Bioinformatics, 2021, 1–3
doi: 10.1093/bioinformatics/btab282
Applications Note

Data and Text Mining

MassExplorer: a computational tool for analyzing desorption electrospray ionization mass spectrometry data

Vishnu Shankar [1,*], Robert Tibshirani[2] and Richard N. Zare[3]

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA, [2]Department of Statistics and Biomedical Data Science, Stanford University, Stanford, CA 94305, USA and [3]Department of Chemistry, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Abstract

Summary: In the last few years, desorption electrospray ionization mass spectrometry imaging (DESI-MSI) has been increasingly used for simultaneous detection of thousands of metabolites and lipids from human tissues and biofluids. To successfully find the most significant differences between two sets of DESI-MSI data (e.g., healthy vs disease) requires the application of accurate computational and statistical methods that can pre-process the data under various normalization settings and help identify these changes among thousands of detected metabolites. Here, we report MassExplorer, a novel computational tool, to help pre-process DESI-MSI data, visualize raw data, build predictive models using the statistical lasso approach to select for a sparse set of significant molecular changes, and interpret selected metabolites. This tool, which is available for both online and offline use, is flexible for both chemists and biologists and statisticians as it helps in visualizing structure of DESI-MSI data and in analyzing the statistically significant metabolites that are differentially expressed across both sample types. Based on the modules in MassExplorer, we expect it to be immediately useful for various biological and chemical applications in mass spectrometry.

Availability and implementation: MassExplorer is available as an online R-Shiny application or Mac OS X compatible standalone application. The application, sample performance, source code and corresponding guide can be found at: https://zarelab.com/research/massexplorer-a-tool-to-help-guide-analysis-of-mass-spectrometry-samples/.

Contact: vishnus1@cs.stanford.edu

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

1 Introduction

Desorption electrospray ionization mass spectrometry imaging (DESI-MSI) is increasingly used for simultaneous, rapid characterization of hundreds to thousands of lipids and metabolites from tissues and biofluids (Vijayalakshmi et al., 2020). DESI-MSI yields a chemical map by scanning the sample surface in the x and y directions using a spray of charged microdroplets. Each pixel (150–200 μm) provides a chemical fingerprint. DESI-MSI operates under ambient conditions and requires minimal sample preparation (Wu et al., 2013). Its applications range from understanding the effect of new drug treatments on rare diseases (Wert et al., 2020) to identifying positive margins for surgical excision of tumors (Vijayalakshmi et al., 2020). While successful application of DESI-MSI presents new possibilities for rapid chemical tissue analysis, one limiting challenge has been identifying the most important molecular changes among the many detected metabolites.

MassExplorer is a computational tool that has been optimized for identifying molecular differences between two groups of DESI-MSI data (e.g., healthy versus diseased). Our tool accelerates pre-processing of raw DESI-MSI data under various normalization settings, builds predictive statistical machine learning lasso models to identify the most important molecular differences, and helps interpret selected differences. To identify the important changes, supervised machine learning classification approaches have been increasingly employed, where models use the metabolite abundances in a mass spectra to predict the outcome of interest. Although a variety of classification approaches have been used to identify metabolic changes, including linear models (Song et al., 2020), support vector machines (Zhang and Liu, 2018) and random forests (Zhang and Liu, 2018), our tool uses the lasso, due to its empirical performance in many DESI-MSI applications, robustness to overfitting, selection of a sparse set of metabolites, and simple interpretability in understanding the selected metabolites' contribution

