IsoMatchMS: Open-Source Software for Automated Annotation and Visualization of High Resolution MALDI-MS Spectra

David J. Degnan[1,*], Kevin J. Zemaitis[2], Logan A. Lewis[1], Lee Ann McCue[1], Lisa M. Bramer[1], James M. Fulcher[2], Dušan Veličković[2], Ljiljana Paša-Tolić[2], Mowei Zhou[2,*]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99354, USA, and [2]Environmental Molecular Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99354, USA.

*David J. Degnan: david.degnan@pnnl.gov and Mowei Zhou: mowei.zhou@pnnl.gov

Mass spectrometry. Isotope profile. MALDI-MS. Trelliscope. R package.

ABSTRACT: Due to its speed, accuracy, and adaptability to various sample types, matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) has become a popular method to identify molecular isotope profiles from biological samples. Often MALDI-MS data does not include tandem MS fragmentation data, and thus the identification of compounds in samples requires external databases so that the accurate mass of detected signals can be matched to known molecular compounds. Most relevant MALDI-MS software tools developed to confirm compound identifications are focused on small molecules (*e.g.,* metabolites, lipids), and cannot be easily adapted to protein data due to their more complex isotopic distributions. Here, we present an R package called *IsoMatchMS* for the automated annotation of MALDI-MS data for multiple datatypes: intact proteins, peptides, and glycans. This tool accepts already derived molecular formulas, or for proteomics applications, can derive molecular formulas from a list of input peptides or proteins including proteins with post-translational modifications. Visualization of all matched isotopic profiles are provided in a highly accessible HTML format called a trelliscope display, which allows users to filter and sort by several parameters such as match scores and the number of peaks matched. *IsoMatchMS* simplifies the annotation and visualization of MALDI-MS data for downstream analyses.

INTRODUCTION

MALDI-MS (matrix-assisted laser desorption/ionization mass spectrometry) is a popular technology for characterizing molecular compounds and their location within biological systems (e.g., tissues, microbial colonies, etc.) due to its high sensitivity and high throughput, relative to other spatially resolved MS technologies.[1, 2] In a typical workflow, peak assignments are based on matching the experimental masses to databases of masses for known molecules, either from parallel experiments or *in silico* prediction. To precisely match an entry in one of these databases, high resolution data with isotopic profiles are preferrable for high confidence identifications, henceforth referred to as annotations. Of particular interest for MALDI-MS analysis are proteins and their modified forms (proteoforms), which have roles in epigenetics (gene regulation), cell signaling, protein degradation, and signal transduction.[3, 4] However, only recently have developments addressing ion transmission and resolving power limitations[5-9] enabled isotopically resolved MALDI-MS spectra for intact proteins (~<20 kDa). Therefore, most existing software for annotating MALDI-MS data is largely focused on molecules with relatively small isotopic profiles (*e.g.,* lipids, metabolites),[10-12] with limited support for peptides (i.e., digested proteins).[13-16] Consequently, users interested in peptide and protein MALDI-MS data often must manually combine the capabilities of several different tools for peak assignment, which is a time-consuming process.

Here, we developed an R package called *IsoMatchMS* (https://github.com/PNNL-HubMAP-Proteoform-Suite/IsoMatchMS) to support both the analysis of MALDI-MS data with molecular formula annotation, and the analysis of intact protein and peptide MALDI-MS data from ProForma[17] strings. *IsoMatchMS* derives molecular formulas from ProForma[17] strings or the modification formats from five different proteomics identification tools: MSPathFinder[14], ProSightPC[18], pTop[19], TopPIC[20], and MS-GF+.[21] Then, *IsoMatchMS* calculates full theoretical isotope profiles, matches them to a summed spectrum, and visualizes the overlap between the two in an HTML display (called a "trelliscope[22] display") where each plot is an overlay of a single isotope profile for a molecule on the summed experimental spectra. Users can then sort the display by high-scoring distributions to confirm annotations. Unique to existing open-source software tools with functions to identify high-quality annotations,[10-16] *IsoMatchMS* supports spectral summing, common proteomics datatypes (both peptide and intact protein), complex compounds with modifications and unknown mass shifts, any adducts with a known mass, and visualizes the results in highly shareable and sortable trelliscope displays. By unifying the

analysis process for multiple types of biological molecules into a singular pipeline, *IsoMatchMS* reduces the time and effort for the annotation of MALDI-MS data. Trelliscope visualizations rank all possible annotations from highest to lowest quality, based on Pearson correlation, and support user-defined cut-off values.

