Bioinformatics, 33(15), 2017, 2419–2420
doi: 10.1093/bioinformatics/btx159
Advance Access Publication Date: 11 April 2017
Applications Note

MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data

Thomas Naake* and Emmanuel Gaquerel*

Centre for Organismal Studies, University of Heidelberg, Heidelberg 69120, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 21, 2016; revised on February 14, 2017; editorial decision on March 16, 2017; accepted on March 18, 2017

Abstract

Summary: Among the main challenges in metabolomics are the rapid dereplication of previously characterized metabolites across a range of biological samples and the structural prediction of unknowns from MS/MS data. Here, we developed `MetCirc` to comprehensively align and calculate pairwise similarity scores among MS/MS spectral data and visualize these across a range of biological samples. `MetCirc` comprises functionalities to interactively organize these data according to compound familial groupings and to accelerate the discovery of shared metabolites and hypothesis formulation for unknowns. As such, `MetCirc` provides a significant advance to address biological questions in areas where chemodiversity plays a role.

Availability and Implementation: `MetCirc`, implemented in the open-source `R` language, together with its vignette are available in the Bioconductor project and at https://github.com/PlantDefenseMetabolism/MetCirc.

Contact: thomasnaake@googlemail.com or emmanuel.gaquerel@cos.uni-heidelberg.de

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

1 Introduction

One of the often neglected bottlenecks in the high-throughput profiling of metabolites is the dereplication of identification knowledge. Critical to this procedure is the ability to confidently assign metabolite identity and to score similarities across metabolite-derived analytical signals. Such procedure, as it necessitates discriminating previously known from unknown metabolites, is also expected to provide insights for the annotation of unknowns. Tandem mass spectrometry (MS/MS) of metabolite-derived fragmentation patterns has become the method of choice to accomplish this task. Innovative methods are needed to organize the vast amount of data generated by modern MS/MS analytics, especially when the analysis is performed across a vast range of samples. Recently, the comprehensive analysis of MS/MS spectral similarity has emerged as a translational dereplication approach in metabolomics (Li *et al.*, 2015; Watrous *et al.*, 2012). Ideally, this approach should be employed to generate data-rich visual outputs that can be interactively navigated by non-bioinformatics users. Such data exploration should help to detect metabolites which are shared across the focal biological classes and more ‘specific’ ones for which *a priori* classification can be provided. To our knowledge, interactive visualizations fulfilling these criteria have not yet been developed. Inspired by the Circos visualization developed to align gene and syntenic blocks across genomes (Krzywinski *et al.*, 2009), we developed `MetCirc` for MS/MS similarity-based interactive analysis of within- and between-sample metabolic relatedness.

2 Available functionality/implementation

`MetCirc` is an `R` open-source analysis and interactive visualization tool for high-resolution MS/MS data. It is especially designed to organize data from comparative cross-organisms/-tissues experiments. The comprehensive calculation of pairwise similarities of MS/MS spectra within `MetCirc` is based on a normalized dot product (NDP Watrous *et al.*, 2012). `MetCirc` circular visualization is based on the `circlize` (Gu *et al.*, 2014) and the `shiny` (https://CRAN.R-project.org/package=shiny) packages. The developed tool represents *per se* a novel application since the original Circos visualization approach is currently restricted to the analysis of genomics and transcriptomics data and is not yet amenable for MS/MS metabolomics data. `MetCirc` is available via the Bioconductor project (https://bioconductor.org).

3 ‘A typical typeline’

The pipeline includes the creation of an MSP object based on fragmentation data, binning of fragment ions (mass-to-charge ratios referred to as *m/z* features), the calculation of the MS/MS similarity score (NDP) for assignment to a similarity matrix and the interactive visualization of this information using the `circlize` framework (Gu *et al.*, 2014).

Creation of the MSP object. The MSP object can be created from MSP files, typically used for MS/MS library building, or from other sources. The Supplementary file and the accompanying vignette describe several cases on how to create a MSP object.

Binning. Due to expected technical variations in *m/z* values across measurement (<1-2 ppm with modern high-resolution MS), these values are first binned according to a tolerance parameter. Finally, a matrix with binned fragment *m/z* values as columns, descriptors of MS/MS as rows and where entries are intensities in % is created via the following command:

```
binnedMSP <- binning(MSP)
```

Calculation of the MS/MS similarity matrix. Pairwise similarity between MS/MS spectra is calculated via a NDP score ranging from 0 to 1 (identical MS/MS). The NDP calculation uses binned *m/z* values and neutral losses corresponding to mass differences between the precursor and fragment ions within a spectrum. For a considered MS/MS pair, peak intensities of shared *m/z* values for precursor/fragment ions and neutral losses (here the intensity of the resulting fragment ion is used) are employed as weights $W_{S1,i}$ and $W_{S2,i}$ within the following NDP formula:

$$NDP = \frac{\sum_{i=1}^{j} (W_{S1,i} \cdot W_{S2,i})^2}{\sum_{i=1}^{j} (W_{S1,i}^2) \cdot \sum_{i=1}^{j} (W_{S2,i}^2)},$$

with S1 and S2 the spectra 1 and 2, respectively, of the i*th* of j common peaks differing by the tolerance parameter specified in `binning`. Weights are calculated according to $W = [peak\ intensity]^m \cdot [m/z]^n$, with m = 0.5 and n = 2 as default values as suggested by MassBank (Li *et al.*, 2015). The similarity matrix is created via:

```
similarity <- createSimilarityMatrix(binnedMSP)
```

