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

David J. Degnan1,*, Kevin J. Zemaitis2, Logan A. Lewis1, Lee Ann McCue1, Lisa M. Bramer1, James M. Fulcher2, Dušan Veličković2, Ljiljana Paša-Tolić2, Mowei Zhou2,*

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


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 software tools 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 (e.g., intact proteins, lipids, metabolites, etc.). This tool accepts already computed molecular formulas, or for proteomics applications, can compute 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.1-3. 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 preferable for high confidence identifications. 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 limitations5-9 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 (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 ProForma17 strings. IsoMatchMS derives molecular formulas from ProForma17 strings or the modification formats from five different proteomics identification tools: MSPath-Finder14, ProSightPC18, pTop19, TopPIC20, 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 “treliscope” 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.
EXPERIMENTAL SECTION

Package Design. IsoMatchMS was written in R\textsuperscript{23} 4.2.2 and has a main pipeline function which requires three inputs: 1) molecular formulas or ProForma\textsuperscript{17} strings, 2) MS peak data, and 3) a settings file (an excel .xlsx). The molecular formulas or ProForma\textsuperscript{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 psecpecterlib\textsuperscript{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 psecpecterlib\textsuperscript{24} wrapper function, which uses mzR\textsuperscript{25} and rawrr.\textsuperscript{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\textsuperscript{17} strings can originate from experimental (e.g., LC-MS/MS) or protein sequence (e.g., UniProt\textsuperscript{27}) databases. If ProForma strings are not provided by an identification tool such as TopPIC\textsuperscript{28}, IsoMatchMS contains a function to derive them from mzid files generated by tools such as MS-GF\textsuperscript{21} or from various modification formats, including those used in the tools MSPathFinder\textsuperscript{14}, ProSightPC\textsuperscript{18}, and pTop.\textsuperscript{19} Molecular formulas are also accepted in lieu of ProForma strings, written simply as element and number of atoms with no spaces (e.g., “C2H5NO2”). 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\textsuperscript{28} or isopat\textsuperscript{29}), range of m/z values to search, an abundance noise filter, a minimum number of isotope peaks to identify, a maximum number of charge states to investigate, and PPM and abundance tolerances to be considered a “match.” 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\textsuperscript{28} isopating algorithm due to its higher precision at heavier masses. For smaller isotope profiles, 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.

\[
\frac{1}{n} \sum \frac{|A_M - A_C|}{A_M}
\]  

(1)

For higher mass profiles (intact proteomics), we suggest a Pearson correlation of 0.9 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 regardless of score, and the processed mass spectra.

MS Data. MALDI-MS data from intact protein and tryptic peptide analyses were collected from sections of a fresh-frozen and non-embedded female Sprague Dawley rat brains (Rattus norvegicus).\textsuperscript{30,31} Briefly, tissues were purchased from BioIVT (Westbury, NY) and cryosectioned as 10 μm sections onto either ITO (indium-tin oxide) for MALDI-MS or PEN (polyethylene napthalate)-membrane slides for laser capture microdissection. Briefly, intact proteins were analyzed after a tissue fixation and lipid depletion with a protocol, and tryptic peptides were analyzed after in-situ deposition of trypsin directly on the tissue and overnight incubation, this is described in great depth elsewhere.\textsuperscript{30,32} Glycan data was collected from formalin-fixed paraffin-embedded (FFPE) soybean nodules (Glycine max cv. Williams 82). Briefly, FFPE soybean nodules were cryosectioned as 7 μm sections onto ITO slides, where PNGase was applied and incubated using a protocol described in greater depth elsewhere.\textsuperscript{33} Both glycans and intact proteins were analyzed on a custom UHMR Q Exactive HF Orbitrap (Thermo Scientific, Bremen, DE) fitted with an elevated pressure MALDI source (Spectroglyph, Kennewick, WA) operated under a custom privilege license.\textsuperscript{32} Tryptic peptides were analyzed on a 12T solariX Fourier transform ion cyclotron resonance mass spectrometer (Bruker Daltonics, Bremen, DE). All raw MALDI-MS data were summed with Mozaic (v.2022.5.1.b1; Spectroswiss, Lausanne, CH). Experimental proteof orm/peptide databases were based on LC-MS/MS data from published results of intact\textsuperscript{30} and digested\textsuperscript{31} LC-MS/MS. All MS Data is included within the R package.

