Candidate methods for microbiome set testing

Quang Nguyen

April 14, 2020

1 Taxonomic aggregation using isometric log-ratio transformations

1.1 Introduction

1.1.1 Taxonomic profiling using high throughput sequencing

Limitations of culturing techniques have prevented scientists from investigating the dynamics of highly complex microbial communities, especially human associated microbiomes. Advances in high-throughput sequencing have enabled the culture-free analysis of such communities, however sequencing data comes with additional statistical challenges.
One major difficulty of analyzing microbiome data is that it is strictly compositional [4]. This is because each sample has a different library size, induced through the PCR procedure embedded in short-read sequecning technologies. However, unlike RNAseq or even scRNAseq, microbiome data does not have "consistent features", such as UMIs or housekeeping genes, that can be used to estimate "size factors", allowing RNAseq-type data to break open the composition [6]. As such, microbiome data primarily exists in the form of relative abundances, where the principles of compositional data analysis (CoDA) applies [1].

1.1.2 Variable aggregation with microbiome data

Another challenge for microbiome data analysis is that it is high dimensional. A common way to reduce this dimensionality burden is to aggregate taxa to higher Linnean taxonomic levels, obtained through genomic annotation. Additionally, aggregating variables using apriori classificiation can assist in reducing the issue of multiple testing, as well as improving interpretability. Coincidentally, this is also the goal of gene set analysis (or gene set testing) methods often used in RNAseq-type data sets.
Currently, 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 any downstream statistical analysis, these aggregated counts are then transformed back to compositional form. As such, we can define the classical taxonomic aggregation processs as simply the element-wise summation of the relative abundances. Let P_i be the relative abundance of higher taxonomic (HT) rank P (indexed by $\boldsymbol{P}$) in sample i with raw counts x_{ij} where j is the column index of the lower taxonomic (LT) proportions, as such:

$$P_i = \frac{\sum_{j \in \boldsymbol{P}} x_{ij}}{\sum_j x_{ij}} = \sum_{j \in \boldsymbol{P}} \frac{x_{ij}}{\sum_j x_{ij}} = \sum_{j \in \boldsymbol{P}} c_{ij} \tag{1}$$

Downstream analysis of aggregated compositions are termed "groups of amalgamated parts analysis" in the CoDA literature [2]. However, as Egozcue et al. [2] pointed out, amalgamated compositions using sums 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

$$cen(x) = \mathcal{C}\left[(\prod^n x_1)^{1/n}, (\prod^n x_2)^{1/n}, (\prod^n x_3)^{1/n}\right]$$

while the center of the aggregated composition is

$$cen(y) = \mathcal{C}\left[(\prod^{n}(x_1 + x_2))^{1/n}, (\prod^{n} x_3)^{1/n}\right]$$

These two centers are very different, which also then translates to differences in the inter-sample Aichison distance after aggregation. In other words, distances between samples are not invariant to the aggregation procedure. Interestingly in Figure 1, the authors demonstrated that even though the distance between two samples is the same after a perturbation, this preserving nature goes away after the variables are aggregated. In the context of microbiome data analysis, this means that technical noise under the original (unaggregated) composition can be inflated. As such, we argued that elementwise summation is not statistically consistent with Aichison geometry of compositions, despite other studies suggesting differently [5]. Specifically, this is important in microbiome data analysis as studies often aggregate variables to more than one higher taxonomic rank, which means that analyses done in the Genus level might create different results under Family or Phylum levels. As such, there is a need for a compositionally coherent method to perform taxonomic aggregation, or taxonomic rank enrichment analysis.

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

1.1.3 Isometric log-ratio transformation

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, Egozcue et al. also showed that we can define a viable orthonormal basis from a sequential binary partition (SBP - which is a tree) [3]. The transformed ilr coordinates are the tree nodes, which represent "balances" between two sides of the node. 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:

$$x_i^* = \sqrt{\frac{l \cdot r}{r + l}} \log\left(\frac{g(\boldsymbol{x}_{j \in \boldsymbol{L}})}{g(\boldsymbol{x}_{j \in \boldsymbol{R}})}\right) \tag{2}$$

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:

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

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

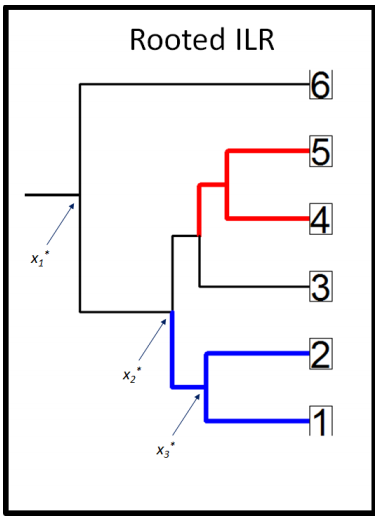

Figure 2: A sample SBP in tree form which is also the phylogenetic tree

the microbiome field, targeting transformations and dimension reduction along the phylogenetic tree [9, 7]. Washburne et al. [9] argued that the *ilr* uses ratios of geometric means, which is a more compositionally meaningful way to aggregate variables. Since the *ilr* transformation naturally incorporates component comparison (as with all log-ratio techniques), it is a natural extension to perform competitive gene set enrichment (or taxonomic set enrichment - TSE), which tests the null hypothesis that genes in the gene set show more association with the outcome than those outside the gene set [8]. Roughly speaking, that null hypothesis can be rewritten as:

$$H_0 : \frac{\mathcal{A}(g \in G)_X}{\mathcal{A}(g \notin G)_X} = \frac{\mathcal{A}(g \in G)_Y}{\mathcal{A}(g \notin G)_Y} \tag{3}$$

where $\mathcal{A}$ is a general aggregation function, g represents genes, G is a candidate gene set, X and Y are the case/control status. In other words, the competitive null hypothesis is that the relative enrichment of genes in gene set G compared to those not in the gene set is the same across two conditions. As per the definition of the *ilr* transformation, this is equivalent to testing the difference in x_i^* in equation (2) between conditions X and Y. As such, here we define a method that performs TSE using the *ilr* transform as the test statistic, naturally incorporating both compositional data analysis as well as competitive set enrichment.

