

Systems biology

Miso: an R package for multiple isotope labeling assisted metabolomics data analysis

Yonghui Dong ¹, Liron Feldberg² and Asaph Aharoni^{1,*}

¹Department of Plant Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel and ²Department of Analytical chemistry, Israel Institute for Biological Research, Ness Ziona 7410001, Israel

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on October 28, 2018; revised on January 18, 2019; editorial decision on February 3, 2019; accepted on February 5, 2019

Abstract

Motivation: The use of stable isotope labeling is highly advantageous for structure elucidation in metabolomics studies. However, computational tools dealing with multiple-precursor-based labeling studies are still missing. Hence, we developed Miso, an R package providing automated and efficient data analysis workflow to detect the complete repertoire of labeled molecules from multiple-precursor-based labeling experiments.

Results: The capability of Miso is demonstrated by the analysis of liquid chromatography-mass spectrometry data obtained from duckweed plants fed with one unlabeled and two differently labeled tyrosine (unlabeled tyrosine, tyrosine-²H₄ and tyrosine-¹³C₉¹⁵N₁). The resulting data matrix generated by Miso contains sets of unlabeled and labeled ions with their retention time, *m/z* values and number of labeled atoms that can be directly utilized for database query and biological studies.

Availability and implementation: Miso is publicly available on the CRAN repository (<https://cran.r-project.org/web/packages/Miso>). A reproducible case study and a detailed tutorial are available from GitHub (https://github.com/YonghuiDong/Miso_example).

Contact: asaph.aharoni@weizmann.ac.il

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Molecular identification is one of the major challenges associated with mass spectrometry (MS)-based metabolomics (De Vijlder *et al.*, 2017; Shahaf *et al.*, 2016). Although the elemental composition of small molecules up to 500 Da can be directly determined using ultra-high-resolution and accurate mass measurement, it is often difficult and error-prone to identify the molecular structure due to the presence of a large number of structural isomers. For instance, a database search of chemical formula C₁₃H₁₈N₂O₄ in SciFinder (<https://scifinder.cas.org/>) revealed over 14 000 compounds possessing this elemental composition.

Stable isotopes have the same number of protons as common elements but differ in mass due to difference in the number of neutrons (Chokkathukalam *et al.*, 2014). The mass difference between isotopologues (unlabeled molecule and its labeled counterparts) could

precisely define the number of labeled atoms in each molecule. Stable isotope labeling is a promising approach for metabolite identification. In particular, labeling the metabolic precursor has been widely used to assist structure elucidation of metabolites. We have previously reported an *in vivo* stable isotope labeling strategy, termed Dual Labeling of Metabolites for Metabolome Analysis (DLEMMA), in which two differently labeled forms of the same molecule (or metabolite) were used as precursors (Feldberg *et al.*, 2009, 2018). Compared to single-precursor-based labeling approach, DLEMMA significantly reduces the number of candidates for a given elemental composition by matching the altered labeling patterns of the same analyte derived from two forms of differentially labeled precursor. Moreover, the dual labeling performed in DLEMMA can be extended to multiple labeled precursors which further improves the confidence in metabolite annotation.

A plethora of computational tools have been developed to detect isotopically labeled molecules in single-precursor-based labeling studies, such as mzMatch-ISO (Chokkathukalam *et al.*, 2013), X¹³CMS (Huang *et al.*, 2014), geoRge (Capellades *et al.*, 2016) and IsotopicLabelling (Ferrazza *et al.*, 2017). However, to our knowledge, computational solutions dealing with multiple-precursor-based labeling studies have not been reported. Hence, we developed Miso, an R package providing automated and efficient data analysis workflow to detect the complete repertoire of labeled molecules from multiple-precursor-based labeling experiments.

2 Materials and methods

Miso was initially developed for dual labeling in DLEMMA, yet, it can be easily applied to single- or multiple-precursor-based labeling studies. Miso enables detecting molecules labeled with various biologically relevant stable isotopes such as hydrogen (²H), carbon (¹³C), oxygen (¹⁸O), nitrogen (¹⁵N) and sulfur (³⁴S). The complete Miso workflow is demonstrated by the analysis of liquid chromatography-MS data from duckweed plants fed with one unlabeled and two differently labeled tyrosines (unlabeled tyrosine, tyrosine-²H₄ and tyrosine-¹³C₉¹⁵N₁), respectively (Fig. 1a). The raw data are converted into mzXML format using MSconvert (Kessner *et al.*, 2008) and pre-processed with XCMS R package (Smith *et al.*, 2006). The resulting XCMS peak-table is used as input for Miso package.