EXPERIMENTAL SECTION

IsoMatchMS was written in R[23] 4.2.2 and has a main pipeline function which requires three inputs: 1) molecular formulas or ProForma[17] strings, 2) MS peak data, and 3) a settings file (an excel .xlsx). The molecular formulas or ProForma[17] strings need only be a vector of characters in R and thus can be read from any file format that can be read into R. Similarly, the MS peak data is a *pspecterlib*[24] peak_data object that can be created from any two numeric vectors (*m/z* and intensity), or extracted from a mzML or a ThermoFisher raw file with a *pspecterlib*[24] wrapper function, which uses *mzR*[25] and *rawrr*.[26] The settings file specifies several parameters such as *m/z* range and *m/z* error; examples are provided within the package. The run_isomatchms() function uses these objects to automatically perform the isotope calculations, peak matching, and trelliscope display generation.

The ProForma[17] strings can originate from experimental (e.g., LC-MS/MS) or protein sequence (e.g., UniProt[27]) databases. If ProForma strings are not provided by an identification tool such as TopPIC[20], *IsoMatchMS* contains a function to derive them from mzid files generated by tools such as MS-GF+[21] or from various modification formats, including those used in the tools MSPathFinder[14], ProSightPC[18], and pTop.[19] Molecular formulas are also accepted in lieu of ProForma strings, written simply as element and number of atoms with no spaces (*e.g.,* "C69H118N18O24S1"). MS1 spectra can be summed with other software (e.g., instrument vendor software), or calculated from mzML files with *IsoMatchMS.* Within the settings files, users can set a variety of parameters to optimize performance, including the isotope algorithm to use (either *Rdisop*[28] or *isopat*[29]), range of *m/z* values to search, an abundance noise filter, a minimum number of isotopic peaks to identify, a maximum number of charge states to investigate, and PPM and abundance tolerances to be considered a "match." For unknown mass shifts, the isotopic distribution is calculated for the derived formula, and then adjusted by the mass shift of interest. The example settings files in *IsoMatchMS* contain more details about each parameter, along with suggested defaults. For intact protein datasets, we suggest a minimum of 5 identified isotopes, higher noise filters (2.5-10%), and using the *Rdisop*[28] isotoping algorithm due to its higher precision at heavier masses. For peptides and glycans, we suggest a minimum of 2 identified isotopes and lower noise filters (0-2.5%). Available isotope scores include Pearson correlation and the absolute relative error (1) where A_M is the measured abundance, A_C is the calculated abundance, and n is the number of peaks.

$$\frac{1}{n} * \sum \frac{|A_M - A_C|}{A_M} \quad (1)$$

For higher mass profiles (intact proteomics), we suggest a Pearson correlation of 0.95 or greater, and 0.7 or greater otherwise. Alongside the trelliscope display, the main pipeline exports csv files with molecular formulas and masses, all matched isotope peaks, and the processed mass spectra. Details regarding MS acquisition of example datasets can be found in Supporting Information.

RESULTS AND DISCUSSION

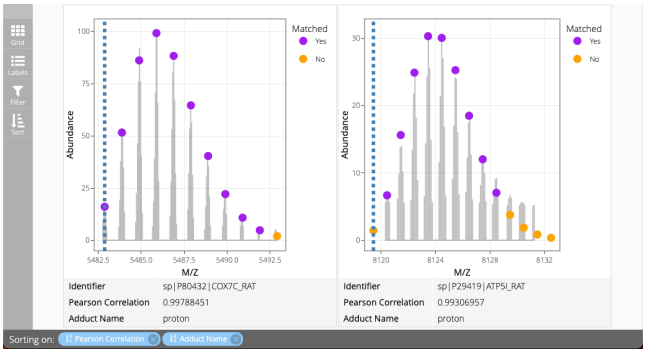


Figure 1. Trelliscope display of the isotopic distributions of two relevant protein species from the intact protein dataset. The blue line is the monoisotopic mass, and the black lines are the experimental mass spectra. The points represent the theoretical isotope distribution where the purple points fit isotopes, and the orange do not.

IsoMatchMS results, in which isotope profiles were calculated, matched, and visualized for intact protein, peptide and glycan datasets, are described below. All trelliscope displays of results can be found at https://pnnl-hubmap-proteoform-suite.github.io/IsoMatchMS_Trelliscope_Examples/.

