Open access • Posted Content • DOI:10.1101/546150

Implementations of the chemical structural and compositional similarity metric in R and Python — Source link

Asker Daniel Brejnrod, Madeleine Ernst, Piotr Dworzynski, Piotr Dworzynski ...+5 more authors

Institutions: University of Copenhagen, University of Montana, Statens Serum Institut, University of California, Berkeley ...+1 more institutions

Published on: 11 Feb 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Chemical similarity

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Implementations of the chemical structural and compositional similarity metric in R and Python

Asker Brejnrod [1,*], Madeleine Ernst [2,3], Piotr Dworzynski [1,4], Lasse Buur Rasmussen [1], Pieter C. Dorrestein [2,3,6,], Justin J.J. van der Hooft [5], Manimozhiyan Arumugam [1]

[1] Novo Nordic Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, 2200, Denmark
[2] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
[3] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
[4] Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
[5] Bioinformatics Group, Plant Sciences Group, Wageningen University, Wageningen, The Netherlands
[6] Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA 92093, USA

*brejnrod@sund.ku.dk

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Tandem mass spectrometry (MS/MS) has the potential to substantially improve metabolomics by acquiring spectra of fragmented ions. These fragmentation spectra can be represented as a molecular network, by measuring cosine distances between them, thus identifying signals from the same or similar molecules. Metrics that enable comparison between pairs of samples based on their metabolite profiles are in great need. Taking inspiration from the successful phylogeny-aware beta-diversity measures used in microbiome research, integrating chemical similarity information about the features in addition to their abundances could lead to better insights when comparing metabolite profiles. Chemical Structural and Compositional Similarity (CSCS) is a recently published similarity metric comparing the full set of signals and their chemical similarity between two samples. Efficient, scalable and easily accessible implementations of this algorithm is currently lacking. Here, we present an easily accessible and scalable implementation of CSCS in both python and R, including a version not weighted by intensity information.
Results: We provide a new implementation of the CSCS algorithm that is over 300 times faster than the published implementation in R, making the algorithm suitable for large-scale metabolomics applications. We also show that adding chemical information enriches existing methods. Furthermore, the R implementation includes functions for exporting molecular networks directly from the mass spectral molecular networking platform GNPS for ease of use for downstream applications.
Contact: brejnrod@sund.ku.dk
Availability: github.com/askerdb/rCSCS, github.com/askerdb/pyCSCS

1 Introduction

Liquid chromatography tandem - mass spectrometry (LC-MS/MS) is gaining more and more popularity in the metabolomics field with a wide range of applications (e.g. [11, 9, 13, 12]). As an extension of LC-MS, one of the most frequently used analytical platforms in mass spectrometry-based metabolomics [27, 4], LC-MS/MS does not only provide a high coverage and sensitivity towards semi-polar metabolites, but also provides chemical structural information of the metabolites investigated. In LC-MS/MS, metabolites are ionized and fragmented and the resulting fragmentation fingerprints of mass-to-charge ratios (m/z) are characteristic to the molecular structure of the metabolite. Furthermore, metabolites resulting in similar fragmentation patterns presumably exhibit similar chemical structures [24, 25]. The similarity of these fingerprints can be calculated using the cosine score and represented as a graph, resulting in so-called mass spectral molecular networks (recently reviewed in [14]). This approach, popularized by the user-friendly Global Natural Products Social Molecular Networking (GNPS) platform [24], has been highly effective in visualizing structural chemical relatedness across samples as well as in aiding the structural identification of metabolites. Several studies have focused on deriving and applying new distance metrics to

metabolomics profiles to compare pairs of samples [16, 15, 8]. Among these, the chemical structural and compositional similarity (CSCS) as proposed by Sedio and collaborators [19, 17, 18], accounts for the chemical structural similarity across metabolites by integrating the similarity of their MS/MS fragmentation patterns through the cosine score. Even though metabolites can be identified for only about 2 % of all signals in a typical LC-MS/MS experiment [1], CSCS has the advantage to integrate chemical structural information without depending on metabolite identification, thus enabling a comparison of the entire range of detected molecules. CSCS can be used to visualize chemical differences across samples by using Principal Coordinate Analysis (PCoA) plots [5] or distance based hypothesis testing of differences. Here, we implement CSCS in user-friendly python and R packages, and demonstrate its performance by computing commonly used metrics on nearly 500 LC-MS/MS samples of the American Gut Project [9] as well as fecal samples obtained from children with Crohn's disease at different time points during nutritional therapy [22].