Instead of directly tracking stable isotopes by iterating over all MS signals using expected mass differences between unlabeled and labeled molecules, Miso first compares the MS signals among unlabeled and two differently labeled equivalent sample groups using Tukey's honest significance test. The ion intensity of an isotopically labeled *m/z* peak in the labeled group is expected to be significantly up-regulated (Fig. 1b). A default *P*-value threshold of 0.05 is used to filter out non-significant peaks. Additionally, to account for any possible statistical insignificance due to high variation among replicates, a fold change parameter (fold change = 10 by default) is used to retain those non-significant peaks if their ion intensity fold changes relative to the unlabeled and differently peaks are above this threshold (Fig. 1b). An alternative pre-filtering strategy is applied in cases when the experiment does not contain any replicates and/or the variations among replicates are extremely high. To achieve this, an ion intensity cutoff parameter is applied to set ion intensities of the 'background noise peaks' to zero when their ion intensities are lower than the cutoff value. A peak is then regarded as a labeled-precursor-derived peak only when it is detected in a minimum number of sample replicates (the minimum number is one if there are no replicates) from that specific labeled-precursor fed group (Fig. 1b). The pre-filter step efficiently eliminates false positives and reduces overall data analysis time.

Next, an accurate isotope filtering is performed following the assumption that the pre-filtered peaks from the labeled-precursor fed sample groups are all isotopically labeled. Miso then searches for their unlabeled counter ions in the unlabeled sample group according to the defined labeling patterns with a default retention time (RT) window of 6 s and a mass error of 10 ppm. This concept follows the paradigm that in MS analysis isotopologues are similar in RT but different in *m/z* values.

Miso generates two types of outputs. The first is a comprehensive data matrix that contains all detected isotopologue pairs with their RT, *m/z* and number of labeled atoms. Additionally, Miso provides an interactive MS spectrum for rapid visualization and comparison of the abundance of isotopologues (Fig. 1c).

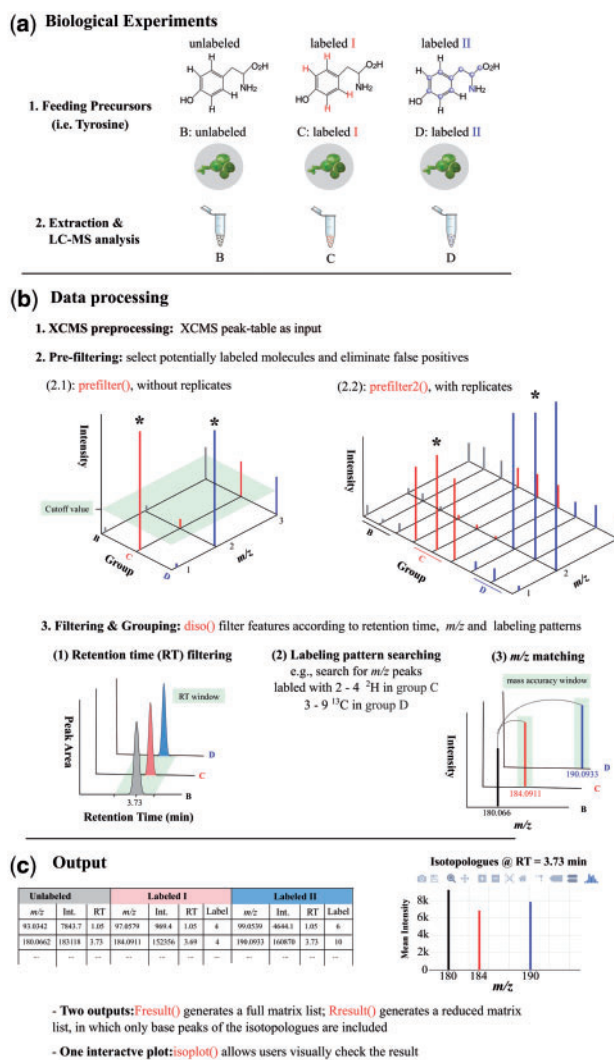


Fig. 1. Representative example of the Miso workflow. (a) Duckweed plants were fed for 24 h with three forms of the aromatic amino acid tyrosine: unlabeled, labeled I (tyrosine-²H₄) and label II (tyrosine-¹³C₉¹⁵N₁). The extracts were analyzed with liquid chromatography coupled to high resolution MS. (b) The raw data were converted into mzXML format and pre-processed with XCMS package. The resulting peak-table was used as input to automatically detect stable isotope labeled metabolites with Miso. (c) Two types of outputs were generated, which could be used for downstream analysis

