Another look at microbe–metabolite interactions: how scale invariant correlations can outperform a neural network

Thomas P. Quinn[1*] and Ionas Erb[2]

[1]Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia
[2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
* *contacttomquinn@gmail.com*

Abstract

Many scientists are now interested in studying the correlative relationships between microbes and metabolites. However, these kinds of analyses are complicated by the compositional (i.e., relative) nature of the data. Recently, Morton et al. proposed a neural network architecture called mmvec to predict metabolite abundances from microbe presence. They introduce this method as a scale invariant solution to the integration of multi-omics compositional data, and claim that "mmvec is the only method robust to scale deviations". We do not doubt the utility of mmvec, but write in defense of simple linear statistics. In fact, when used correctly, correlation and proportionality can actually outperform the mmvec neural network.

1 Response

Many scientists are now interested in studying the correlative relationships between microbes and metabolites (e.g., [1, 2, 3, 4]). However, these kinds of analyses are complicated by the compositional (i.e., relative) nature of the data [5, 6]. Recently, Morton et al. proposed a neural network architecture called mmvec to predict metabolite abundances from microbe presence [7]. They introduce this method as a scale invariant solution to the integration of multi-omics compositional data, and claim that "mmvec is the only method robust to scale deviations". We do not doubt the utility of mmvec, but write in defense of simple linear statistics. In fact, when used correctly, correlation and proportionality can actually outperform the mmvec neural network.

Scale invariance is important because we do not want a method that is sensitive to (i.e., is variant to) changes in technical factors like sequencing depth (i.e., differences in scale). In compositional data analysis (CoDA), scale invariance is forced by using a log-ratio transformation that recasts the data with respect to an internal reference [8]. The resultant log-ratios are scale invariant, and so any analysis of log-ratios is scale invariant. This is true for multi-omics data too, but only if the transformation is performed correctly. Let us consider two possible transformations of the multi-omics data, presented here as functions of the input:

$$\begin{aligned}
\mathcal{A}(\mathbf{u}_i, \mathbf{v}_i) &= \text{clr}\left(\left[u_{i1}, ..., u_{iM}, v_{i1}, ..., v_{iN}\right]\right) \\
&= \log\left(\frac{\left[u_{i1}, ..., u_{iM}, v_{i1}, ..., v_{iN}\right]}{\sqrt[M+N]{\Pi_j^M u_{ij} \Pi_j^N v_{ij}}}\right) \\
\mathcal{B}(\mathbf{u}_i, \mathbf{v}_i) &= \left[\text{clr}\left(\left[u_{i1}, ..., u_{iM}\right]\right), \text{clr}\left(\left[v_{i1}, ..., v_{iN}\right]\right)\right] \\
&= \left[\log\left(\frac{\left[u_{i1}, ..., u_{iM}\right]}{\sqrt[M]{\Pi_j^M u_{ij}}}\right), \log\left(\frac{\left[v_{i1}, ..., v_{iN}\right]}{\sqrt[N]{\Pi_j^N v_{ij}}}\right)\right]
\end{aligned}$$

for sample i, where $\mathbf{u}_i$ measures the $1...M$ microbes and $\mathbf{v}_i$ measures the $1...N$ metabolites. Only approach $\mathcal{B}$ is scale invariant, but Morton et al. use approach $\mathcal{A}$ when they claim that correlation and proportionality are unreliable.

Why is approach $\mathcal{B}$ valid, but not approach $\mathcal{A}$? It is because the microbe and metabolite data are generated from two separate sampling processes: they are individually, not jointly, constrained to sum to 1. In other words, the abundance of microbe 1 is limited by the abundance of microbes 2-to-M, but is in no way limited by the abundance of metabolites 1-to-N. Consequently, the denominator from approach $\mathcal{A}$ has no meaning. On the other hand, the denominators from approach $\mathcal{B}$ have the property that they cancel any constant factor multiplied with the sample values in each numerator. As such, they cancel the implicit biases that arise from the sequencing procedure and cause the samples to be on different scales. An additional property of these numerators is that they are useful normalization factors themselves [9]: under the assumption that the majority of features are unchanged, approach $\mathcal{B}$ will make the transformed data proportional to the original (absolute) data (thus performing an effective library-size normalization).

How important is the choice between approach $\mathcal{A}$ and approach $\mathcal{B}$? We repeated the authors' analysis to measure the F1-score (precision and recall) for the top microbe-metabolite associations, except this time we used approach $\mathcal{B}$. Figure 1 shows the updated performance of correlation and proportionality, both of which outperform mmvec on their simulated benchmark. Interestingly, correlation (Spearman and Pearson) performed best, suggesting that the "ground truth" includes power-law relationships between microbes and metabolites (i.e., log-linear relationships with slopes other than 1). Since ϕ and ρ are designed for intercept-free linear relationships, pairs in which one feature is proportional to another when taken to an exponent will usually go undetected.

We do not disagree that neural networks can add value to multi-omics data integration. Their ability to learn non-linear relationships could improve metabolite prediction by directly modeling complex microbe-microbe interactions. However, neural networks do not offer a magical solution to the problems of compositional data analysis [10]. They are merely a nested series of transformed linear operators. As such, they may be prone to yield spurious results whenever a simple linear method would. It seems to us that mmvec's primary advantage is how it handles the compositional data, not its neural network architecture. Indeed, our analysis shows that when we transform the multi-omics data correctly, simple linear methods outperform mmvec in their own benchmark. We conclude by reminding our readers that not all problems in biology are solved by adding layers of complexity: sometimes it is sufficient to use our simplest solutions more carefully.