2 Methods

2.1 Mathematical exposition

Consider two samples A and B with the union of features [1, ..., C], and intensity vectors A = [I_{a1}, ..., I_{aC}] and B = [I_{b1}, ..., I_{bC}]. Calculation of the Chemical Structural and Compositional Similarity (CSCS) takes the following inputs: a chemical similarity matrix, CSS, and the two vectors of ion intensities, A and B. CSS is defined as a matrix of pairwise cosine distances

$$CSS = \begin{bmatrix} cos\Theta_{1,1} & ... & cos\Theta_{1,C} \\ cos\Theta_{2,1} & ... & cos\Theta_{2,C} \\ cos\Theta_{C,1} & ... & cos\Theta_{C,C} \end{bmatrix}$$

In practice this matrix can be calculated from a file specifying pairwise cosine distances across nodes in a mass spectral molecular network, which can be downloaded from GNPS.

The CSCS as defined by Sedio and collaborators [19], referred to here as weighted CSCS ($CSCSw$) is then defined as:

$$CSCSw = \frac{CSS \times AB^T}{max(CSS \times AA^T, CSS \times BB^T)}$$

where $\times$ is element-wise multiplication.

Analogously, we here define the unweighted $CSCSu$ where all elements of A and B are dichotomized into 1 or 0 upon presence or absence of a signal in the sample, respectively. CSCS is a metric of similarity. However, many downstream analyses within metabolomics rely on a dissimilarity matrix (e.g. PCoA), thus our libraries return a dissimilarity matrix, corresponding to 1-CSCS.

2.2 Implementation

Implementations are available at `https://github.com/askerdb/rCSCS` as an R [21] package, which can be installed through devtools [26], and at `https://anaconda.org/askerdb/pycscs` as a python package, which can be installed through conda. The R package depends on the packages foreach [2], igraph [3] and Rcurl[7], wheras the python package depends on numpy [23], pandas [10], scikit bio (`http://scikit-bio.org`) as well as the sparse matrix implementation in scipy [6].

2.3 Data

Run time benchmarks were computed on 18 fecal samples obtained from children with Crohn's disease at different time points during nutritional therapy [22]. This dataset is publicly available at `https://massive.ucsd.edu/` under the MassIVE accession number MSV000081120. A total of 1245 MS/MS features were retrieved for this dataset using the GNPS networking parameters publicly accessible at `https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b0524246804a4b50a8a4ec6244a8be2e`. Samples from the American Gut Project data are publicly available under the MassIVE accession number MSV000080179 [9]. This dataset consisted of 489 samples and 16349 features using GNPS networking parameters publicly accessible at `https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a07557dc26cc4d3f8a2076d5ae0898a2`.

Implementation	Time (s)
rCSCS	1.53
pyCSCS	1.66
Sedio et al.	507.31

Table 1. CSCS timing benchmarks for a dataset consisting of 18 fecal samples obtained from children with Crohn's disease at different time points during nutritional therapy [22].

3 Results

3.1 Benchmarks

We computed CSCS for a dataset consisting of 18 fecal samples obtained from children with Crohn's disease at different time points during nutritional therapy [22] and a total of 1245 MS/MS features, a realistically sized dataset on a single core of a Macbook Pro (2.8 GHz i7), and compared it against the published implementation from Sedio and collaborators [19]. Results shown in Table 1 include run time overhead that should be minimal in realistic applications. To demonstrate the utility of this implementation on a scale that is relevant to high-throughput collection of data, we used the American Gut Project that consists of 489 samples with 16349 MS/MS features when downloaded from GNPS. Analysis of this dataset with our python implementation finished in 7.5 hours on 40 CPU cores. We did not compare run time of the original implementation [19], as this will likely not finish in a reasonable time.