Intact Protein. Of the possible 12,750 annotations, 52 had a Pearson correlation greater than 0.95, 482 were greater than 0.9, and 878 were greater than 0.7 (Table S1). Proteins at correlation scores less than 0.95 tended to have more missing peaks or lower intensities due to limitations in MALDI-MS technologies relative to LC-MS/MS, specifically the reduced dynamic range of MALDI-MS (Table S2). Many of the isotope distributions with Pearson correlations below 0.95 had high levels of noise, thus we recommended capping correlation scores at higher values and manually reviewing top-scoring isotopic distributions. Note that *IsoMatchMS* does not collapse isotope distributions that are similar due to their combinations of molecular formulas, mass shifts, and adduct masses. Instead, all possible combinations are returned.

Peptides. From 7782 possible annotations, 54 were identified at a Pearson correlation of 0.95 or higher, 58 at 0.9, and 151 at 0.7 (Table S1). Isotope distribution with correlation scores less than 0.9 tended to have lower intensities and thus only matched one peak, typically the monoisotopic peak (Table S3). The reliability of these "single match" annotations and their biological relevancy can be determined by the user when visualized in the trelliscope display.

Glycans. Of 1766 possible annotations, 18 were identified at a Pearson correlation of 0.95 or higher, 21 at a correlation of 0.9 or higher, and 29 at a correlation of 0.7 or higher (Table S1). Similar to the peptide dataset, there is a correlation score

threshold where isotope distributions tend to have lower intensities and a single matched peak (Table S4).

Overall, based on these studies, we have provided suggested defaults for each parameter for intact protein, peptide, and glycan datasets within the R package. Proper parameter selection depends on the expertise of the user and their knowledge of their own data.

CONCLUSION

Because manual assignment and validation of molecular isotopic envelopes is labor-intensive and error-prone, we built *IsoMatchMS*, an open-source R package, to identify high quality annotations from high resolution MALDI-MS data. *IsoMatchMS* combines the steps of a bioinformatics workflow, including sequence to molecular formula conversion, the addition of known molecular formulas, mass shifts, and adducts, and isotopic matching, to provide fast and reproducible annotation of high-resolution peptide, intact protein, and glycan MALDI-MS data. Limitations to this approach include a dependence on the quality of upstream identification tools, and an inability to separate biomolecules with identical or highly similar isotopic profiles. Future development could focus on fine-tuning the isotope matching score to account for these issues. Overall, *IsoMatchMS* provides a critical first screening effort for MALDI-MS data, and the trelliscope displays support necessary manual verification of annotations.

AUTHOR INFORMATION

Author Contributions

LPT and MZ conceptualized and managed the project; KJZ, LMB, JMF, DV, and MZ contributed to the tool and its design; LAM conceptualized the software and supervised development; DJD designed the tool; LAL contributed to the code and documentation; and DJD and KJZ wrote the first draft of the manuscript. All authors have edited and given approval to the final version of the manuscript.

Funding Sources

This research was funded by the National Institutes of Health (NIH) Common Fund, Human Biomolecular Atlas Program (HuBMAP) grant UG3CA256959-01. A portion of this work was performed under project doi.org/10.46936/staf.proj.2020.51770/60000309 (LPT) at the Environmental Molecular Science Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830. The authors declare no competing financial interest.

SUPPORTING INFORMATION

Descriptions for mass spectrometry data acquisition and annotation of intact protein, peptide, and glycan datasets; (Supplemental_File_1.docx)

IsoMatchMS performance on intact protein, peptide, and glycan datasets; full results from the isotope matching algorithm for intact protein, peptide, and glycan data (Supplemental_File_2.xlsx)

A tutorial on how to use trelliscope displays for *IsoMatchMS* output (Supplemental_File_3.mp4)

ACKNOWLEDGEMENTS

We thank Camryn Pettenger-Wiley for her assistance in acquiring the peptide dataset.

REFERENCES

(1) Taylor, M. J.; Lukowski, J. K.; Anderton, C. R. Spatially Resolved Mass Spectrometry at the Single Cell: Recent Innovations in Proteomics and Metabolomics. *J Am Soc Mass Spectrom* **2021**, *32* (4), 872-894. DOI: 10.1021/jasms.0c00439.

(2) Singhal, N.; Kumar, M.; Kanaujia, P. K.; Virdi, J. S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front Microbiol* **2015**, *6*, 791. DOI: 10.3389/fmicb.2015.00791.

