The Journal of Open Source Software

subMALDI: an open framework R package for processing irregularly-spaced mass spectrometry data

Kristen Yeh[1], **Sophie Castel**[4], **Naomi L. Stock**[2], **Theresa Stotesbury**[3], **and Wesley Burr**[4]

1 Forensic Science Program, Trent University **2** Water Quality Center, Trent University **3** Faculty of Science, Forensic Science & Applied Bioscience, Ontario Tech University **4** Faculty of Science, Mathematics, Trent University

DOI: 10.21105/joss.02694

Software

- Review
- Repository
- Archive

Editor: Amy Roberts

Reviewers:

- @jspaezp
- @sigven
- @tystan

Submitted: 31 August 2020
Published: 10 September 2021

Summary

Mass spectrometry (MS) is an essential analytical technique used in many fields of science, including chemistry, biology, medicine, and more (Gross, 2011). Its uses are varied, from biotechnology studies of biomolecular sequencing (Maux et al., 2001), genetic analysis of human DNA (Null et al., 2001), exploration of the structure of single cells (Jones et al., 2003) and even examination of extraterrestrial objects (Fenselau & Caprioli, 2003). This incredible breadth of applications using MS results in highly complex data, which often requires significant processing in order to obtain actionable insights.

Modern instrumentation often includes proprietary software for spectral processing and analysis (e.g. Bruker Daltonics' Data Analysis). These tools, though convenient, often fail to provide sufficient documentation of the algorithms employed in the software and have limited analytical capabilities. Other commercial tools are available to supplement these programs (e.g. Agilent Technologies' MassHunter Profinder and Thermo Scientific's SIEVE^{TM}), however, they come at a cost. Open source software for analysis of MS data is also available online. These applications are often implemented in a variety of statistical computing languages, including Python (e.g. pyOpenMS) (Rost et al., 2014), Matlab (e.g. LIMPIC) (Mantini et al., 2007), C++ (e.g. ProteoWizard) (Chambers et al., 2012) and R (e.g. MSnbase, MALDIquant) (Gatto & Lilley, 2012; Gibb & Strimmer, 2012). While more accessible and well-documented than proprietary software, these available open source applications (Gibb, 2016) often utilize complex data structures (e.g. S3 and S4 class objects in R), which can make it difficult for researchers without strong coding backgrounds to access their raw spectral data. In order to simplify the organization and processing of mass spectrometry data, we propose the R package `subMALDI`.

`subMALDI` is an open framework tool that permits organization, pre-processing (smoothing, baseline correction, peak detection), and normalization of spectral data sets without masking into S3 or S4 class objects. After every step of processing, the m/z and intensity data of each spectrum is readily accessible, providing researchers with a more thorough understanding of the data manipulation that occurs during analysis. As a result of the package's open framework, subMALDI data sets are compatible with functions from a wide variety of other R packages, and user-defined functions are easier to implement and test.

Statement of Need

`subMALDI` permits the direct comparison of irregularly spaced spectral replicates in an open framework, an important feature that other open source tools do not contain. While matrix-assisted laser desoprtion/ionization (MALDI) mass spectra (and, also, any single spectra

aquired data) are often visualized on a continuous scale, the data observed are positive intensity values, corresponding to discretely measured mass-to-charge (m/z) values (Stanford et al., 2016). When spectral replicates are acquired of a sample, there is variation in the number and value of m/z responses with accompanying peaks due to spectra centroiding in the mass analyzer. This results in irregularly spaced data. This has implications for the statistical interpretations of inter-and intra-sample comparisons. In order to generate meaningful results from unevenly spaced data, it is essential that the data set be standardized by some means. In statistical computing languages, replicates often must be aligned against the same data structure: for our purposes, this will be the default data structure in R, the data.frame (Wickham, 2014).

`subMALDI` processes each raw spectrum with one of several smoothing filters, baseline correction methods, and peak detection algorithms included in the package. The processed spectral intensity values are then aligned to an array of all the theoretically possible m/z values in the observed mass range, at a specified resolution. The resulting data frame contains all m/z data in the first column, with the intensity data of each spectral replicate in adjacent columns.

