Annals of Epidemiology 26 (2016) 322–329

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

The Microbiome and Epidemiology

It's all relative: analyzing microbiome data as compositions

Gregory B. Gloor PhD [a],*, Jia Rong Wu BSc [a], Vera Pawlowsky-Glahn PhD [b], Juan José Egozcue PhD [c]

[a] Department of Biochemistry, University of Western Ontario, London, Ontario, Canada
[b] Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain
[c] Department of Civil and Environmental Engineering, Technical University of Catalonia, Spain

ARTICLE INFO

Article history:
Received 29 October 2015
Accepted 23 March 2016
Available online 2 April 2016

Keywords:
Compositional data
16S rRNA gene sequencing
Microbiome
Multivariate analysis

ABSTRACT

Purpose: The ability to properly analyze and interpret large microbiome data sets has lagged behind our ability to acquire such data sets from environmental or clinical samples. Sequencing instruments impose a structure on these data: the natural sample space of a 16S rRNA gene sequencing data set is a simplex, which is a part of real space that is restricted to nonnegative values with a constant sum. Such data are compositional and should be analyzed using compositionally appropriate tools and approaches. However, most of the tools for 16S rRNA gene sequencing analysis assume these data are unrestricted.
Methods: We show that existing tools for compositional data (CoDa) analysis can be readily adapted to analyze high-throughput sequencing data sets.
Results: The Human Microbiome Project tongue versus buccal mucosa data set shows how the CoDa approach can address the major elements of microbiome analysis. Reanalysis of a publicly available autism microbiome data set shows that the CoDa approach in concert with multiple hypothesis test corrections prevent false positive identifications.
Conclusions: The CoDa approach is readily scalable to microbiome-sized analyses. We provide example code and make recommendations to improve the analysis and reporting of microbiome data sets.

Introduction

High-throughput sequencing has provided the laboratory tools needed for the large-scale culture-independent analysis of microbial communities. However, studies often fail to replicate earlier work even when similar technologies and strategies are used. For example, four recent articles indicating a link between autism and the gut microbiota have implicated many different associated genera. In mouse models, de Angelis et al. [1] identified 74 differential operational taxonomic units (OTUs) composing at least 90% of the sequence reads obtained, whereas de Theije et al. [2] reported only three, and Hsiao et al. [3] reported 67. Although it is difficult to directly compare results across studies because of different sequencing and bioinformatic platforms, it is interesting to note that the three groups had very little overlap in taxonomically assigned genera, and the same named genus could exhibit statistically significant change in different directions in different experiments. In an additional study on humans, Kang et al. [4] identified five additional taxa that distinguished Autism Spectrum Disorder from neurotypical control patients. However, as shown below, examination of these data sets suggests that their conclusions could be explained by chance alone. Although these autism studies serve as facile examples, the literature on the microbiome of other conditions are replete with similar experiments with the same shortcomings.

Hanage [5] recently called for a skeptical re-examination of microbiome research results by posing five questions, with the first four being: Can experiments detect differences that matter? Are we examining correlation or causation? Is there a mechanism? Do the experiments reflect reality? These questions are beginning to be examined and addressed in detail by others.

Hanage's final question, "could anything else explain the results," was the most troubling because it is clear that the answer to the last question is a resounding yes! For example, work from the Microbiome Quality Control Consortium (http://www.mbqc.org) shows that the wet laboratory, sequencing and computational approaches used can affect the quality, quantity, and scope of the data, and thus, the conclusions reached. The consortium effort has already led to changes that make the analyses more reproducible (see e.g., [6–8]). A second example has been the realization that the

* Corresponding author. Department of Biochemistry, University of Western Ontario, London, Ontario N6A5C1, Canada. Tel.: +1-5196613526; fax: +1-5196613175.
E-mail address: ggloor@uwo.ca (G.B. Gloor).

http://dx.doi.org/10.1016/j.annepidem.2016.03.003
1047-2797/Crown Copyright © 2016 Published by Elsevier Inc. All rights reserved.

reagents themselves often contribute contaminants to samples that contain low amounts of inputs [9,10].

A further unappreciated problem is that of the sample space itself. We contend the data generated by high-throughput sequencing are multivariate and constrained by an arbitrary constant sum. The constant sum constraint is imposed because all sequencing instruments have a fixed upper bound on the number of reads delivered. Data with this constraint are referred to as compositional data (CoDa) [11], and the only information that can be obtained from such a data set is that of the ratios between the parts. We argue that it is more powerful and appropriate to examine these data using the existing tools designed for CoDa that have been used in the fields of geology and ecology [12]. Furthermore, we argue that sequencing data should be treated when possible in a Bayesian manner as probability distributions rather than as point estimates. Finally, we contend that proper statistical practice of correcting for multiple hypotheses is essential since the data are multivariate.

It may be helpful to draw an analogy between CoDa and proportional mortality. In both instances, the true denominator is not known and only relative information is obtained. In addition, variables with small numbers of events (OTUs represented by low counts, or deaths by rare causes) will be seen to be highly variable. Thus, a cautious interpretation of the data set is required for CoDa sets as it is for proportional mortality data sets.