(3) Forgrave, L. M.; Wang, M.; Yang, D.; DeMarco, M. L. Proteoforms and their expanding role in laboratory medicine. *Pract. Lab. Med.* **2022**, *28*, e00260, DOI: 10.1016/j.plabm.2021.e00260.

(4) Lee, M. J.; Yaffe, M. B. Protein Regulation in Signal Transduction. *Cold Spring Harb. Perspect. Biol.* **2016**, *8*(6), DOI: 10.1101/cshperspect.a005918.

(5) Zemaitis, K. J.; Velickovic, D.; Kew, W.; Fort, K. L.; Reinhardt-Szyba, M.; Pamreddy, A.; Ding, Y.; Kaushik, D.; Sharma, K.; Makarov, A. A.; et al. Enhanced Spatial Mapping of Histone Proteoforms in Human Kidney Through MALDI-MSI by High-Field UHMR-Orbitrap Detection. *Anal Chem* **2022**, *94* (37), 12604-12613. DOI: 10.1021/acs.analchem.2c01034.

(6) Spraggins, J. M.; Rizzo, D. G.; Moore, J. L.; Rose, K. L.; Hammer, N. D.; Skaar, E. P.; Caprioli, R. M. MALDI FTICR IMS of Intact Proteins: Using Mass Accuracy to Link Protein Images with Proteomics Data. *J Am Soc Mass Spectrom* **2015**, *26* (6), 974-985. DOI: 10.1007/s13361-015-1147-5.

(7) Nicolardi, S.; Switzar, L.; Deelder, A. M.; Palmblad, M.; van der Burgt, Y. E. Top-down MALDI-in-source decay-FTICR mass spectrometry of isotopically resolved proteins. *Anal Chem* **2015**, *87 (6)*, 3429-3437. DOI: 10.1021/ac504708y.

(8) Prentice, B. M.; Ryan, D. J.; Van de Plas, R.; Caprioli, R. M.; Spraggins, J. M. Enhanced Ion Transmission Efficiency up to m/ z 24 000 for MALDI Protein Imaging Mass Spectrometry. *Anal Chem* **2018**, *90(8)*, 5090-5099. DOI: 10.1021/acs.analchem.7b05105.

(9) Han, J.; Permentier, H.; Bischoff, R.; Groothuis, G.; Casini, A.; Horvatovich, P. Imaging of protein distribution in tissues using mass spectrometry: An interdisciplinary challenge. TrAC Trends in *Anal Chem* **2019**, *112*, 13-28.

(10) Tortorella, S.; Tiberi, P.; Bowman, A. P.; Claes, B. S. R.; Scupakova, K.; Heeren, R. M. A.; Ellis, S. R.; Cruciani, G. LipostarMSI: Comprehensive, Vendor-Neutral Software for Visualization, Data Analysis, and Automated Molecular Identification in Mass Spectrometry Imaging. *J Am Soc Mass Spectrom* **2020**, *31(1)*, 155-163. DOI: 10.1021/jasms.9b00034.

(11) Goracci, L.; Tortorella, S.; Tiberi, P.; Pellegrino, R. M.; Di Veroli, A.; Valeri, A.; Cruciani, G. Lipostar, a Comprehensive Platform-Neutral Cheminformatics Tool for Lipidomics. *Anal Chem* **2017**, *89(11)*, 6257-6264. DOI: 10.1021/acs.analchem.7b01259.

(12) Palmer, A.; Phapale P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandorov, T. FDR-controlled metabolite annotation for high-resolution imaging mass spectroemtry. *Nat Methods* **2017**, *14,* 57-60.

(13) Greer, J. B.; Early, B. P.; Durbin, K. R.; Patrie, S. M.; Thomas, P. M.; Kelleher, N. L.; LeDuc, R. D.; Fellers, R. T. ProSight Annotator: Complete control and customization of protein entries in UniProt XML files. *Proteomics* **2022**, *22(11-12)*, e2100209. DOI: 10.1002/pmic.202100209.

(14) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; et al. Informed-Proteomics: open-source software package for top-down proteomics. *Nat Methods* **2017**, *14(9)*, 909-914. DOI: 10.1038/nmeth.4388.

(15) Fornelli, L.; Srzentic, K.; Huguet, R.; Mullen, C.; Sharma, S.; Zabrouskov, V.; Fellers, R. T.; Durbin, K. R.; Compton, P. D.; Kelleher, N. L. Accurate Sequence Analysis of a Monoclonal Antibody by Top-Down and Middle-Down Orbitrap Mass Spectrometry Applying Multiple Ion Activation Techniques. *Anal Chem* **2018**, *90(14)*, 8421-8429. DOI: 10.1021/acs.analchem.8b00984.

