cILR: Taxonomic Enrichment Analysis with competitive Isometric Log Ratios

Quang P. Nguyen, Dartmouth College Anne G. Hoen, Dartmouth College
H. Robert Frost, Dartmouth College

Introduction

Materials and Methods

Methods

Competitive Isometric Log-ratio (cILR)

The cILR method generates sample-specific enrichment scores for microbial sets using the isometric log-ratio transformation [1]. The cILR method takes two inputs:

- **X**: n by p matrix of positive counts for p taxa and n samples measured through either targeted sequencing p matrix of positive counts for p taxa and n samples measured through either targeted sequencing (such as 16S rRNA) or whole genome shotgun sequencing. Usually **X** is generated from standard sequence processing pipelines such as DADA2 [2] and MetaPhlAn2 [3].
- **A**: p by m indicator matrix annotation the membership of each taxa p to m sets of interest. These sets can be Linnean taxonomic classifications annotated using databases such as SILVA [4] or those based on more functionally driven categories such as the functional tropism of microbes ($a_{i,j} = 1$ indicates that microbe i belongs to set j.

The cILR method generates one output:

- **E**: n by m matrix indicating the enrichment score of m pre-defined sets as specified in **A** across n samples.

The procedure is as follows:

1. **Compute the cILR statistic**: Let **M** be a n by m matrix of cILR scores. Let $\mathbf{M}_{i,k}$ be the be cILR scores for set k of sample i:

$$\mathbf{M}_{i,k} = \sqrt{\frac{\sum_k A_{ik}(p - \sum_k A_{ik})}{p}} \ln\left(\frac{g(\mathbf{X}_{i,j}|\mathbf{A}_{j,k} = 1)}{g(\mathbf{X}_{i,j}|\mathbf{A}_{j,k} \neq 1))}\right)$$

   where $g()$ is the geometric mean. This represents the ratio of the geometric mean of the relative abundance of taxa assigned to set k and remainder taxa.
2. **Compute the cILR statistic on permuted X**: We seek to evaluate the empirical null distribution of the cILR statistic under H_o that relative abundances in **X** of members of set k are not enriched compared to those not in set k. Let $\mathbf{X}_p$ be the column permuted relative abundance matrix, and $\mathbf{M}_p$ be the corresponding cILR scores generated from $\mathbf{X}_p$.

3. **Fit Gaussian mixture distribution for each column of $\mathbf{M}_p$**
4. **Calculate finalized cILR scores as CDF values of the fitted mixture distribution**

Results

Discussion

Conclusion

1. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology. 2003;22.

2. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13:581–3.

3. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods. 2015;12:902–3.

4. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research. 2013;41:D590–6.