Taxonomic Enrichment Analysis with Isometric Log-Ratios

Quang P. Nguyen[1,2], Anne G. Hoen[1,2], and H. Robert Frost[1]

[1] *Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA*

[2] *Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA*

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque arcu libero, suscipit sed enim nec, posuere eleifend nunc. Phasellus vitae augue orci. Sed vestibulum nisi id augue porta, a sagittis magna accumsan. Sed augue mi, venenatis sed fringilla nec, ultrices id ipsum. Suspendisse ac sapien eu mi laoreet fringilla. Nulla facilisi. Sed eget feugiat erat, et efficitur risus. Duis sit amet nulla at leo dignissim porta. Morbi nec ligula non sapien fringilla congue. Proin consequat volutpat nulla, eu convallis leo tempor in. Mauris elit sem, dignissim sit amet sapien sed, varius laoreet felis. Etiam elementum vulputate justo non malesuada. Suspendisse a libero id massa pellentesque convallis at in nibh. Ut nec consequat ante, vitae convallis dui. Ut eu pharetra nisi.

Background

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 [5]. Even though methods have emerged in the scRNAseq domain addressing estimation of effective library size independent of persistent variables, the assumptions of these approaches have not been verified in the context of the microbiome. As such, microbiome data primarily exists in the form of relative abundances, where the principles of compositional data analysis (CoDA) applies [1].

Variable aggregation with microbiome data

However, microbiome data is also high dimensional. A common way to reduce this burden is to aggregate taxa, most naturally to higher Linnean taxonomic levels. This would reduce the number of hypotheses being tested, as well as improving interpretation. 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 sum-based 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 $\mathcal{P}$ in sample i with raw counts x_{ij} where j is the column index of the lower taxonomic (LT) proportions. Let $\boldsymbol{P}$ be the set of column indices that belong to the HT rank $\mathcal{P}$ of interest. As such we have:

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

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, where the transformation distorts the Aichison distance between samples. Since microbiome data analysis relies on distance-based methods, distortions in the distance metric can potentially increase noise and therefore inflating type I error. This is particuarly important as often analyses include aggregation to multiple taxonomic ranks.

0.1 Isometric log-ratio transformation

One approach to solve this amalgamation issue is 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 is different than the often use *alr* and *clr* transformations, where the 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

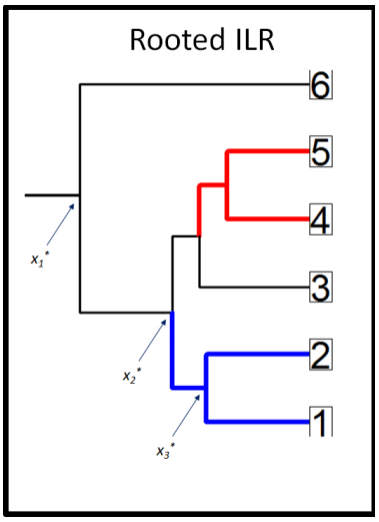

Figure 1: A sample SBP in tree form which is also the phylogenetic tree

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

the microbiome field, targeting transformations and dimension reduction along the phylogenetic tree [**?**, **?**]. Washburne et al. [**?**] 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 [6]. 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 Methods

1.1 Taxonomic enrichment analysis using isometric log-ratio transformations (TEA-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 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.

1.3 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.

Simulation studies

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. *Frontiers in Microbiology*, 8, 2017.

[5] 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.

[6] 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. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, September 2005.