDIquant to provide a complete open-source analysis pipeline on the R platform (R Development Core Team, 2012) comprising all steps from importing of raw data, preprocessing (e.g. baseline removal), peak detection and non-linear peak alignment to calibration of mass spectra. MALDIquant is written as a standalone package using S4 object-oriented programming to facilitate further extension.

MALDIquant was initially developed for clinical proteomics using Matrix-Assisted Laser Desorption/Ionization (MALDI) technology. However, the algorithms implemented in MALDIquant are generic and may be equally applied to other 2D mass spectrometry data.

2 DISTINCTIVE FEATURES

In comparison with related R packages for mass spectrometry analysis, MALDIquant features a number of unique capabilities.

In particular, it implements a sophisticated non-linear peak alignment algorithm (He *et al.*, 2011; Wang *et al.*, 2010) as well as a calibration procedure for normalization of peak intensities across spectra that are modeled on a related method for sequence count data (Anders and Huber, 2010). In addition, MALDIquant allows to analyze technical replicates and spectra with unequal resolution, a crucial feature in clinical mass spectrometry where spectra from multiple sources need to be compared.

3 DETAILS ON ALGORITHMS

An example workflow for mass spectrometry analysis using MALDIquant is depicted in Fig. 1, starting with a raw unprocessed MALDI spectrum (A), followed by smoothing, baseline correction and peak detection (B), local alignment of peaks across spectra by warping (C–E) and merging and visualization (F). In the following, we briefly provide some background on the respective algorithms.

3.1 Data import

MALDIquant is carefully designed to be independent of any specific mass spectrometry hardware. Nonetheless, native input of binary data files (as well as complete folder hierarchies) from Bruker flex series instruments and input of the mzXML data format is supported through the associated R packages readBrukerFlexData and readMzXmlData.

3.2 Data preprocessing

For preprocessing spectral data, MALDIquant offers a complete set of routines for smoothing, variance stabilization, baseline correction and peak detection. MALDIquant implements several approaches to adjust the baseline and uses per default the SNIP algorithm (Ryan *et al.*, 1988) that returns a smooth baseline and leads to positive corrected intensities (Fig. 1B).

3.3 Peak alignment

For comparison of peaks across different spectra, it is essential to conduct alignment. In order to match peaks belonging to the same mass, MALDIquant uses a statistical regression-based approach combining the algorithms of He *et al.* (2011) and Wang *et al.* (2010). Specifically, first landmark peaks are identified that occur in most spectra. Subsequently, a non-linear warping function is computed for each spectrum by fitting a local regression

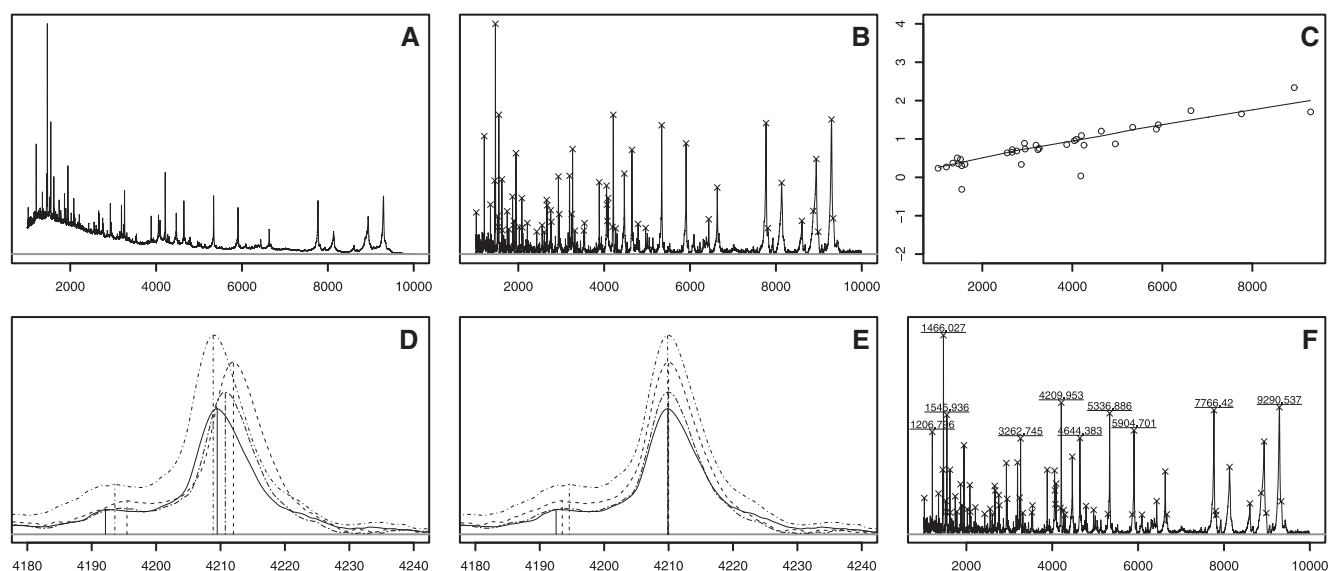[*]To whom correspondence should be addressed.

