rMSIproc: an R package for mass spectrometry imaging data processing

Pere Ràfols[1,2*], Bram Heijs[3,4], Esteban del Castillo[1], Oscar Yanes[1,2], Liam A. McDonnell[3,4,5], Jesús Brezmes[1,2], Iara Pérez-Taboada[2,6], Mario Vallejo[2,6] María García-Altares[1,2] and Xavier Correig[1,2]

[1]Department of Electronic Engineering, Rovira i Virgili University, IISPV, Tarragona, Spain, [2]Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain, [3]Center for Proteomics & Metabolomics, Leiden University Medical Center, Leiden, The Netherlands, [4]Department of Pathology, Leiden University Medical Center, Leiden The Netherlands, [5]Fondazione Pisana per la Scienza ONLUS, Pisa, Italy, [6]Instituto de Investigaciones Biomédicas Alberto Sols, Consejo Superior de Investigaciones Científicas (CSIC)/Universidad Autónoma de Madrid, Spain.

Abstract

Summary: Mass spectrometry imaging (MSI) can reveal biochemical information directly from a tissue section. MSI generates a large quantity of complex spectral data which is still challenging to translate into relevant biochemical information. Here we present rMSIproc, an open-source R package that implements a full data processing workflow for MSI experiments performed using TOF or FT-based mass spectrometers. The package provides a novel strategy for spectral alignment and recalibration, which allows to process multiple datasets simultaneously. This enables to perform a confident statistical analysis with multiple datasets from one or several experiments. rMSIproc is designed to work with files larger than the computer memory capacity and the algorithms are implemented using a multi-threading strategy. rMSIproc is a powerful tool able to take full advantage of modern computer systems to completely develop the whole MSI potential.

Availability and Implementation: rMSIproc is freely available at https://github.com/prafols/rMSIproc

Contact: pere.rafols@urv.cat

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Mass spectrometry imaging (MSI) is capable of mapping the spatial distributions of molecular ions in biological tissues(Caprioli *et al.*, 1997). The size and complexity of MSI data requires specialized software to extract relevant information. Several software packages, reviewed in (Ràfols, Vilalta, *et al.*, 2018), have been released to address these demands. More recently, new releases of MSiReader (Bokhart et al., 2017) and SpectralAnalysis (Race *et al.*, 2016) have been published. Both programs are Matlab written packages featuring intuitive graphical tools, and are freely available. The major limitation of MSIreader is that the full dataset is loaded into computer memory, which impedes processing a dataset larger than the available memory. SpectralAnalysis (Race *et al.*, 2016) overcomes the memory limitations and provides spectral processing algorithms. In comparison to these Matlab-based software solutions, the R platform is a truly open alternative that allows a straightforward modification and combination of different tools. Indeed, the last version of the R-based Cardinal package has improved the data model in order to handle larger-than-memory datasets (Bemis and Vitek, 2017). Nevertheless, Cardinal does not provide a graphical user interface (GUI) to explore the MSI data. None of these software tools have exploited the full potential of multicore processors by designing an optimized MSI data processing workflow capable of distributing the workload across available processors to maximize the processing throughput. In addition, none of these packages provides a mechanism able to combine various MSI datasets in the same data structure enabling a straightforward, robust and simple statistical analysis of multiple biological samples.

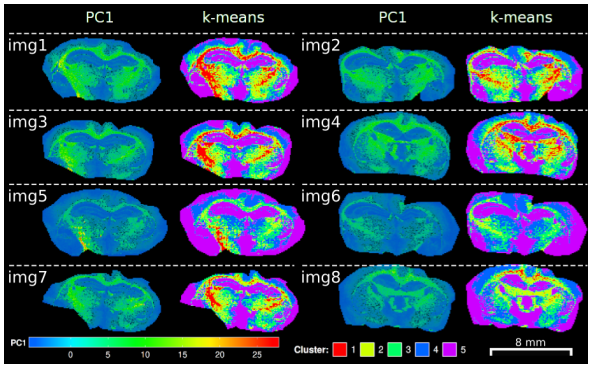

Fig. 1. Overview of rMSIproc results in a dataset containing eight mouse brain tissue sections acquired in the same batch. Spatial distribution of the first principal component (**PC1**) of a PCA and a standard **k-means** clustering algorithm calculated using the complete peak matrix of the eight tissue sections.

Here we present rMSIproc, an open-source MSI data processing R package to complement the previously released rMSI package (Ràfols *et al.*, 2017). The rMSI package was designed to allow efficient access to large MSI datasets combined with a data visualization GUI. rMSIproc expands on this by providing a data processing workflow designed to extract the relevant mass-to-charge (*m/z*) features from multiple MS images. To achieve this goal, rMSIproc automatically corrects pixel-to-pixel mass shifts for detected *m/z* features across MSI data using our novel spectral alignment algorithm (Ràfols, Castillo, *et al.*, 2018). Lastly, rMSIproc reduces the data processing time using a multi-thread approach designed to take advantage of modern multicore processors whilst keeping a low memory usage and controlled disk access.

