Notes on applying gene set enrichment to microbiome analyses

Quang Nguyen

April 9, 2020

1 Gene set analysis in the context of the microbiome

2 Methods involving trees

3 Some hypotheticals and further considerations

3.1 Approach 1: Taxonomic aggregation using isometric log-ratio transformations

3.1.1 Statistical interpretation of taxonomic aggregation using elementwise summations

Most microbiome studies have performed taxonomic aggregation through element-wise summation of the count vectors for all taxa assigned to the taxonomic rank of interest. Prior to statistical analysis, these aggregated counts are then usually converted to relative abundances by dividing by the total counts per sample, effectively normalizing for differences in library size. This compositional closure operation is unavoidable as ”size-factors” can't be estimated from most microbiome data sets due to a lack of ”consistent features” similar to that of housekeeping genes or UMIs in RNA-seq type studies. As such, we can define the classical taxonomic aggregation processs as simply the element-wise summation of compositional parts (or proportions). Let P_i be the relative abundance of higher taxonomic rank (HTR) P (indexed by $\mathbb{P}$) in sample i with raw counts x_{ij} where j is the column index of the lower taxonomic rank (LTR), as such:

$$P_i = \frac{\sum_{j \in \mathbb{P}} x_{ij}}{\sum x_{ij}} = \sum_{j \in \mathbb{P}} \frac{x_{ij}}{\sum x_{ij}} = \sum_{j \in \mathbb{P}} c_{ij}$$

Downstream analyses follow principles of compositional data analysis [4], which usually involves analyzing ratios of aggregated LTRs. These log-ratio analyses with aggregated compositions are termed ”groups of amalgamated parts” analysis in the CoDA literature [2].

However, as Egozcue et al. [2] pointed out, amalgamated compositions are not equivalent to their original form. For example, take a simple composition of 3 parts $x = [x_1, x_2, x_3]$ and the aggregated composition $y = [x_1 + x_2, x_3]$ (with n samples). The center [1] of the initial composition is

$$cen(x) = \mathcal{C}\left[\prod^{n} x_1, \prod^{n} x_2, \prod^{n} x_3\right]$$

while the center of the aggregated composition is

$$cen(y) = \mathcal{C}\left[\prod^{n}(x_1 + x_2), \prod^{n} x_3\right]$$

These two centers are very different, which also then translates to differences in the inter-sample Aichison distance after transformation. This invariant in inter-sample distance after aggregation means that aggregation distorts the original composition, and can attenuate differences in samples even if they're just noise (Figure 1) [2]. Even though other studies contend that aggregation is still viable [5], we contend that this distortion effect is significant in microbiome analyses, which relies a lot on distance based methods.

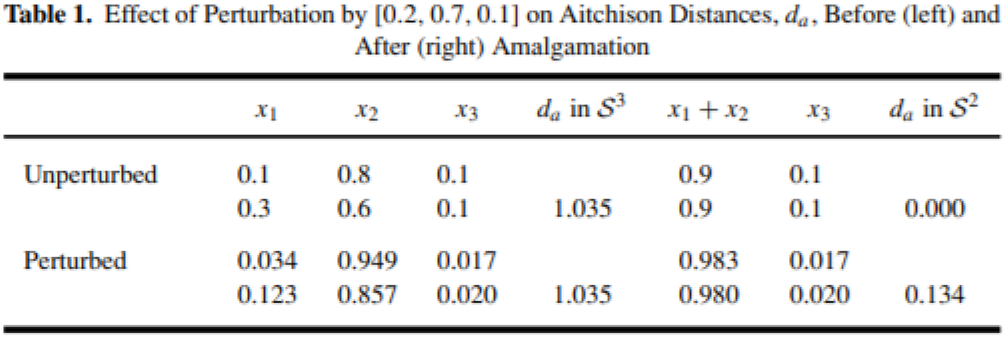

Table 1. Effect of Perturbation by [0.2, 0.7, 0.1] on Aitchison Distances, d_a, Before (left) and After (right) Amalgamation

	x_1	x_2	x_3	d_a in $\mathcal{S}^3$	$x_1 + x_2$	x_3	d_a in $\mathcal{S}^2$
Unperturbed	0.1	0.8	0.1		0.9	0.1	
	0.3	0.6	0.1	1.035	0.9	0.1	0.000
Perturbed	0.034	0.949	0.017		0.983	0.017	
	0.123	0.857	0.020	1.035	0.980	0.020	0.134

