ANALYSIS OF PROTEOMIC DATA USING MALDIQUANT

Sebastian Gibb[1] *and Korbinian Strimmer*[1]

[1]Institute for Medical Informatics, Statistics and Epidemiology (IMISE),
University of Leipzig, Härtelstr. 16–18, D-04107 Leipzig, Germany
Contact: sebastian.gibb@studserv.uni-leipzig.de

ABSTRACT

MALDI-TOF is a well established technology for mass spectrometric profiling of proteomic data. Here, we introduce the MALDIquant R package that implements an analysis pipeline for quantitative analysis of clinical MALDI-TOF data. We provide a brief overview of its current and planned capabilities. We provide an outline of the standard steps in the analysis of MALDI-TOF data and specifically discuss baseline correction algorithms implemented in our software. MALDIquant is freely available from the R archive CRAN and is distributed under the GNU General Public License.

1. INTRODUCTION

Proteomic mass spectrometric analyzes are now starting to be used in clinical diagnostics. For example, Hand [1] employed mass spectrometric profiling to determine proteomic biomarkers for breast cancer, and Alexandrov and colleagues [2] recently report a study concerned with mass spectrometric analysis of colorectal cancer tissue.

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) is a common technique in proteomic mass spectrometry. A particular feature of MALDI-TOF is that a "matrix" is used in the ionization of the probe which allows the measurement of large biomolecules that would otherwise be difficult to ionize. MALDI-TOF hardware (e.g. the Bruker *flex series) is accompanied by commercial special purpose analysis programs. However, due to the closed nature of this software it is difficult to control — and if needed modify and adjust — all steps of the statistical analysis. The great success of biostatistical and bioinformatics tools for gene and genome chips on the CRAN and Bioconductor R [3] archives shows that an open source analysis pipeline for MALDI-TOF data is highly desirable.

Here we present MALDIquant, an effort to establish a simple to use yet highly effective set of R functions for quantitative analysis of MALDI-TOF data. In the following we first provide an outline of standard steps in the analysis of mass spectrometric data. Subsequently, we compare for illustration of the capabilities of MALDIquant a variety of baseline removal algorithms. Finally, we provide an outlook on the future development of MALDIquant.

2. ANALYSIS STEPS

A standard workflow in proteomic mass spectrometry analysis is depicted in Fig. **1** — see also [4] for a much more detailed discussion.

The first step after input of the raw data (Fig. **2** top row) consists of smoothing and baseline removal. MALDI-TOF data often suffer from strong distortions due to effects from the matrix, i.e. the physical carrier of the proteins. This may lead to a pronounced deviation of the baseline from the zero line. Second, after correction of baseline effects peaks need to be identified (Fig. **2** bottom row). A third step comprises calibration and normalization of peak intensities to enable quantitative comparison across multiple spectra. This is followed by peak alignment. In these steps technical replicates may be merged so that only the biological variability remains. Finally, after preprocessing statistical analysis such as classification or biomarker identification is conducted.

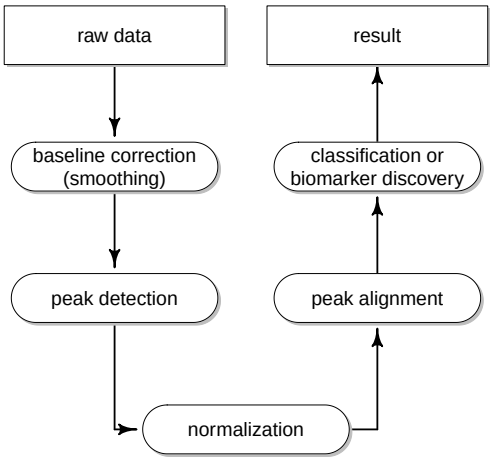

Figure 1. General workflow in mass spectrometric analysis.

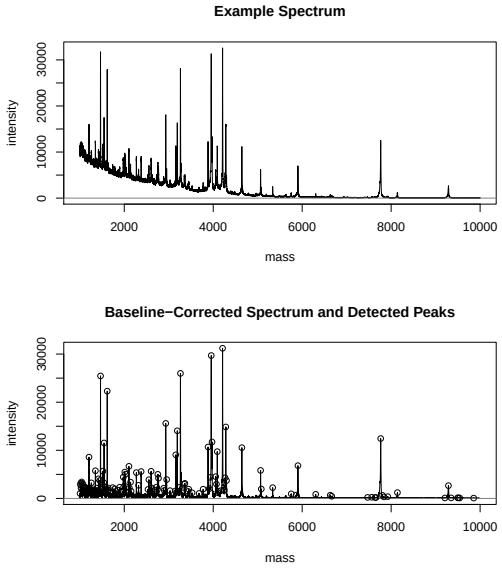

Figure 2. Raw protein spectrum (top row) versus baseline removed spectrum with identified peaks.