1.2 Methods

1.2.1 Taxonomic rank enrichment analysis using isometric log-ratio transformations (TRE-ILR)

Here we propose a competitive taxonomic enrichment method based on an *ilr* transformation of microbial compositions. Shortened as TRE-ILR.

The TRE-ILR method takes in two matrices:

- $\boldsymbol{X}$: The $n \times p$ matrix of relative abundances of p LT proportions in n samples
- $\boldsymbol{A}$: The $m \times p$ matrix denoting the assignment of p LT proportions into m HT sets

TRE-ILR generates the following matrix:

- $\boldsymbol{S}$: The $n \times m$ matrix denoting the enrichment scores of m HT sets by n samples.

With inputs $\boldsymbol{X}$ and $\boldsymbol{A}$, we compute $\boldsymbol{S}$ as follows:

1. Let $\boldsymbol{R}$ be a $n \times m$ matrix of competitive *ilr* statistic for each HT set which is defined as follows:

$$R_{ij} = \sqrt{\frac{\sum_k A_{jk}(p - \sum_k A_{jk})}{p}} \log\left(\frac{g(\boldsymbol{x}_{ik}|A_{jk} \neq 0)}{g(\boldsymbol{x}_{ik}|A_{jk} = 0)}\right) \tag{4}$$

2. To capture the distribution of the *ilr* test statistic under the null hypothesis that the relative weights of LT proportions in the HT sets to the overall composition is no different than LT proportions not in the HT sets, the competitive *ilr* statistic is computed for each HT set with permuting row labels of matrix $\boldsymbol{A}$, simulating random assignments of LTs into HT sets. Denote $\boldsymbol{A}_{perm}$ be the row-permuted version of $\boldsymbol{A}$ and $\boldsymbol{R}_{perm}$ be row-permuted version of $\boldsymbol{R}$

3. We fit a gaussian distribution using the method of maximum likelihood for each column $\boldsymbol{R}_{perm}$. This stem from previous research treating the *ilr* coordinates as normally distributed [2].

4. Use the cumulative distribution (CDF) for the normal distribution to compute specific enrichment scores for HT sets. We formulate the target matrix S as CDF scores for the normal distribution fitted on columns of $\boldsymbol{R}_{perm}$:

$$\boldsymbol{S}[,k] = F_{N(\hat{\mu}_k, \hat{\sigma}_k)}(\boldsymbol{R}_{perm}[,k]) \tag{5}$$

1.2.2 TRE-ILR and standard isometric log-ratio transformations

Let $\boldsymbol{M}$ be the $p-1 \times p$ sign matrix representing a sequential binary partition for p ASVs across $p-1$ orders, with the first order being the first node from the root of the tree. For each sample i we define $\mathcal{M}_i$ as the set of $\boldsymbol{M}$s such that

$$\mathcal{M}_i = \left\{ \boldsymbol{M} | \boldsymbol{M}_{1j} = \begin{cases} 1 \text{ if } A_{ij} = 1 \\ -1 \text{ if } A_{ij} = 0 \end{cases} \right\}$$

As such, $\mathcal{M}$ represents the set of SBPs such that the first order partition splits between the LT belonging to the HT set and those that don't. The *ilr* coordinate of the first order partition is equal across all $\boldsymbol{M} \in \mathcal{M}$. In other words, we're interested in the coordinates for the projection of the composition $\boldsymbol{x}_i$ onto a very specific unit vector defined by the first order split as explained above.

$$\boldsymbol{e}_i = \mathcal{C}[\exp(\underbrace{a, a, a, ..., a, a, a}_{\sum_k A_{jk} \text{ elements}}, \overbrace{b, b, b, ..., b, b, b}^{p - \sum_k A_{jk} \text{ elements}}))]$$

where $a = \sqrt{\frac{\sum_k A_{jk}}{p \cdot (p - \sum_k A_{jk})}}$ and $b = \sqrt{\frac{-(p - \sum_k A_{jk})}{p \cdot \sum_k A_{jk}}}$ with $\sum_k A_{jk}$ being the size of the HT set k and $p - \sum_k A_{jk}$ being the number of LTs not in the HT set k. This unit vector can be part of various other orthonormal bases defined by the subtrees following the initial split. Since this vector is redefined for every HT set, the TRE-ILR *ilr* coordinates can't be compared across sets without some sort of transformation.

1.2.3 Statistical properties of TRE-ILR

Due to the equivalent of the TRE-ILR scores to the *ilr* coordinates of the composition onto the unit vector $\boldsymbol{e}$ defined above, it enjoys the various statistical properties of the *ilr* coordinate, specifically that it can be assumed to be normally distributed [3, 2]. As such, the raw TRE-ILR scores can be used for hypothesis testing for any specific HT set across two known case/control conditions. However, in order to use these scores together in a statistical model like a regression framework, we further transformed the scores as the CDF of the row-permuted distribution, which transforms the scores into a common scale. Furthermore, p-values associated to our null hypothesis can be obtained with a simple operation of $1 - S_{ij}$. Finally, it bounds the scores between 0 and 1, and is robust to large outliers.

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] Thomas P. Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F. Richardson, and Tamsyn M. Crowley. A field guide for the compositional analysis of any-omics data. Preprint, Bioinformatics, December 2018.

[7] 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.

[8] Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102(38):13544–13549, September 2005.

[9] 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.