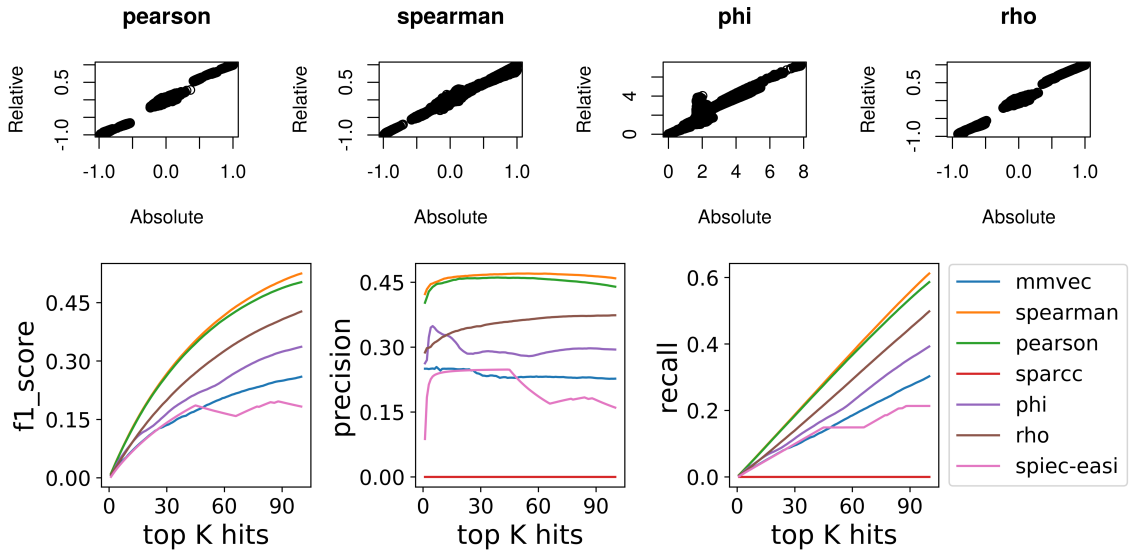

Figure 1: Our re-analysis of the Morton et al. simulated data shows that, with the correct log-ratio transformation, simple linear statistics are scale invariant and sufficient for use in the integration of multi-omics compositional data. In the top panels, we see excellent agreement between absolute and relative correlation, as well as between absolute and relative proportionality. In the bottom panels, we see the updated performances from the simulated data benchmark. When used correctly, correlation and proportionality can actually outperform the mmvec neural network. All scripts available from https://doi.org/10.5281/zenodo.3544999.

References

[1] Agnieszka Smolinska, Danyta I. Tedjo, Lionel Blanchet, Alexander Bodelier, Marieke J. Pierik, Ad A. M. Masclee, Jan Dallinga, Paul H. M. Savelkoul, Daisy M. A. E. Jonkers, John Penders, and Frederik-Jan van Schooten. Volatile metabolites in breath strongly correlate with gut microbiome in CD patients. *Analytica Chimica Acta*, 1025:1–11, September 2018.

[2] Jason Lloyd-Price, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, Nadim J. Ajami, Kevin S. Bonham, Colin J. Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G. Graeber, A. Brantley Hall, Kathleen Lake, Carol J. Landers, Himel Mallick, Damian R. Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A. White, Jonathan Braun, Lee A. Denson, Janet K. Jansson, Rob Knight, Subra Kugathasan, Dermot P. B. McGovern, Joseph F. Petrosino, Thaddeus S. Stappenbeck, Harland S. Winter, Clary B. Clish, Eric A. Franzosa, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655, May 2019.

[3] Zheng-Zheng Tang, Guanhua Chen, Qilin Hong, Shi Huang, Holly M. Smith, Rachana D. Shah, Matthew Scholz, and Jane F. Ferguson. Multi-Omic Analysis of the Microbiome and Metabolome in Healthy Subjects Reveals Microbiome-Dependent Relationships Between Diet and Metabolites. *Frontiers in Genetics*, 10, 2019.

[4] Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, Fumie Hosoda, Hirofumi Rokutan, Minori Matsumoto, Hiroyuki Takamaru, Masayoshi Yamada, Takahisa Matsuda, Motoki Iwasaki, Taiki Yamaji, Tatsuo Yachida, Tomoyoshi Soga, Ken Kurokawa, Atsushi Toyoda, Yoshitoshi Ogura, Tetsuya Hayashi, Masanori Hatakeyama, Hitoshi Nakagama, Yutaka Saito, Shinji Fukuda, Tatsuhiro Shibata, and Takuji Yamada. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature Medicine*, 25(6):968, June 2019.

[5] Andrew D. Fernandes, Jennifer Ns Reid, Jean M. Macklaim, Thomas A. McMurrough, David R. Edgell, and Gregory B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, 2014.

[6] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Computational Biology*, 11(3), March 2015.

[7] James T. Morton, Alexander A. Aksenov, Louis Felix Nothias, James R. Foulds, Robert A. Quinn, Michelle H. Badri, Tami L. Swenson, Marc W. Van Goethem, Trent R. Northen, Yoshiki Vazquez-Baeza, Mingxun Wang, Nicholas A. Bokulich, Aaron Watters, Se Jin Song, Richard Bonneau, Pieter C. Dorrestein, and Rob Knight. Learning representations of microbe–metabolite interactions. *Nature Methods*, pages 1–9, November 2019.

[8] J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London, UK, UK, 1986.

[9] Thomas P. Quinn, Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, August 2018.

[10] Raimon Tolosana Delgado, Hassan Talebi, Mahdi Khodadadzadeh, and K. Gerald van den Boogaart. On machine learning algorithms and compositional data. *CoDaWork2019*, 2019.