In this article, we demonstrate the utility of the CoDa frame of reference by comparing the buccal mucosa and tongue dorsum samples from the Human Microbiome Project. The CoDa approach almost completely separates samples from these adjacent sites, and many distinguishing OTUs can be identified. We then re-evaluate an available autism data set with this framework and make recommendations for future work.

The origin of high-throughput sequencing data

Data sets for 16S rRNA gene sequencing are generated from polymerase chain reaction amplified random environmental samples of DNA molecules. We know that the total number of molecules varies by sample. For example, total bacterial load is quite different based on the environment (e.g., stool vs. mucous membranes), or across different phases of a cellular growth cycle. The data returned are random samples of the molecules in the environment, and each sample is subject to an arbitrary constant sum constraint imposed by the sequencer itself. Thus, the total number of reads assigned to an OTU can provide no information about the number of molecules in the original sample, and we can only investigate relative changes. This limitation is acknowledged when investigators treat 16S rRNA gene sequencing as proportions, percentages, or “relative abundances” in the data analysis, and by the RNA-seq convention of normalizing the counts across samples, an approach that has been advocated for microbiome studies [13]. However, these two approaches both effectively normalize all samples to a common denominator: relative abundance uses a constant denominator of 100, and the various count normalization approaches use an empirically determined denominator unique for each experiment [14].

Problems in the analysis of compositional data

A composition quantitatively describes parts of some whole. Parts are grouped in a vector of D positive components, $x = (x_1,x_2,\ldots,x_D)$. The composition is said to be closed when the sum of all components add up to a constant, for instance 1, 100, or a million. The major issue with these data is that the only relevant information is contained in the ratios between components [11,15]. This property means that a composition can be multiplied by any positive constant without any change in its meaning. Thus, vectors with proportional positive components are equivalent from the compositional point of view [15–17]. Logarithmic transformation of the ratios ensures that the scale of ratios is symmetrical so that the permutation of numerator and denominator only causes a change of sign and places values into an absolute scale. This so-called log-ratio approach [11] allows inferences about CoDa to be performed on logarithms of ratios, which do not change when the composition is multiplied by a positive constant. Examples of these combinations are simple log-ratios as $\ln(x_i/x_j)$, or more complex log-ratios as $\ln((x_1x_2)^{1/2}/(x_3x_4x_5)^{1/3})$; they are called log-contrasts. Remarkably, log-contrasts can be computed from nonclosed compositions.

The analysis of compositions in a raw form has several inconsistencies. As first noted by Pearson in 1897 [18], the analysis of compositions with a constant sum results in spurious correlations. Several pitfalls can be detected even when the compositions are not closed.

Figure 1 uses a simple example to illustrate one issue with the constant sum. Commonly an experimentalist wishes to count and compare the total number of molecules in two samples. Samples may contain many OTUs which values are independent and with similar abundance, and here one OTU has a 10-fold difference. A scatter plot of these two samples in the “Counts” panel shows that we would infer by counting that the OTU represented by the blue circle has increased 10-fold in sample B compared with sample A, but that the rest are essentially unchanged.

The “Proportions” panel shows the same data when these samples are subjected to DNA sequencing and random sampling in which case counting is no longer valid. The data are now constrained to a constant sum: note that any constant sum is exactly equivalent to a proportion and a constant scaling factor. All information regarding the total number of molecules in each sample is lost and only relative information remains. In the scatter plot, the distortion introduced by the constant sum is readily apparent: the invariant majority appear to have become less abundant, whereas the blue OTU appears to become more abundant: however, it increases only from 6.8% in A to 41.9% in B. Note that this is not linearly related to the actual 10-fold difference. If we had many such samples, we would infer that the red OTUs are positively correlated (they all went down together!), and that they are negatively correlated with the blue OTU. This spurious correlation is caused only by restricting the data to have a constant sum.

Spurious correlation also appears when reducing a full composition to a composition with fewer parts: that is, dealing with a so-called subcomposition. Note that 16S rRNA gene sequencing data sets are always subcompositions because the constituent OTUs that are found in the samples depend on arbitrary decisions made during the data analysis pipelines (see the filtering and cutoff values recommended by Quantitative Insights Into Microbial Ecology [19] and mothur [20] for examples).

As an example of this problem, consider a vector of proportions like $x = (x_1,x_2,x_3,x_4)$, which is reduced to $y = C(x_1,x_2,x_3)$, where $C = 1/(x_1+x_2+x_3)$. Thus, y is a vector of proportions composed of only three first parts of x: clearly y is a subcomposition of x. Table 1 illustrates the sample covariance matrices for x and y for a synthetic data set of counts produced by simulating four log-normal variables with a given covariance structure (see the subcomposition code block in the supplement). If we are interested in evaluating the sample correlation between the first three components, we can see that these values are markedly different in the complete composition x than in the subcompositon y. In fact, none of the correlations involving the three initial components coincide, and differ substantially with no clear rule, thus deserving the name of spurious correlation. While not observed here, it is worth noting that a change of sign in the correlation is commonly found.