Examining microbe–metabolite correlations by linear methods

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*

1 Response

Analyzing correlative relationships between microbes and metabolites is a timely topic [1, 2, 3], but complicated by the compositional (i.e., relative) nature of the data [4, 5]. Recently, Morton et al. proposed a neural network architecture called mmvec to predict metabolite abundances from microbe presence [6]. We do not doubt the utility of mmvec, but write in defense of simple linear statistics. When used correctly, correlation and proportionality [5, 7] can be scale invariant, and can outperform mmvec in certain conditions.

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

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

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

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 not 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 their respective numerators. As such, they cancel the implicit sequencing biases that cause the samples to be on different scales. An additional property of these denominators 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 and thus performs effective library-size normalization.

We repeated the authors' analysis to measure the F1-score (precision and recall) for the top microbe-metabolite associations using approach $\mathcal{B}$. Figure 1 shows the performance of correlation and proportionality, both of which outperform mmvec on their simulated benchmark. Interestingly, correlation 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 which, for example, could mean that although an increase in two microbe units associate with a doubling of metabolites, four units associate with a quadrupling). Since ϕ and ρ are designed for intercept-free linear relationships, these power-law relationships will usually go undetected. Note that although SPIEC-EASI already implements $\mathcal{B}$ in "multi-source" mode, it makes a strong assumption that the true ecological association network is sparse [10]. This assumption does not appear to hold true for the simulated data (see [6]). If one instead calculates covariance via a second inversion of the regularized inverse covariance matrix, the model performs well (see "QUIC-cov" in Figure 1).

Data sparsity, by which we mean an excess of zero counts, presents a major challenge to microbiome data analysis. For one, a log-ratio transformation fails for a zero entry. Many methods have been proposed to address compositional zeros, including Bayesian imputation strategies [11] and alternative transformations [12]. The simplest approach involves replacing all zeros with a very small number. Every zero handling strategy has limitations. However, it remains unclear whether a neural network will necessarily perform better. For the *simulated* microbiome data used in Figure 1, about 14% of the values are zero (i.e., the data are 14% sparse). We increase the sparsity by sampling new counts from an equivalent multinomial distribution where we use the closed counts as parameters at 1/20 the sequencing depth. This sampling generates new relative data with 71% sparsity, without any change to the corresponding absolute data. Figure 2 shows how simple correlations at 71% sparsity–despite a considerable drop in accuracy–still outperform the mmvec baseline at 14% sparsity. Interestingly, Spearman's rank correlation is impaired most by data sparsity, likely because any change to small counts would distort ranks more than parametric covariance estimates.

It is worth noting that neither the precision nor the recall is high for any of these methods. This is consistent with how information gets lost when producing compositional counts, especially if under-sampling leads to an excess of zero counts. It is also worth noting that CLR-based correlations, by definition, describe how microbes and metabolites behave relative to their respective sample means. Although the CLR can, under some circumstances, provide a useful normalization of the data, analysts must take care not to forget that the geometric mean is foremost a *reference frame* [13], a kind of yardstick against which we compare the relative abundances in order to establish a scale-invariant analysis of the data. If the CLR transform does not perfectly normalize the relative data, then we might see some discrepancies between our estimates and the true associations [7].

Note that even when the CLR is not a perfect normalization tool, proportionality is designed to still reveal some linear associations without having to make the relationship between the variables and the reference explicit. On the other hand, CLR-based correlations depend more on the chosen reference because any expression of power laws will necessarily involve that reference. To see this, we assume the correlation coefficient was high enough to detect a linear relationship between the logarithms of two features $\mathbf{x}$ and $\mathbf{y}$, both having the same reference $\mathbf{r}$. We have the log-linear model

$$\log \frac{\mathbf{y}}{\mathbf{r}} = m \ \log \frac{\mathbf{x}}{\mathbf{r}} + b + \boldsymbol{\epsilon}$$

(with offset b and error term $\boldsymbol{\epsilon}$). This implies that $\mathbf{y} = e^{b+\boldsymbol{\epsilon}}\mathbf{r}^{1-m}\mathbf{x}^{m}$. From this we can see how the reference (e.g., geometric mean from CLR) influences the relationship between variables when the slope m is not 1.

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-metabolite interactions [14]. However, neural networks do not offer a magical solution to the problems of compositional data analysis [15]. 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 yield spurious results. It seems to us that mmvec's primary advantage is how it handles the compositional data, not its neural network architecture *per se*. For example, the use of a softmax transformation, which is equivalent to an inverse CLR transformation, might imply that the linear operations from previous layers actually occur in CLR coordinates [6].

We conclude by reminding readers that not all problems in biology are solved by adding layers of complexity: sometimes it is sufficient to use the simplest solutions more carefully.