3.2 Chemical information produces distinct similarities

To evaluate how the CSCS metrics compare with other popular metrics, we compared them with Bray-Curtis, Jaccard and Canberra metrics estimated for a dataset of 594 metabolome samples from the American Gut Project [9]. We used Procrustes analysis[20] to measure the similarity between the distance matrices and used this information as input for a PCoA plot (Figure 1). Metrics with chemical information are clearly separated from those without, and on this dataset the weighting of chemical information with ion intensities drives the separation on the axis that explains the most variance, while the unweighted CSCS is distinctly separated on the second principal component.

4 Conclusion

We have implemented the calculation of weighted and unweighted CSCS distances in R and Python. Through benchmarking of publicly available data we have demonstrated the highly significant run time improvement, which expands the application of CSCS to large datasets.

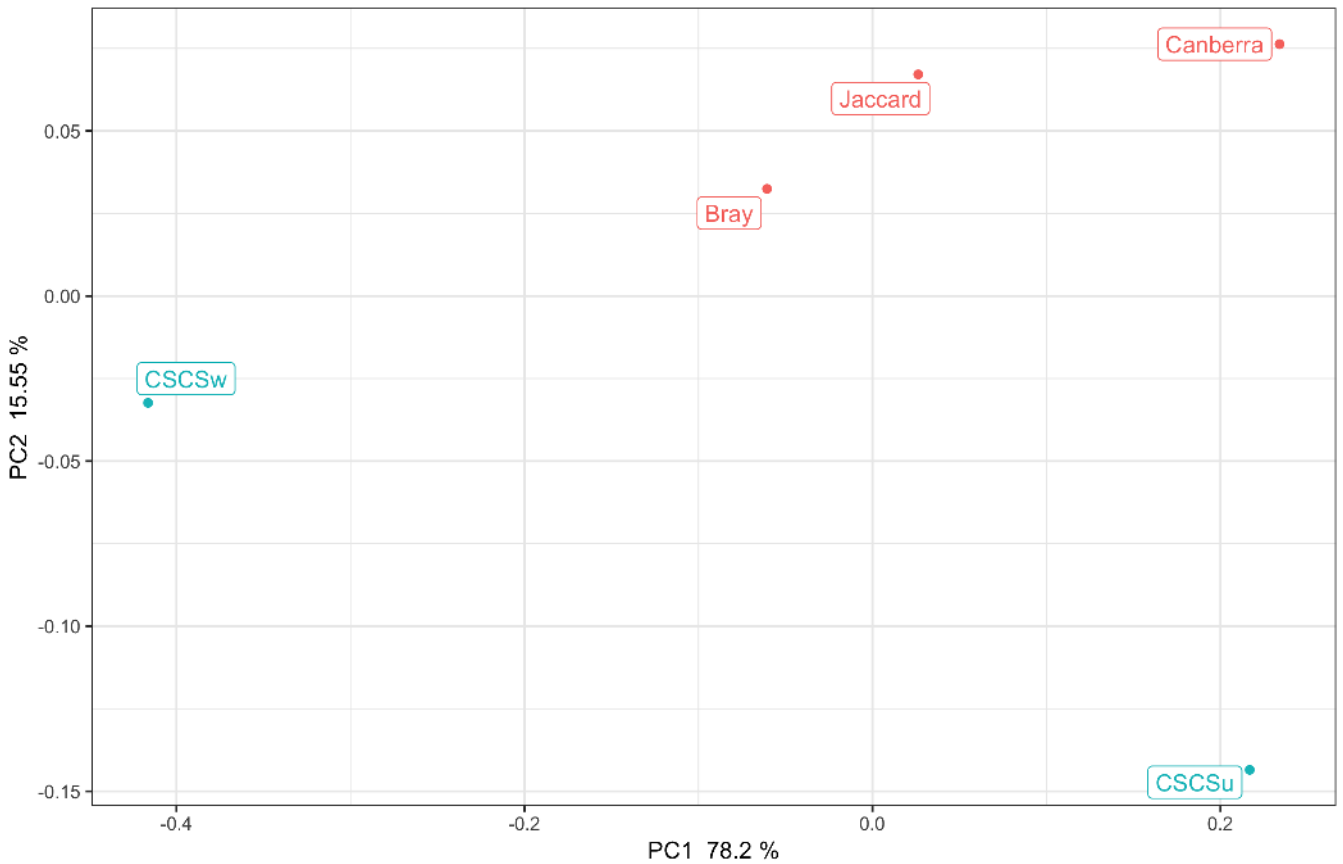

Fig. 1. Figure 1: PCoA of the procrustes distances between various distance metrics commonly used in metabolomics. Metrics colored in green include chemical information, either weighted by ion intensities or unweighted. Metrics with no chemical information are colored red. information

Acknowledgements

We would like to acknowledge Brian Sedio and collaborators for proposing and implementing the original chemical structural and compositional similarity metric, and making software available for direct comparisons.

Funding

Asker Brejnrod was supported by Independent Research Fund Denmark (DFF-6111-00471). JJJvdH was supported by an ASDI eScience grant (ASDI.2017.030) from the Netherlands eScience Center (NLeSC).

References

[1]Alexander A Aksenov, Ricardo da Silva, Rob Knight, Norberto P Lopes, and Pieter C Dorrestein. Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry*, 1(7):0054, 2017.

[2]Revolution Analytics and Steve Weston. foreach: Foreach looping construct for r. *R package version*, 1(1):2013, 2013.