2 Features

The main goal of the rMSIproc package is to process MSI data to produce a reduced peak matrix that is: (i) a robust representation of the relevant information contained by the full dataset, and (ii) small enough to fit within the computer's memory. This data reduction enables the efficient statistical analysis of larger-than-memory MSI datasets to be performed directly in R, and reduces the contribution due to chemical background. The processed spectral data is stored along the raw files allowing the rMSI visualization tools to enhance ion image reconstruction. To achieve these goals,

rMSIproc implements a spectral processing workflow including: mass spectra smoothing and alignment, *m/z* recalibration, normalization of intensities, peak detection and binning. Moreover, rMSIproc can merge and process various datasets simultaneously, producing a single peak matrix from all of the aligned mass spectra. The processing workflow can be easily adjusted to adapt it to all possible MSI data particularities. More details are available in the Supplementary Information Appendix 1 and Fig. S1.

All the user-relevant methods of rMSIproc are exposed as R functions following the structure of a standard R package but unique characteristics designed for computational efficiency are included as well. The internals of rMSIproc are mainly implemented in C++ language to provide efficient memory management and highly optimized multi-threading execution. This enables the end user to easily integrate rMSIproc routines in their own R scripts whilst keeping a high CPU performance. At the bottom level, the rMSI package is used to efficiently handle MSI data. Therefore, the same data formats as rMSI are supported. This includes the open-standard format imzML (Schramm *et al.*, 2012) in both 'continuous' and 'processed' mode.

3 Results

Several MSI datasets up to 200 gigabytes in size have been processed using rMSIproc. In all cases we obtained a balanced CPU load distributed across all processing cores. The memory consumption was managed by exclusively loading the data chunk being processed at each given time. We observed a processing performance that increases with the number of processing threads up to a point where the HDD throughput becomes the bottleneck. This suggests that MSI data processing performance will certainly benefit from modern solid-state disk technologies. The computational performance of rMSIproc is reported in Fig. S2.

To demonstrate rMSIproc's data merging capabilities we processed a dataset acquired with a MALDI-TOF mass spectrometer containing eight coronal mouse brain tissue sections (experimental details are available in Supplementary Information Appendix 2). This resulted in a unified peak matrix containing the relevant *m/z* features for the eight tissue sections. Along with the peak matrix we stored the peak lists before and after the spectral alignment stage. This allowed us to graphically represent the mass shifts observed at each *m/z* feature retained in the peak matrix (see mass alignment comparisons in Fig. S3-S11). The alignment algorithm has proven to be able to compensate for *m/z* shifts in both TOF-MS and FT-MS datasets. After the alignment, all mass spectra shared a common mass axis and were re-calibrated together. Our alignment routine could properly resolve isobaric *m/z* species in ultra-high mass resolution MALDI-FTICR datasets that were otherwise impossible to detect accurately (an example is provided in Fig. S12). Finally, to demonstrate the integration of rMSIproc with R, a PCA analysis and a k-means clustering was performed using the resulting peak matrix and R's built-in functions (Fig. 1). The reliable and straightforward rMSIproc workflow facilitates the biological interpretation of the MSI data.

4 Conclusions

rMSIproc is a valuable tool for efficiently processing MSI files containing both high mass and high spatial resolution in the R environment. It enables the possibility to combine multiple biological samples in the same MSI workflow allowing to conduct a confident statistical analysis using a higher number of tissue sections. This will surely help to establish MSI as powerful technique for biological and clinical research. The combination of rMSI and rMSIproc provides a full MSI data visualization and processing platform that uses modern computers in a novel and open-source manner.

Funding

This work has been supported by the Spanish Ministry of Economy and Competitiveness through project TEC2015-69076-P, PR´s predoctoral grant No. BES-2013-065572, project BFU2017-89336-R, the Direcció General de Recerca of the Government of Catalonia through project 2017 SGR 1119and IPT's fellowship from the Spanish Ministry of Education, Culture and Sports gran No. FPU 14/04457.

Conflict of Interest: none declared.

References

Bemis,K.A. and Vitek,O. (2017) matter: an R package for rapid prototyping with larger-than-memory datasets on disk. *Bioinformatics*, **33**, 3142–3144.

Bokhart,M.T. *et al.* (2017) MSiReader v1.0: Evolving Open-Source Mass Spectrometry Imaging Software for Targeted and Untargeted Analyses. *J. Am. Soc. Mass Spectrom.*, 1–9.

Caprioli,R.M. *et al.* (1997) Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.*, **69**, 4751–4760.

Race,A.M. *et al.* (2016) SpectralAnalysis: Software for the Masses. *Anal. Chem.*, **88**, 9451–9458.

Ràfols,P., Castillo,E. del, *et al.* (2018) Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer. *Anal. Chim. Acta*, **1022**, 61–69.

Ràfols,P. *et al.* (2017) rMSI: an R package for MS imaging data handling and visualization. *Bioinformatics*, **33**.

Ràfols,P., Vilalta,D., *et al.* (2018) Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications. *Mass Spectrom. Rev.*, **37**, 281–306.

Schramm,T. *et al.* (2012) imzML — A common data format for the flexible exchange and processing of mass spectrometry imaging data. *J. Proteomics*, **75**, 5106–5110.