For each of the steps above a number of statistical algorithms are available. An open platform such as R facilitates a customized and modular analysis fitting the data at hand, without the constraints of a commercial software.

3. BASELINE CORRECTION

In the following we discuss the first step, baseline correction, of a typical MALDI-TOF analysis. Fig. **3** shows the effect of four different baseline removal algorithms on the raw spectrum of Fig. **2** (top row).

One of the simplest and widely used strategies is to estimate the baseline by computing the median of the intensities in a moving window — see the first row of Fig. **3**. The resulting baseline is not smooth and may also lead to negative intensity values in the corrected spectrum.

The PROcess algorithm [5] is a commonly used procedure. It starts by dividing a spectrum in equal parts. In each block all intensity values are set to the respective minimum intensity, and subsequently a loess curve is fitted to determine the estimated baseline. Similarly as the moving median baseline the PROcess algorithm may result in negative corrected intensities. In addition to window size PROcess requires an extra smoothing parameter for the loess algorithm.

Another useful strategy for baseline removal is to compute the convex hull of the spectrum via monotonic regression [6]. This is shown in the third row of Fig. **3**. This procedure guarantees a positive corrected spectrum and also provides a smooth baseline. Another advantage of the convex hull algorithm is that it is completely parameter-free.

A further baseline correction approach is the SNIP algorithm [7]. SNIP is an iterative algorithm that computes the baseline by computing the local minima and the local mean intensities in windows of increasing size. This approach returns a smooth baseline and leads to positive corrected intensities, cf. Fig. **3** last row. The SNIP algorithm requires specification of a single window size parameter.

Generally, if the matrix effect in a spectrum is weak both SNIP and the convex hull algorithms lead to similar baselines. However, if there is a pronounced mode in the spectrum, as in Fig. **4**, then the convex hull completely fails to provide a satisfactory baseline. Therefore, in MALDIquant we selected the SNIP procedure as the default algorithm for baseline estimation.