Figure 1: Table from Egozcue et al. demonstrating alterations of sample distances after aggregation

3.1.2 Competitive gene set testing using isometric log-ratios

In order to solve this amalgamation issue, Egozcue et al. proposed the isometric log ratio (ilr) transformation [3]. In essence, ilr transform is a projection of the composition from the Aichison space to an orthonormal basis that exists in the simplex. This allows for the usage of standard statistical techniques as the composition is ”opened” as well as being geometrically coherent compared to other flavors of log-ratio transforms. Conveniently, a sequential binary partition (SBP - which is a tree) is a valid orthonormal basis where the composition can be projected onto [3]. The transformed ilr coordinates are the tree nodes, which represent ”balances” between two sides of the node. In figure 2 is a toy example the ilr transformation on top of a phylogenetic tree. Each node x_1^*, x_2^*, x_3^* represents the transformed ilr coordinates. The ilr transformation is defined as follows:

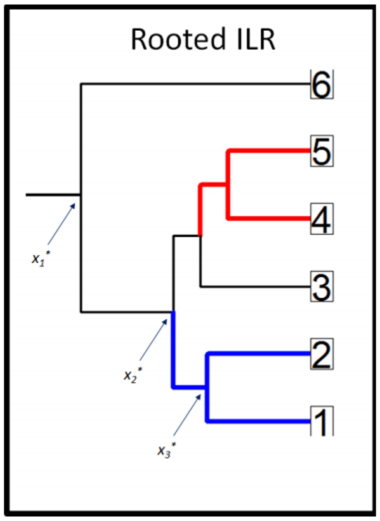

Figure 2: A sample SBP in tree form which is also the phylogenetic tree

$$x_i^* = \sqrt{\frac{l}{r+l}} \log \left(\frac{g(\boldsymbol{x}_{j \in \boldsymbol{L}})}{g(\boldsymbol{x}_{j \in \boldsymbol{R}})} \right)$$

where $\boldsymbol{x}$ is the compositional vector, $g()$ is the geometric mean, $\boldsymbol{L}$ is the set of size l of all parts on the left side of the node, and $\boldsymbol{R}$ is the set of size r of all parts on the left side of the node. Note that $\boldsymbol{L}$ and $\boldsymbol{R}$ are non-overlapping sets. For the example in Figure 2, we have:

$$x_2^* = \sqrt{\frac{2}{2+3}} \log \left(\frac{(x_1 x_2)^{1/2}}{(x_3 x_4 x_5)^{1/3}} \right)$$

The ilr coordinate can be interpreted as the overall relative contribution of variables in $\boldsymbol{L}$ to the composition of $\boldsymbol{L} \cup \boldsymbol{R}$ weighted by the sizes of $\boldsymbol{L}$ and $\boldsymbol{R}$. This concept of balances have gained recent attention by the microbiome field, targeting transformations and dimension reduction along the phylogenetic tree [7, 6]. The ilr transformation
Since the analysis of compositional data involves ratios of variables instead of individual factors, CoDA lends itself naturally to the gene set testing, specifically competitive gene set testing.

References

[1] John Aitchison. A Concise Guide to Compositional Data Analysis. page 134.

[2] J. J. Egozcue and V. Pawlowsky-Glahn. Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, 37(7):795–828, October 2005.

[3] J J Egozcue, V Pawlowsky-Glahn, G Mateu-Figueras, and C Barcelo-Vidal. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, page 22, 2003.

[4] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.*, 8, 2017.

[5] Michael Greenacre. Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Applied Computing and Geosciences*, 5:100017, March 2020.

[6] Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:e21887, February 2017.

[7] Alex D. Washburne, Justin D. Silverman, Jonathan W. Leff, Dominic J. Bennett, John L. Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A. David. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5:e2969, February 2017.