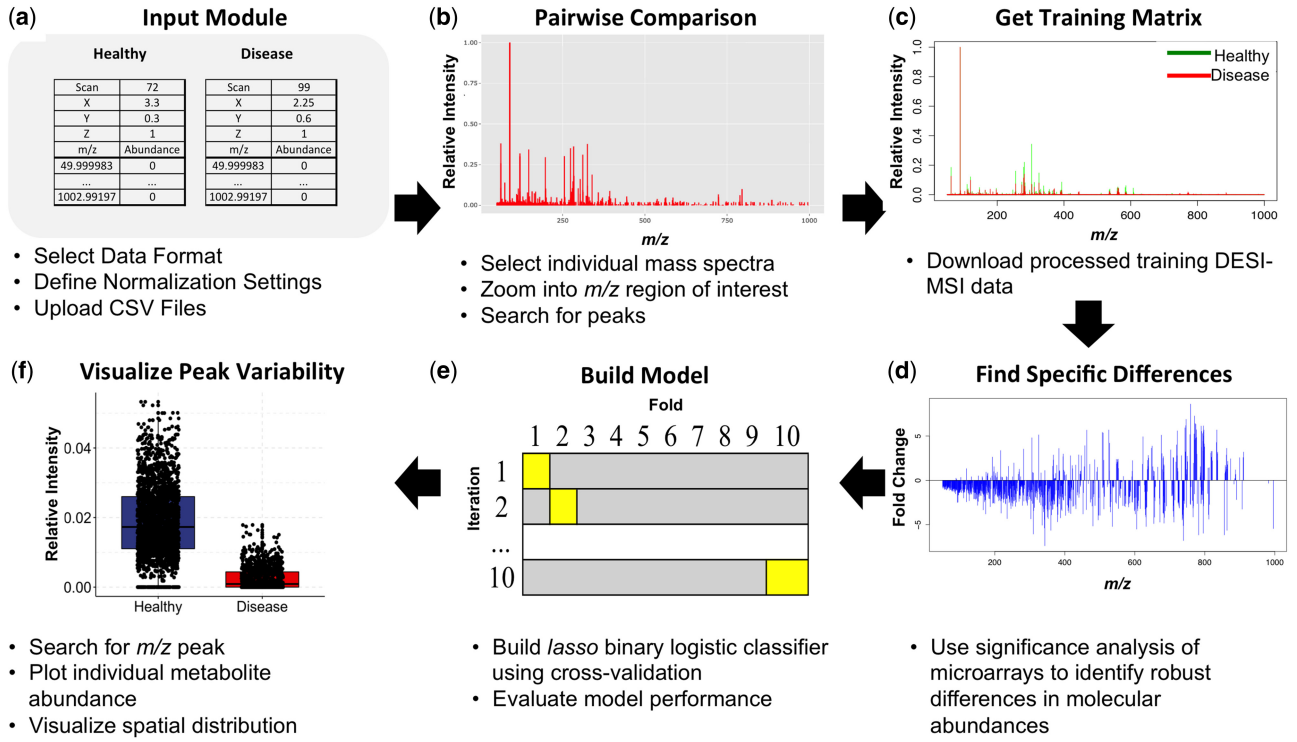

Fig. 1. Summary of MassExplorer workflow: (**a**) input two groups of raw mass spectrometry or DESI-MSI data; (**b**) pairwise comparison of individual spectra; (**c**) process raw mass spectrometry data to build training matrix with rows corresponding to a unique mass spectra identifier and columns to all detected metabolites; (**d**) identification of statistically significant molecular differences using significance analysis of microarrays (SAM); (**e**) lasso model selection through cross-validation; and (**f**) interpretation of selected model peaks

to the model predictions. Figure 1 presents a summary of the application modules and possible workflow.

2 Tool features

The "Input Module" accepts multiple raw mass spectrometry data files in comma-separated values (CSV) format in either "normal" or "imaging" mode, where "imaging" looks for mass spectra corresponding to a unique patient, *x*, *y* coordinate and scan identifier (Fig. 1). The input module includes various settings for processing the data, such as an internal standard peak to normalize the molecular abundances, tolerance to account for how much can the same peak vary between samples, and a threshold that can be used to exclude metabolites detected in only few samples.

The "Pairwise Comparison" module allows the user to select and visualize any individual mass spectra from both groups. This module's interactivity enables zooming into each spectrum, focusing the comparison between groups to a particular *m/z* region, and looking for specified chemical fragments in a searchable and sortable data table.