[3]Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[4]Madeleine Ernst, Denise Brentan Silva, Ricardo Roberto Silva, Ricardo Z. N. Vncio, and Norberto Peporine Lopes. Mass spectrometry in plant metabolomics strategies: from analytical platforms to data acquisition and processing. *Nat. Prod. Rep.*, 31:784–806, 2014.

[5]John C Gower and Pierre Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of classification*, 3(1):5–48, 1986.

[6]Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: open source scientific tools for Python. 2014.

[7]Duncan Temple Lang. Rcurl: General network (http/ftp/...) client interface for r. *R package version*, 1(1), 2012.

[8]Kang Liu, Azian Azamimi Abdullah, Ming Huang, Takaaki Nishioka, Md Altaf-Ul-Amin, and Shigehiko Kanaya. Novel approach to classify plants based on metabolite-content similarity. *BioMed research international*, 2017, 2017.

[9]Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.

[10]Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pp. 51–56. Austin, TX, 2010.

[11]Don D Nguyen, Alexey V Melnik, Nobuhiro Koyama, Xiaowen Lu, Michelle Schorn, Jinshu Fang, Kristen Aguinaldo, Jr Lincecum, Tommie L, Maarten G K Ghequire, Victor J Carrion, Tina L Cheng, Brendan M Duggan, Jacob G Malone, Tim H Mauchline, Laura M Sanchez, A Marm Kilpatrick, Jos M Raaijmakers, René De Mot, Bradley S Moore, Marnix H Medema, and Pieter C Dorrestein. Indexing the pseudomonas specialized metabolome enabled the discovery of poaeamide b and the bananamides. *Nature microbiology*, 2:16197–16197, 10 2016.

[12]Florent Olivon, Pierre-Marie Allard, Alexey Koval, Davide Righi, Gregory Genta-Jouve, Johan Neyts, Cécile Apel, Christophe Pannecouque, Louis-Félix Nothias, Xavier Cachet, Laurence Marcourt, Fanny Roussi, Vladimir L. Katanaev, David Touboul, Jean-Luc Wolfender, and Marc Litaudon. Bioactive natural products prioritization using massive multi-informational molecular networks. *ACS Chemical Biology*, 12(10):2644–2651, 10 2017.