Visualization. Interactive visualization is started by running:

```
selectedFeatures <- shinyCircos(similarity, MSP)
```

One of the key features of the interactive framework is that pre-defined biological classes can be compared (see Fig. 1). Classes specify the affiliation of a given MS/MS feature to any biological identifier relevant to the experiment conducted, e.g. organism, cell/tissue type, treatment/sampling time, etc. `shinyCircos`, which combines the `circlize` and `shiny` frameworks, is used to visualize pairwise MS/MS similarities within- and/or between biological classes. Upon selection of a MS/MS feature, the displayed auxiliary information can be updated by the user within the browser. NDP MS/MS similarity scores visualized as edges can be thresholded. Additionally, on the sidebar panel of the browser, the type of biological inference link to be displayed can be selected. MS/MS ordering (according to precursor *m/z*, retention time or from clustering by MS/MS similarity) within a biological class can be rapidly inter-changed.

4 Conclusion

In summary, `MetCirc` is a novel open-source package available via the `R`/Bioconductor project to make biological sense of mass spectral similarities from metabolomics data. MetCirc provides a dedicated data analysis infrastructure and visualization interface to explore small molecules that mediate functionally important phenotypes. Possible applications range from MS/MS-based metabolic pathway elucidation (in which biochemical conversions of metabolic intermediates are tracked in an unbiased manner by spectral similarity) to the diagnosis of metabolic markers in clinical metabolomics projects. Finally, stepwise data mining via `MetCirc` can be used to pinpoint and formulate first structural hypotheses on previously non-characterized metabolites associated with a given phenotype.

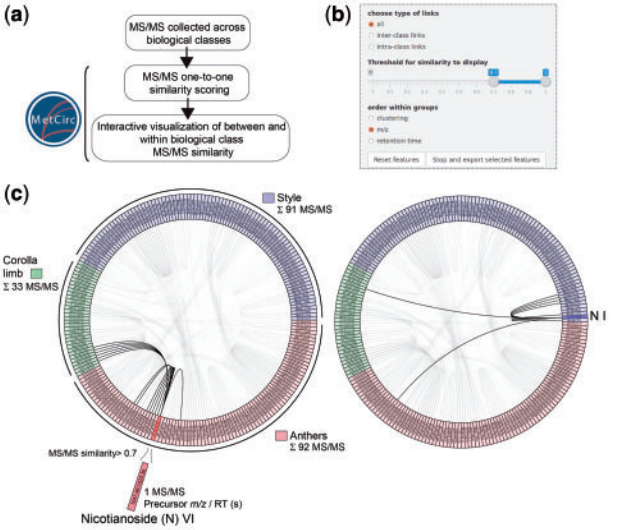

Fig. 1. Visualization of spectral similarity for MS/MS collected from flower organs of a wild tobacco species. (**a**) `MetCirc` workflow. (**b**) Bar within the browser to modulate MS/MS ordering, similarity thresholding and class belonging. (**c**) Selected are MS/MS for two previously identified metabolites: Nicotianoside VI (left panel) and Nicotianoside I (right panel). Edges for MS/MS similarity scores from 0.7 to 1 are visualized. Nicotianoside I is exclusively detected in the floral style data-set (no edge for a similarity of 1 connecting to other tissue types), but its spectrum shares high similarity scores with known and unknown metabolites in this (score of 0.8 with the previously identified Nicotianoiside XI) and the two other tissues. Nicotianoside VI, selected in the anther (pollen carrying tissue of the flower), is re-identified in the limb of the floral corolla and ‘connects’ several unknown metabolites

Acknowledgements

We thank the technical support of the Metabolomics Core Technology Platform (University of Heidelberg) and Dapeng Li (Max Planck Institute for Chemical Ecology) for help with compiling the MS/MS data.

Funding

Research of E.G. and T.N. in Heidelberg is supported within the framework of the Deutsche Forschungsgemeinschaft Excellence Initiative to the University of Heidelberg.

Conflict of Interest: none declared.

References

Gu,Z. *et al*. (2014) `circlize` implements and enhances circular visualization in `R`. *Bioinformatics*, **30**, 2811–2812.

Krzywinski,M.I. *et al*. (2009) Circos: An information aesthetic for comparative genomics. *Genome Res*., **19**, 1639–1645.

Li,D. *et al*. (2015) Navigating natural variation in herbivory-induced secondary metabolism in coyote tobacco populations using MS/MS structural analysis. *Proc. Natl. Acad. Sci. U. S. A*., **112**, E4147–E4155.

Watrous,J. *et al*. (2012) Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A*., **109**, E1743–E1752.