Table 1. IsoMatchMS performance on intact protein, peptide, and glycan datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Intact Protein</th>
<th>Peptide</th>
<th>Glycans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular formulas searched</td>
<td>2125</td>
<td>2594</td>
<td>3667</td>
</tr>
<tr>
<td>Adducts</td>
<td>Proton</td>
<td>Proton</td>
<td>Proton, Sodium</td>
</tr>
<tr>
<td>Number of manually verified high confidence annotations</td>
<td>21</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>Number of manual identifications also identified by IsoMatchMS at a Pearson correlation score of 0.70</td>
<td>21 / 21</td>
<td>28 / 28</td>
<td>26 / 26</td>
</tr>
<tr>
<td>Number of manual identifications also identified by IsoMatchMS at a Pearson correlation score of 0.90</td>
<td>19 / 21</td>
<td>27 / 28</td>
<td>21 / 26</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Number of manual identifications also identified by IsoMatchMS at a Pearson correlation score of 0.95</td>
<td>15 / 21</td>
<td>27 / 28</td>
<td>18 / 26</td>
</tr>
<tr>
<td>Seconds to complete using an Apple M1 Chip with 10 cores</td>
<td>72</td>
<td>54</td>
<td>33</td>
</tr>
</tbody>
</table>

**RESULTS AND DISCUSSION**

For each of the three tested datasets (intact protein, peptide, and glycan data), IsoMatchMS calculated, matched, and visualized isotope profiles in under 90 seconds (Table 1). Datasets where molecular formula information was already calculated from a theoretical database (glycans) completed faster than formulas derived from experimental ProForma\(^\text{17}\) strings. Across all datasets, highly confident annotations derived by using other tools\(^\text{12, 14}\) were also identified by IsoMatchMS at a Pearson score threshold of 0.7 (Table 1). Even at higher Pearson score thresholds of 0.9 and 0.95, there was high overlap between the number of manual and IsoMatchMS annotations. An example of the trelliscope display output for the intact protein data was provided (Fig. 1). While IsoMatchMS was able to annotate several dozen high-confidence isotopic profiles within the presented datasets, the number of matches per number of molecular formulae searched (Table 1) was relatively low. This was due to the reduced dynamic range of MALDI-MS detect several dozen to a couple hundred species in comparison to LC-MS/MS which can routinely detect several hundred intact proteins or several thousand digested peptides.

While most intact proteins detected by MALDI-MS can be readily assigned from experiment intact protein LC-MS/MS databases, the proteome coverage will never perfectly align even in the case of the best experimental databases generated from serial sections of tissue.

Regardless, manual assignment and validation of molecular isotopic envelopes can take up to several hours or more per analysis and require the use of multiple tools depending on the specific restrictions per tool and the aims of the analysis. Centralizing the steps of the bioinformatics workflow, which includes sequence to molecular formula conversion, the addition of known and unknown mass shifts, and isotopic matching, within IsoMatchMS allows for a preliminary analysis of high-resolution MALDI-MS with an experimental or protein sequence database within minutes. This improvement in speed allows for more high-throughput analyses of MALDI-MS data and reduces the knowledge burden associated with using several tools to process data for different types of molecules.

While the curation of an experimental or protein sequence database still requires sample dependent supplemental efforts, this tool vastly improves the throughput and reproducibility of MALDI-MS workflows for biomolecules. As more
experimental data become available in public repositories, scoring matches may be improved to the point of fully automated peak annotations.

CONCLUSION

IsoMatchMS is an open-source R package for identifying molecular annotations from high resolution MALDI-MS data. Trelliscope displays allow users to quickly sort through results to select interesting annotations for further study in downstream analyses, including differential abundance statistics or mass spectrometry imaging. Overall, IsoMatchMS reduces the time burden for identifying high quality annotations, and is an example of a flexible tool built for multiple biological molecules (e.g., metabolites, intact or digested proteins, lipids). IsoMatchMS provides a critical first screening effort for MALDI-MS data, but manual verification of annotations in the trelliscope display should still occur as settings are variable.

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-AC05-76RL1830. The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

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

REFERENCES


(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(9), 5990-5999. DOI: 10.1021/acs.analchem.7b05105.


IsoMatchMS

External Tools:

Mass Spectra

Identification Results

C253 H415
N71 O72 S1

Derive molecular formulas

Filter Noise

Calculate isotope profiles

Return trelliscope display of matched peaks