(16) Guo, G.; Papanicolaou, M.; Demarais, N. J.; Wang, Z.; Schey, K. L.; Timpson, P.; Cox, T. R.; Grey, A. C. Automated annotation and visualisation of high-resolution spatial proteomic mass spectrometry imaging data using HIT-MAP. *Nat Commun* **2021**, *12(1)*, 3241. DOI: 10.1038/s41467-021-23461-w.

(17) LeDuc, R. D.; Schwämmle, V.; Shortreed, M. R.; Cesnik, A. J.; Solntsev, S. K.; Shaw, J. B.; Martin, M. J.; Vizcaino, J. A.; Alpi, E.; Danis, P. ProForma: a standard proteoform notation. *J. Proteome Res.* **2018**, *17(3)*, 1321-1325.

(18) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res* **2007**, *35*, W701-706. DOI: 10.1093/nar/gkm37.

(19) Sun, R. X.; Wang, R. M.; Luo, L.; Liu, C.; Chi, H.; Zeng, W. F.; He, S. M. Accurate Proteoform Identification and Quantitation Using pTop 2.0. *Methods Mol Biol* **2022**, *2500*, 105-129. DOI: 10.1007/978-1-0716-2325-1_9.

(20) Kou, Q.; Xun, L.; Liu, X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **2016**, *32(22)*, 3495-3497. DOI: 10.1093/bioinformatics/btw398.

(21) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **2014**, *5*, 5277. DOI: 10.1038/ncomms6277.

(22) Hafen, R.; Shloerke, B. trelliscope js: Creative Interactive Trelliscope Displays. *R package version 0.2.6,* **2021**. https://CRAN.R-project.org/package=trelliscopejs (accessed 2023-05-01)

(23) Ihaka, R.; Gentleman, R. R: a language for data analysis and graphics. *JCGS* **1996**, *5(3)*, 299-314.

(24) Degnan, D. J.; Bramer, L. M.; White, A. M.; Zhou, M.; Bilbao, A.; McCue, L. A. PSpecteR: A User-Friendly and Interactive Application for Visualizing Top-Down and Bottom-Up Proteomics Data in R. *J Proteome Res* **2021**, *20(4)*, 2014-2020. DOI: 10.1021/acs.jproteome.0c00857.

(25) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012**, *30(10)*, 918-920. DOI: 10.1038/nbt.2377.

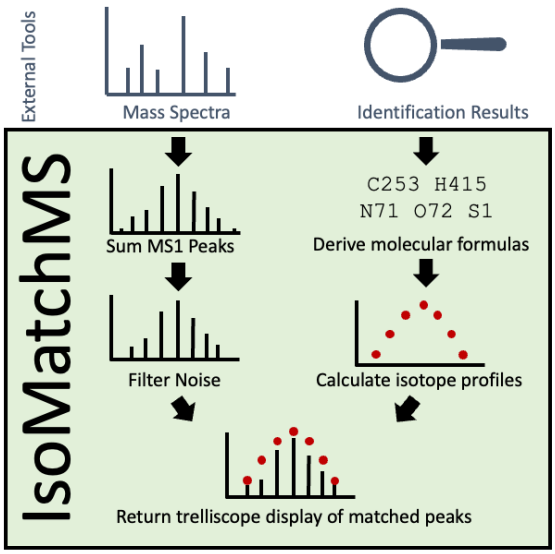
(26) Kockmann, T.; Panse, C. The rawrr R Package: Direct Access to Orbitrap Data and Beyond. *J Proteome Res* **2021**, *20(4)*, 2028-2034. DOI: 10.1021/acs.jproteome.0c00866.

(27) UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **2023**, *51(D1)*, D523-D531. DOI: 10.1093/nar/gkac1052.

(28) Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **2009**, *25(2)*, 218-224.

(29) Loos, M. isopat: Calculation of isotopic pattern for a given molecular formula. *R package version 1.0,* **2012**. https://CRAN.R-project.org/package=isopat (accessed 2023-05-01)

For TOC Only


External Tools
Mass Spectra
Identification Results
IsoMatchMS
Sum MS1 Peaks
C253 H415
N71 O72 S1
Derive molecular formulas
Filter Noise
Calculate isotope profiles
Return trelliscope display of matched peaks