3 Results

A step-by-step tutorial is provided in GitHub (https://github.com/YonghuiDong/MISO_example). Users can install Miso on CRAN with the function `install.packages("Miso")`. A sample dataset is provided along with the package and can be accessed with the function `data(lcms)`. The dataset was acquired in a DLEMMA experiment, in which three forms of tyrosine precursors were used to feed duckweed plants and track tyrosine-derived metabolites (Fig. 1a). The materials and methods are provided in [Supplementary Material](#). The dataset matrix is 161 333 × 22 in size, and the overall workflow requires ~2.2 min using a PC with 16 GB memory and a 3.1 GHz Intel Core i7 processor.

In total, 371 isotopologue sets (one unlabeled and two different labeled forms) were detected ([Supplementary Table S1](#)). However, due to the presence of natural isotopes and the possibility that the same

molecule could be labeled with different number of isotope elements, the result could be redundant for subsequent metabolite identification. It is worth noting that de-isotoping (the elimination of natural isotopic clusters) based on theoretical isotopic distribution is not effective in stable isotope labeling studies since the isotopic distribution of the labeled compound is significantly changed as compared to its unlabeled counterpart. To address this issue, sets containing only base peaks of all the isotopologues were kept. As a consequence, 79 isotopologues sets were selected. A careful manual inspection of the raw data showed that only six false positives remained in the reduced list, including three background noise peaks and three ^{13}C natural isotope peaks. In addition, two true sets were discarded as their isotopically labeled ions correspond to more than one unlabeled ions, and only the unlabeled form with highest ion intensity was kept. However, the missing true pairs can be easily identified and retrieved as their labeled forms are present in the reduced list (Supplementary Table S2).

To conclude, Miso is an efficient and easy-to-use R package allowing automated tracking of isotopically labeled metabolites in data derived from multiple-precursor-based stable isotope labeling studies. It provides a user-friendly output which can be used for further database queries and biological studies.

Funding

This work was supported by the Israel Ministry of Science and Technology [grant number 3-14297]. A.A. is the incumbent of the Peter J. Cohn Professorial Chair.

Conflict of Interest: none declared.

References

- Capellades, J. *et al.* (2016) geoRge: a computational tool to detect the presence of stable isotope labeling in LC/MS-based untargeted metabolomics. *Anal. Chem.*, **88**, 621–628.
- Chokkathukalam, A. *et al.* (2013) mzMatch-ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics*, **29**, 281–283.
- Chokkathukalam, A. *et al.* (2014) Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis*, **6**, 511–524.
- De Vijlder, T. *et al.* (2017) A tutorial in small molecule identification via electrospray ionization-mass spectrometry: the practical art of structural elucidation. *Mass Spectrom. Rev.*, **50**, 1234–1223.
- Feldberg, L. *et al.* (2018) DLEMMA-MS-Imaging for Identification of Spatially Localized Metabolites and Metabolic Network Map Reconstruction. *Anal. Chem.*, **90**, 10231–10238.
- Feldberg, L. *et al.* (2009) Dual labeling of metabolites for metabolome analysis (DLEMMA): a new approach for the identification and relative quantification of metabolites by means of dual isotope labeling and liquid chromatography–mass spectrometry. *Anal. Chem.*, **81**, 9257–9266.
- Ferrazza, R. *et al.* (2017) IsotopicLabelling: an R package for the analysis of MS isotopic patterns of labelled analytes. *Bioinformatics*, **33**, 300–302.
- Huang, X. *et al.* (2014) X13CMS: global tracking of isotopic labels in untargeted metabolomics. *Anal. Chem.*, **86**, 1632–1639.
- Kessner, D. *et al.* (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, **24**, 2534–2536.
- Shahaf, N. *et al.* (2016) The WEIZMASS spectral library for high-confidence metabolite identification. *Nat. Commun.*, **2016** 7, 12423.
- Smith, C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.