[13]Daniel Petras, Louis-Felix Nothias, Robert A. Quinn, Theodore Alexandrov, Nuno Bandeira, Amina Bouslimani, Gabriel Castro-Falcon, Liangyu Chen, Tam

Dang, Dimitrios J. Floros, Vivian Hook, Neha Garg, Nicole Hoffner, Yike Jiang, Clifford A. Kapono, Irina Koester, Rob Knight, Christopher A. Leber, Tie-Jun Ling, Tal Luzzatto-Knaan, Laura-Isobel McCall, Aaron P. McGrath, Michael J. Meehan, Jonathan K. Merritt, Robert H. Mills, Jamie Morton, Sonia Podvin, Ivan Protsyuk, Trevor Purdy, Kendall Satterfield, Stephen Searles, Sahil Shah, Sarah Shires, Dana Steffen, Margot White, Jelena Todoric, Robert Tuttle, Aneta Wojnicz, Valerie Sapp, Fernando Vargas, Jin Yang, Chao Zhang, and Pieter C. Dorrestein. Mass spectrometry-based visualization of molecules associated with human habitats. *Analytical Chemistry*, 88(22):10775–10784, 2016. PMID: 27732780.

[14]Robert A Quinn, Louis-Felix Nothias, Oliver Vining, Michael Meehan, Eduardo Esquenazi, and Pieter C Dorrestein. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends in pharmacological sciences*, 38(2):143–154, 2017.

[15]Anita Rácz, Filip Andrić, Dávid Bajusz, and Károly Héberger. Binary similarity measures for fingerprint analysis of qualitative metabolomic profiles. *Metabolomics*, 14(3):29, 2018.

[16]Lora A. Richards, Lee A. Dyer, Matthew L. Forister, Angela M. Smilanich, Craig D. Dodson, Michael D. Leonard, and Christopher S. Jeffrey. Phytochemical diversity drives plant–insect community diversity. *Proceedings of the National Academy of Sciences*, 112(35):10973–10978, 2015.

[17]Brian E. Sedio. Recent breakthroughs in metabolomics promise to reveal the cryptic chemical traits that mediate plant community composition, character evolution and lineage diversification. *New Phytologist*, 214(3):952–958, 2017.

[18]Brian E. Sedio, Cristopher A. Boya P., and Juan Camilo Rojas Echeverri. A protocol for high-throughput, untargeted forest community metabolomics using mass spectrometry molecular networks. *Applications in Plant Sciences*, 6(3):e1033, 2018.

[19]Brian E Sedio, Juan C Rojas Echeverri, P Boya, A Cristopher, and S Joseph Wright. Sources of variation in foliar secondary chemistry in a tropical forest tree community. *Ecology*, 98(3):616–623, 2017.

[20]Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[21]R Core Team et al. R: A language and environment for statistical computing. 2013.

[22]Justin J. J. van der Hooft, Joe Wandy, Francesca Young, Sandosh Padmanabhan, Konstantinos Gerasimidis, Karl E. V. Burgess, Michael P. Barrett, and Simon Rogers. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. *Analytical Chemistry*, 89(14):7569–7577, 2017. PMID: 28621528.

[23]Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[24]Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828, 2016.

[25]Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S. Heath, Jane Y. Yang, Roland D. Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M. Raaijmakers, Bradley S. Moore, Julia Laskin, Nuno Bandeira, and Pieter C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, 109(26):E1743–E1752, 2012.

[26]Hadley Wickham and Winston Chang. Devtools: Tools to make developing r packages easier. *R package version*, 1(0), 2016.

[27]Jean-Luc Wolfender, Serge Rudaz, Young Hae Choi, and Hye Kyong Kim. Plant metabolomics: from holistic data to relevant biomarkers. *Current Medicinal Chemistry*, 20(8):1056–1090, 2013.