The "Get Training Matrix" module builds a matrix under the selected normalization settings, where the rows correspond to a unique patient, *x*, *y*, scan identifier and columns to all detected metabolites. This module gathers all the detected *m/z* fragments, discretizes the range by hierarchical clustering and cuts the resulting dendogram at the tolerance specified in the "Input" Module. The training matrix, which can be downloaded as a CSV, also includes a column for the "Disease State," indicating to which group each observation belongs.

The "Find Specific Differences," based on the samr (Tibshirani et al., 2015) package in Bioconductor, finds statistically significant molecular differences using significance analysis of microarrays (SAM) (Tusher et al., 2001), which calculates a modified *t*-statistic (i.e., q-statistic) for each metabolite from the assembled training matrix. As SAM accounts for the false discovery rate using permutation tests, it can accurately identify true underlying molecular differences that reflect biochemical signals.

The "Build Model" module uses the glmnet (Friedman et al., 2010) package in R to construct a binary logistic classifier, based on the lasso (Tibshirani, 1996) to help identify the most important differences. This approach constructs a simple, interpretable predictive model that only uses a small subset of detected metabolites. The tool can perform cross-validation, either 10-fold or leave one patient out for larger datasets. This module produces the cross-validation results, selected lasso coefficients and model performance metrics, including sensitivity, specificity, and accuracy.

The "Visualize Peak Variability" module allows the user to type any *m/z* of interest and produces a boxplot comparing the relative intensity pattern across both groups. This module produces a crude spatial map for each group, based on the maximum peak intensity at each *x*, *y* coordinate. Taken together, these plots can visualize how substantial are the selected molecular differences between the groups and what does the inherent variability within each group look like.

We believe this publicly available tool enables rapid processing and analysis of DESI-MSI data in a self-contained manner, where a user can visualize, model and interpret the data in real time. While there have been several tools for visualization and analysis of mass spectrometry data (many reviewed in Weiskirchen et al., 2019), our tool is the first that is specific to DESI-MSI technology users, which can assist in both visualizing and pre-processing raw data for building predictive models. Both MassExplorer and other mass spectrometry computational tools enable comparison of individual mass spectra and visualization of the spatial distribution of any particular *m/z* peak of interest. While MassExplorer does not have capabilities for peak identification, our self-contained tool uniquely enables users to work with high dimensional mass spectrometry datasets by building and interpreting predictive models. We recommend the use of MassExplorer for helping to focus on key signals among thousands of detected metabolites.

Acknowledgements

V.S. acknowledges discussions with members of the Zare lab.
Financial Support: V.S. acknowledges funding support from the Stanford Precision Health and Integrated Diagnostics Center seed grant and National Science Foundation [CHE-1734082].

Conflict of Interest: none declared.

References

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.
Song,X. *et al.* (2020) Oral squamous cell carcinoma diagnosed from saliva metabolic profiling. *Proc. Natl. Acad. Sci. USA*, **117**, 16167–16173.
Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.
Tibshirani,R. *et al.* (2015) Package 'samr'. Available at: https://cran.rproject.org/web/packages/samr/samr.pdf (11 October 2020, date last accessed).
Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.
Vijayalakshmi,K. *et al.* (2020) Identification of diagnostic metabolic signatures in clear cell renal cell carcinoma using mass spectrometry imaging. *Int. J. Cancer*, **147**, 256–265.
Weiskirchen,R. *et al.* (2019) Software solutions for evaluation and visualization of laser ablation inductively coupled plasma mass spectrometry imaging (LA-ICP-MSI) data: a short overview. *J. Cheminf.*, **11**, 16.
Wert,K.J. *et al.* (2020) Metabolite therapy guided by liquid biopsy proteomics delays retinal neurodegeneration. *EBioMedicine*, **52**, 102636.
Wu,C. *et al.* (2013) Mass spectrometry imaging under ambient conditions. *Mass Spectrom. Rev.*, **32**, 218–243.
Zhang,Y. and Liu,X. (2018) Machine learning techniques for mass spectrometry imaging data analysis and applications. *Bioanalysis*, **10**, 519–522.