Annals of Epidemiology 26 (2016) 330–335

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

The Microbiome and Epidemiology

Compositional data analysis of the microbiome: fundamentals, tools, and challenges

Matthew C.B. Tsilimigras, Anthony A. Fodor PhD *

Department of Bioinformatics and Genomics, UNC Charlotte, Bioinformatics Building, The University of North Carolina, Charlotte 9201, University City Blvd, Charlotte

ARTICLE INFO

Article history:
Received 1 December 2015
Accepted 23 March 2016
Available online 31 March 2016

Keywords:
Metagenomics
Data interpretation, statistical
Microbiota
RNA, Ribosomal
16S
Statistics as topic
Selection bias
High-throughput nucleotide sequencing

ABSTRACT

Purpose: Human microbiome studies are within the realm of compositional data with the absolute abundances of microbes not recoverable from sequence data alone. In compositional data analysis, each sample consists of proportions of various organisms with a sum constrained to a constant. This simple feature can lead traditional statistical treatments when naively applied to produce errant results and spurious correlations.
Methods: We review the origins of compositionality in microbiome data, the theory and usage of compositional data analysis in this setting and some recent attempts at solutions to these problems.
Results: Microbiome sequence data sets are typically high dimensional, with the number of taxa much greater than the number of samples, and sparse as most taxa are only observed in a small number of samples. These features of microbiome sequence data interact with compositionality to produce additional challenges in analysis.
Conclusions: Despite sophisticated approaches to statistical transformation, the analysis of compositional data may remain a partially intractable problem, limiting inference. We suggest that current research needs include better generation of simulated data and further study of how the severity of compositional effects changes when sampling microbial communities of widely differing diversity.

Introduction

Compositional data are vectors of nonnegative elements constrained to sum to a constant. This simple feature of compositional data can have surprisingly adverse effects when traditional methods of multivariate statistics are naively used [1]. The dangers of ignoring the effects of compositionality were noted by Pearson, who recognized more than a century ago, that “spurious correlations” would result, should values constructed as proportions be compared haphazardly [2]. Compositional data is subject to the “closure problem” that occurs when components necessarily compete to make up the constant sum constraint [3]. This can cause large changes in the absolute abundance of one component to drive apparent changes in the measured abundance of others, violating the assumption of sample independence and creating inevitable errors in covariance estimates that can lead to bias and flawed inference. Diverse academic disciplines have begun to appreciate the complexity of the analysis of compositional data, ranging from forensics [4,5] and psychology [6] to the assessment of antibiotic use [7] and nutritional epidemiology [8].

In the case of the microbiome sequencing surveys, the compositional nature of the data comes from the fact that a correction must be made for different samples having different numbers of sequences while the total absolute abundance of all bacteria in each sample is unknown. These complications arise from sample collection, polymerase chain reaction (PCR) amplification, and the sequencing technology itself from which the absolute abundances of bacteria are not recoverable from sequence counts, but the proportions of different taxa are still relevant. Numerous schemes are used in the literature to convert the number of sequences for each taxon within each sample to relative abundance with popular techniques, including proportional abundance and rarefying, the latter being the default choice in the popular Quantitative Insights Into Microbial Ecology pipeline [9,10]. Neither of these approaches corrects for compositionality and it has been argued that this lack of correction has led to erroneous analyses that fail to discriminate between true and spurious correlations between taxa [11,12]. However, it remains unclear whether these sorts of normalization schemes routinely produce spurious correlations in the study of complex microbial communities, like the gut, or whether errors due to compositionality are instead restricted to analysis of microbial communities where only a few taxa dominate, such as the vaginal microbiome.

* Corresponding author. Department of Bioinformatics and Genomics, The University of North Carolina; Charlotte 9201, University City Blvd, Charlotte, NC 28223-0001. Tel.: 704-687-8214.
E-mail address: anthony.fodor@gmail.com (A.A. Fodor).

http://dx.doi.org/10.1016/j.annepidem.2016.03.002
1047-2797/© 2016 Elsevier Inc. All rights reserved.

In this review, we examine the historical literature on the compositionality problem and some modern approaches to its solution that have been proposed for the analysis of next-generation sequencing data sets. We track recent progress and indicate where we think more research is needed. We also emphasize that the analysis of compositional data will always be at least a partially intractable problem despite the development of sophisticated statistical transformations as the absolute abundances of microbes before sequencing can never be recovered from sequence data alone, and this will inevitably color inference based on compositional samples.

Compositional data sets are best analyzed after a log-ratio transformation

The initial literature on compositional data analysis has largely been attributed to a pioneering author, John Aitchison, whose classic treatise, "The Statistical Analysis of Compositional Data," has remained enormously influential for nearly 3 decades [3]. However, Aitchison, developing his theory in the 1980s, was analyzing data sets considerably smaller than those of current next-generation sequencing. His examples were often sourced from geology and usually featured problems such as how different mineral components were used to categorize variability in rock specimens. Despite the relative simplicity of the data sets he analyzed, the theory Aitchison developed was surprisingly complex. His work eventually led to the realization that the unit-sum constraint yielded a new geometrical space requiring a substantial background in advanced multivariate linear algebra to fully appreciate. A central challenge for researchers wanting to apply these elegant mathematical formalisms to modern genomics data is the complexity of sequencing data sets, which, unlike simple geology data sets, can have tens of thousands of different categories (high dimensionality), have zeros dominating all other values (sparsity), and have a number of samples substantially fewer than the number of variables (underdetermination) [13,14]. Aitchison recognized these problems but does not offer complete solutions to them in his treatise, and attempts to satisfactorily address these difficult compositional data sets continue to the current day.