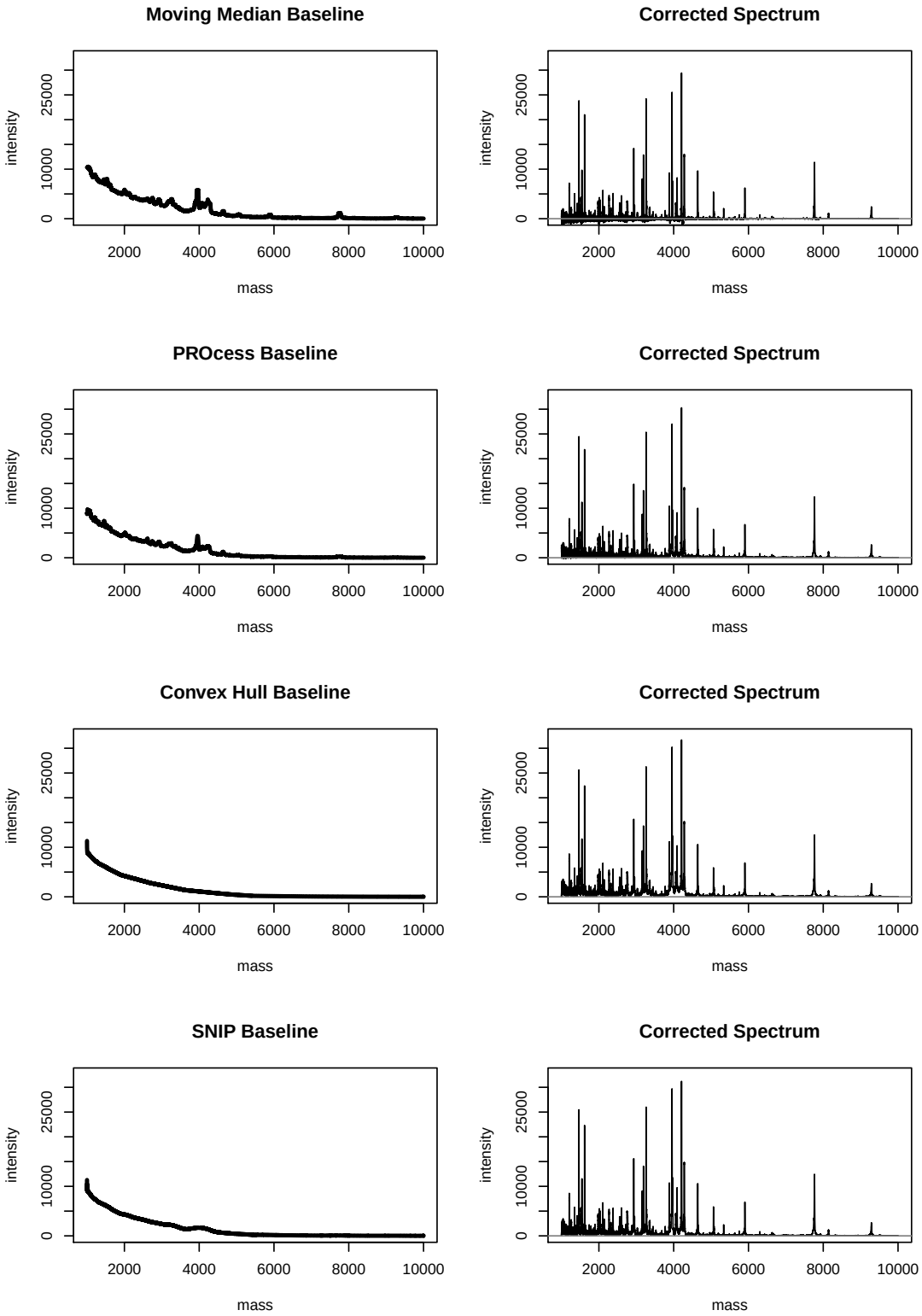

Figure 3. Comparison of four different baselines removal algorithms computed for the raw spectrum of Fig. **2**.

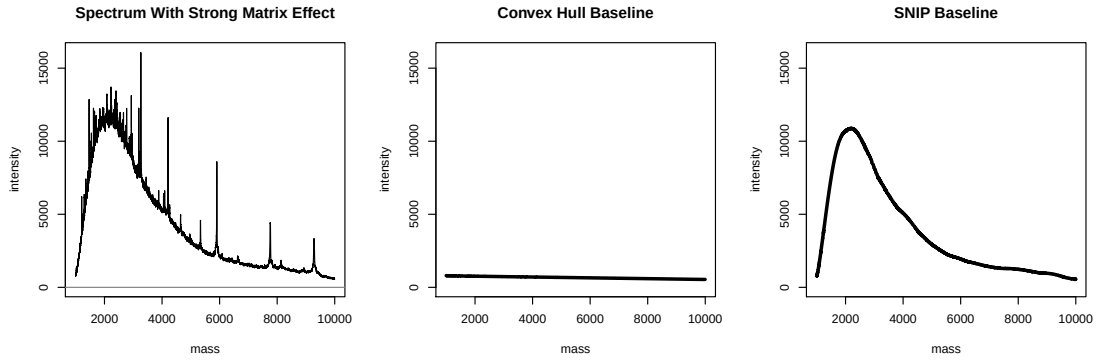

Figure 4. For spectra with pronounced matrix effect the convex hull baseline algorithm is not appropriate.

4. OTHER FEATURES

In addition to the implementation of various baseline removal algorithms MALDIquant supports both native input of binary data files (and complete folder hierarchies) from Bruker *flex series instruments and input of the mzXML data format, and offers procedures for data visualization. Under active development are functions for peak picking, peak alignment, merging of technical replicates, calibration of relative intensity scales. Furthermore, simple interfaces for subsequent multivariate analysis, e.g., classification are provided. More details, specifically on the performance of the implemented calibration algorithm, will be reported elsewhere.

5. CONCLUSION

MALDIquant is an R package that facilitates the analysis of proteomic mass spectrometric data. It provides a set of simple to use functions that allow a complete processing of MALDI-TOF from raw data to processed intensity values useful for clinical diagnostics and biomarker detection.

MALDIquant is available from the R archive CRAN under the GNU General Public License. The software is accompanied by example files demonstrating the application of MALDIquant.

6. ACKNOWLEDGMENTS

We thank Alexander Leichtle for providing MALDI-TOF example data and for many helpful discussions, and Verena Zuber for critical reading of the manuscript.

References

[1] D. J. Hand, "Breast cancer diagnosis from proteomic mass spectrometry data: a comparative evaluation," *Statist. Appl. Genet. Mol. Biol.*, vol. 7, pp. 15, 2008.

[2] T. Alexandrov, J. Decker, B. Mertens, A. M. Deelder, R. A. E. M. Tollenaar, P. Maass, and H. Thiele, "Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation.," *Bioinformatics*, vol. 25, pp. 643–649, 2009.

[3] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0.

[4] J. S. Morris, K. A. Baggerly, H. B. Gutstein, and K. R. Coombes, "Statistical contributions to proteomic research," *Methods in Molecular Biology*, vol. 641, pp. 143–166, 2010.

[5] X. Li, "PROcess: Ciphergen SELDI-TOF processing," Bioconductor R package archive, 2005, R package version 1.26.0.

[6] T. Robertson, F. T. Wright, and R. L. Dykstra, *Order restricted statistica inference*, John Wiley and Sons, 1988.

[7] C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie, and D. R. Cousens, "SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications," *Nucl. Instrument. Meth. B*, vol. 34, pp. 396–402, 1988.