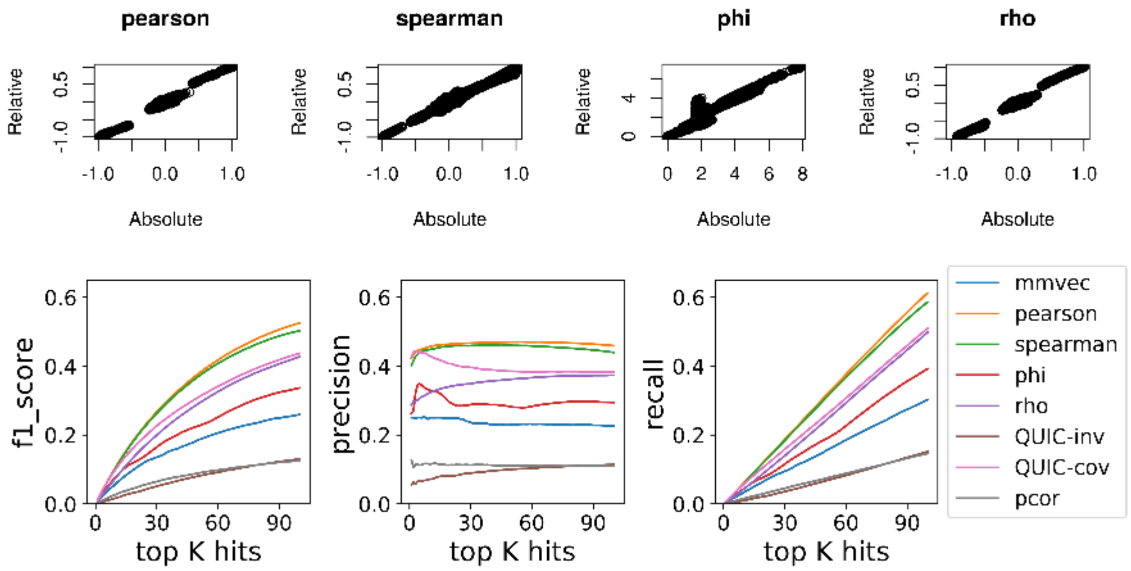

Figure 1: The re-analysis of the Morton et al. simulated data. The top panels show agreement between absolute and relative metrics when using approach $\mathcal{B}$. The bottom panels show the updated performances from the simulated data, where QUIC refers to the regularized inverse covariance matrix, and pcor refers to partial correlations.

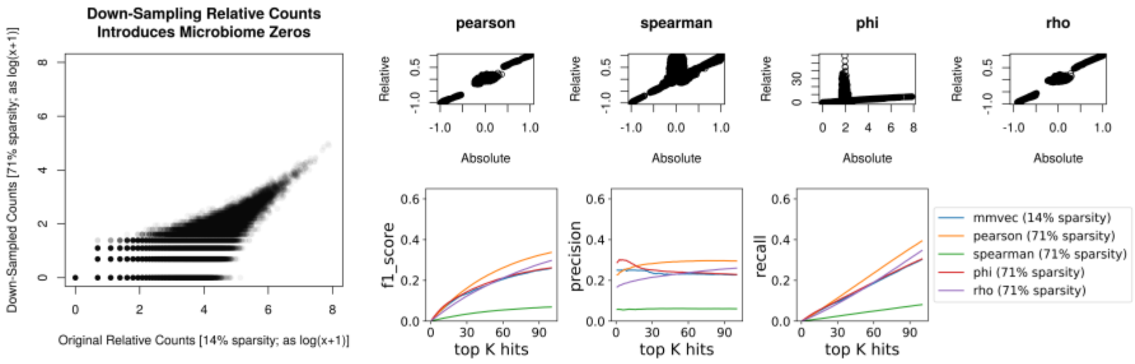

Figure 2: The re-analysis of the sparsified simulated data. The leftmost panel shows a log-log plot of the new relative data with 71% sparsity (y-axis) versus the original relative data with 14% sparsity (x-axis), confirming successful down-sampling. The top-right panels show agreement between absolute and relative metrics when using approach $\mathcal{B}$.

2 Data Availability

Data and scripts used in Figure 1 and Figure 2 are available from https://doi.org/10.5281/zenodo.3610709 and https://doi.org/10.5281/zenodo.3833174, respectively.

3 Method

For Figure 1, we performed a re-analysis of the simulated data by taking the following steps: (1) We loaded in the absolute and relative data sets provided by the authors in the "results/benchmark_output/CF_sims/data" directory; (2) We replaced all zeros with the minimum non-zero value; (3) We performed a CLR of the microbe and metabolite data separately for each of the absolute and relative data sets; (4) We calculated proportionality (using propr package version 4.2.8) and correlation (using base R version 3.6.3) for each of the absolute and relative data sets;

(5) We measured and plotted the precision-recall of the relative data analysis against the mmvec results using a Python script from the authors. For Figure 2, we repeated this same procedure, but added a new step where we down-sampled the relative microbiome data by using a multinomial distribution at 1/20 the sequencing depth, where the expected proportions were set as the original relative microbiome proportions.

4 Competing Interests

The authors declare no competing interests.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Ionas Erb and Cedric Notredame. How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135:21–36, 2016.

[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] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, May 2015.

[11] Javier Palarea-Albaladejo and Josep Antoni Martín-Fernández. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, April 2015.

[12] Cameron Martino, James T. Morton, Clarisse A. Marotz, Luke R. Thompson, Anupriya Tripathi, Rob Knight, and Karsten Zengler. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems*, 4(1):e00016–19, February 2019.

[13] James T. Morton, Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S. Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. Establishing microbial composition measurement standards with reference frames. *Nature Communications*, 10(1):1–11, June 2019.

[14] Vuong Le, Thomas P. Quinn, Truyen Tran, and Svetha Venkatesh. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. *BMC Genomics*, 21(4):256, July 2020.

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