`subMALDI` was designed for use by researchers who wish to organize, process, and analyze single spectra data, particularly MS data, while still being able to access their raw data at various points throughout the process. It has been utilized in a scientific article in the *Journal of Forensic Chemistry* (Yeh et al., 2020) and in our laboratory for analysis of MALDI-MS and electrospray-ionization (ESI) MS data. The open framework format and data structures of `subMALDI` create a more transparent pipeline for processing of MS data, where users can easily access their raw data and better understand the processing algorithms that are being executed on their data sets. The `subMALDI` framework is intended to reduce the "black-box" characteristics of MS data analysis and assist students and researchers in obtaining a more thorough understanding of MS and the complex, diverse data sets that it is used to produce.

Acknowledgement

We are grateful to the Canadian Foundation for Innovation, and the Ontario Research Fund for funding the Bruker SolariX XR MALDI FT-ICR-MS in the Water Quality Centre at Trent University.

Funding

This work was supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant to W. Burr (2017-04741) and a Vice President Research Fund to T. Stotesbury (2019). Author K. Yeh was supported by two NSERC Undergraduate Student Research Awards (USRA), 2019 and 2020.

References

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M., Paulse, C., Creasy, D., … Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*, *30*, 918–920. https://doi.org/10.1038/nbt.2377

Fenselau, C., & Caprioli, R. (2003). Mass Spectrometry in the Exploration of Mars. *J Mass Spectrom*, *38*, 1–10. https://doi.org/10.1002/jms.396

Gatto, L., & Lilley, K. S. (2012). MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing, and quantitation. *Bioinformatics*, *28*, 288–289. https://doi.org/10.1093/bioinformatics/btr645

Gibb, S. (2016). *Open source tools for mass spectrometry analysis*. Online, via http://www.strimmerlab.org/notes/mass-spectrometry.html.

Gibb, S., & Strimmer, K. (2012). MALDIquant: a versatile R pacakge for the analysis of mass spectrometry data. *Bioinformatics*, *28*, 2270–2271. https://doi.org/10.1093/bioinformatics/bts447

Gross, J. H. (2011). *Mass Spectrometry: A Textbook*. Springer. ISBN: 978-3-642-10709-2

Jones, J. J., Stump, M. J., Fleming, R. C., Lay, J. O., & Wilkins, C. L. (2003). Investigation of MALDI-TOF and FT-MS Techniques for Analysis of Escherichia coli Whole Cells. *Anal Chem*, *75*, 1340–1347. https://doi.org/10.1021/ac026213j

Mantini, D., Petrucci, F., Pieragostino, D., DelBoccio, P., Nicola, M. D., Ilio, C. D., Federici, G., Sacchetta, P., Comani, S., & Urbani, A. (2007). LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bionformatics*, *8*, 101. https://doi.org/10.1186/1471-2105-8-101

Maux, D., Enjalbal, C., Martinez, J., Aubagnac, J. L., & Combarieu, R. (2001). Static Secondary Ion MS to Monitor Solid-Phase Peptide Synthesis. *J Am Mass Spectrom*, *12*, 1099–1105. https://doi.org/10.1016/s1044-0305(01)00296-3

Null, A. P., Hannis, J. C., & Muddiman, D. C. (2001). Genotying of Simple and Compound Short Tandem Repeat Loci Using Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal Chem*, *73*, 4514–1521. https://doi.org/10.1021/ac0103928

Rost, H. L., Schmitt, U., Aebersold, R., & Lars, M. (2014). pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics*, *14*, 74–77. https://doi.org/10.1002/pmic.201300246

Stanford, T. E., Bagley, C. J., & Solomon, P. J. (2016). Informed baseline subtraction of proteomic mass spectrometry data aided by a novel sliding window algorithm. *Proteome Sci*, *14*, 19. https://doi.org/10.1186/s12953-016-0107-8

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(1), 1–23. https://doi.org/10.18637/jss.v059.i10

Yeh, K., Burr, W. S., Stock, N. L., & Stotesbury, T. (2020). Preliminary analysis of latent fingerprints recovered from underneath bloodstains using matrix-assisted laser desorption/ionization fourier-transform ion cyclotron resonance mass spectrometry imaging (MALDI FT-ICR MSI). *Forensic Chemistry*, *20*, 100274. https://doi.org/10.1016/j.forc.2020.100274