Aitchison argued that taking the logarithm of ratios is a transformation of compositional data that restores much of the utility of traditional statistical analyses in situations such as relative abundance. This transformation is structured so that the constant sum constraint does not distort the underlying covariance or correlation structure originating from the natural interaction of the components [3]. A natural problem in using a ratio-based transformation is that one has to choose what will be in the denominator; that is to say, which value to use to normalize all the values in a sample. Aitchison considered two possible transformations in his text, both of which are still in use. The simplest transformation is to choose one component as a reference. For example, in a metagenomics experiment analyzed at the phyla level, one could choose as a reference the phyla "Firmicutes." Then all other taxa would be reported as a ratio of each taxa to Firmicutes. Although compositionality was not the motivation, this was in fact the transformation that was used in an early landmark study of the human microbiome, which reported that the ratio of Bacteroidetes to Firmicutes was associated with obesity in a human population (interestingly this observation has proven to be difficult to replicate[15]). Choosing reference taxa has the advantage of simplicity, but there may not always be an obvious reference to choose and results may vary substantially dependent on the choice of reference [13]. One solution might be to systematically perform inference on every possible pair of taxa, but performing N^2 analyses, and then correcting for N^2 multiple hypotheses is not usually feasible given the large numbers of distinct taxa in many metagenomic analysis pathways. Aitchison called this simple choice of using one reference taxon and taking the logarithm an "additive log-ratio" (alr). As an alternative, Aitchison recommended transforming each taxon within a sample by taking the log-ratio of the counts for that taxon divided by the geometric mean of the counts of all taxa, called the centered log-ratio (clr). This approach is necessarily more robust than the additive log-ratio as it does not depend on the choice of an arbitrary reference. This algorithm has found use in the current microbial literature [16] where it was argued that this transformation could be used to successfully analyze microbiome data as well as RNA-seq data and, indeed, any next-generation sequence data set. Egozcue et al. [17] later defined a third isometric log-ratio transformation (ilr), which is the product of the clr and the transpose of a matrix which consists of elements that are clr-transformed components of an orthonormal basis. This ilr transformation is an orthonormal isometry that addresses certain difficulties of alr and clr, but its interpretability is subject to the selection of its basis, which has somewhat limited its adoption [17].

Although the centered log-ratio has mathematical elegance and has found sophisticated champions in the current metagenomic literature, it has potential problems when applied to metagenomic data sets. This difficulty arises from extreme variability of library sizes and the great sparsity of metagenomic data sets. In a highly sparse data set, the geometric mean of all taxa can often be zero or near zero. Obviously, if it is zero, a transformation that involves dividing by the geometric mean is undefined. One can of course correct for this by adding a pseudo-count to each cell, but it is not immediately clear what the value of this pseudo-count should be. For example, if the value 1 is chosen for the pseudo-count, then dividing by the geometric mean in a highly sparse data set is equivalent to simply not normalizing the data (because you are dividing by 1 before the log transformation). Performing statistical inference on unnormalized data will often lead to results that do not reveal biological variability, but merely reflect differences in sequencing depth [18]. For example, Weiss et al. [19] has shown that the first principal coordinate analysis axes of data sets are often well correlated to the number of sequences per sample. This problem is not ameliorated by a transformation such as taking geometric mean while using a small pseudo-count (Fodor lab, unpublished data).

One could choose some other value for the normalizing counts other than the geometric mean. Packages made for RNA-seq data, DESeq, for example, use values based on medians or certain percentiles in the denominator [20,21]. This offers some of the advantages of the geometric mean, but there is still no guarantee that even very high percentiles of a metagenomics data set do not yield zeros subject to the routinely encountered sparsity. One article [18] recommends the use of RNA-seq pipelines for analysis of metagenomic data but does not offer much guidance on how best to set the normalizing threshold to avoid normalizing by zero or the pseudo-count.

Another problem related to sequencing depth in metagenomic experiments is the difficult decision of when to remove samples that have few sequences [18,22]. In general, these samples tend to be outliers. The low number of sequences in such samples may reflect a PCR error or indicate a sample in which there was no input microbial DNA and the sequences reflect kit microbes or other artifacts [23]. However, it is not clear how to define the cutoff value that indicates that a sample has so few sequences that it should be removed from downstream analysis. This difficult decision of sequence count thresholds impacts the corrections for compositional data described previously in ways that are not fully appreciated as the compositionality corrections work in relative space, but the decision to threshold is in absolute space, and the interaction of making decisions in these two spaces is unclear.

It should be stressed that even with all the algorithms that have been developed to appropriately analyze compositional data