Fig. 1. Example of MALDIquant output: **(A)** raw spectrum; **(B)** variance-stabilized, smoothed and baseline-corrected spectrum with detected peaks; **(C)** fitted warping function for peak alignment; **(D)** four unaligned peaks; **(E)** four aligned peaks; and **(F)** merged spectrum with detected and labeled peaks.

to the matched reference peaks (Fig. 1C–E). This also allows to merge aligned spectra from technical replicates. An example of a merged spectrum with identified and labeled peaks is shown in Fig. 1F.

3.4 Calibration

Quantitative analysis of multiple spectra, e.g. to detect differentially expressed peaks, requires calibration. In order to render peak intensities comparable across spectra, a suitable scale factor for each individual spectrum needs to be determined. Experimentally, quantification of intensities is performed by reference to spike-in samples. In absence of spike-ins, MALDIquant offers a way of calibrating relative intensities by adapting an algorithm for calibrating next generation sequencing data (Anders and Huber, 2010). In this procedure first a reference spectrum is created using the median intensity of aligned peaks from all spectra. Subsequently, a scale factor is computed for each spectrum by using a robust estimator of the overall ratio of the peak intensities of the uncalibrated spectrum versus the reference spectrum. Additionally, calibration based on total ion current is available.

3.5 Classification and feature selection

Finally, the resulting calibrated peak intensity matrix may be exported for further use in high-level statistical analysis, for instance classification and feature selection using shrinkage discriminant analysis (Ahdesmäki and Strimmer, 2010).

4 CONCLUSION

`MALDIquant` is a versatile R package providing a flexible analysis pipeline for MALDI-TOF and other mass spectrometry data. It offers a number of distinctive features, in particular for alignment by non-linear warping and simultaneous calibration of peak intensities.

An overview of its capabilities is given by running the included demo script

```
library(``MALDIquant'')
demo(``MALDIquant'')
```

REFERENCES

Ahdesmäki,M. and Strimmer,K. (2010) Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Statist.*, **4**, 503–519.

Anders,S. and Huber,W. (2010) Differential expression for sequence count data. *Genome Biol.*, **11**, R106.

Fiedler,G.M. *et al.* (2009) Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer. *Clin. Cancer Res.*, **15**, 3812–3819.

He,Q.P. *et al.* (2011) Self-calibrated warping for mass spectra alignment. *Cancer Inform.*, **10**, 65–82.

Morris,J.S. *et al.* (2010) Statistical contributions to proteomic research. *Methods Mol. Biol.*, **641**, 143–166.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing.*. R Foundation for Statistical Computing Vienna, Austria.

Ryan,C.G. *et al.* (1988) SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrument. Meth. B*, **34**, 396–402.

Wang,B. *et al.* (2010) DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Anal. Chem.*, **82**, 5069–5081.