Journal of Biotechnology 250 (2017) 45–50

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

A simple statistical test of taxonomic or functional homogeneity using replicated microbiome sequencing samples

Daniel H. Huson[a,*], Mike Steel[b], Mohamed El-Hadidi[a], Suparna Mitra[c], Silke Peter[d,e], Matthias Willmann[d,e]

[a] Center for Bioinformatics, University of Tübingen, Germany
[b] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand
[c] Faculty of Medicine and Health – University of Leeds, Old Medical School, Leeds General Infirmary, Leeds, UK
[d] Institute of Medical Microbiology and Hygiene, University of Tübingen, Germany
[e] German Center for Infection Research (DZIF), Partner Site Tübingen, Germany

ARTICLE INFO

Article history:
Received 17 August 2016
Received in revised form 20 October 2016
Accepted 24 October 2016
Available online 27 October 2016

Keywords:
Metagenomics
Taxonomic composition
Functional diversity
Environmental inhomogeneity
Statistical testing

ABSTRACT

One important question in microbiome analysis is how to assess the homogeneity of the microbial composition in a given environment, with respect to a given analysis method. Do different microbial samples taken from the same environment follow the same taxonomic distribution of organisms, or the same distribution of functions? Here we provide a non-parametric statistical "triangulation test" to address this type of question. The test requires that multiple replicates are available for each of the biological samples, and it is based on three-way computational comparisons of samples. To illustrate the application of the test, we collected three biological samples taken from different locations in one piece of human stool, each represented by three replicates, and analyzed them using MEGAN. (Despite its name, the triangulation test does not require that the number of biological samples or replicates be three.) The triangulation test rejects the null hypothesis that the three biological samples exhibit the same distribution of taxa or function (error probability ≤ 0.05), indicating that the microbial composition of the investigated human stool is not homogenous on a macroscopic scale, suggesting that pooling material from multiple locations is a reasonable practice. We provide an implementation of the test in our open source program MEGAN Community Edition.

1. Introduction

There is growing interest in microbiome analysis, the study of microorganisms in a particular environment, such as the human body, water or soil. Taxonomic profiling of microbiome samples is usually based on DNA sequencing, using either an amplicon or a shotgun approach (Handelsman, 2004). While early studies typically involved small numbers of samples, there is now an increased expectation that replicate samples are sequenced (Knight et al., 2012) so as to allow statistical analysis of the significance of obtained results.

One important question is how to assess the homogeneity of the microbial composition in a given environment. Do different samples taken from the same environment follow the same taxonomic distribution of organisms? Do they follow the same distribution of functional features? Presumably, well-mixing environments such as air or water will display a higher degree of compositional homogeneity than more structured environments such as soil, stool or waste water flocks.

We need to introduce two key concepts. First, we assume that multiple samples are taken from the same given environment. For example, in the case of human stool, one might collect samples from a number of different locations in the same piece of fecal matter. We refer to these samples as *biological samples*. Second, we assume that each biological sample gives rise to two or more *replicates*, which are obtained by constructing multiple DNA libraries for each biological sample.

In addition, we emphasize that the whole chain of steps employed, from sampling, DNA extraction, sequencing, calculation of taxonomic or functional profiles, to the comparison of such profiles using a dissimilarity measure, is part of the input to the problem, and we will use M to denote the complete analysis procedure.

We describe a novel statistical test, which we call the *triangulation test*, that aims at testing the null hypothesis that different biological samples taken from a given environment exhibit the

* Corresponding author.
E-mail address: Daniel.Huson@uni-tuebingen.de (D.H. Huson).

http://dx.doi.org/10.1016/j.jbiotec.2016.10.020

same distribution of taxa, with respect to M. The triangulation test can be applied to a wide range of different sampling regimes involving any number of biological and replicates. This is a nonparametric test and is thus robust against assumptions of population distribution.

In Section 2, we first formulate the triangulation test for the 3×3 case of three biological samples, each represented by three replicates. We then generalize the triangulation test to the case of any number of biological samples, each represented by an arbitrary number of replicates.

In Section 3, we demonstrate the application of the triangulation test in the 3×3 case using human stool samples. In more detail, we collected three samples from different locations (approximately one centimeter apart) in the same piece of stool and sequenced three replicates for each. Application of the triangulation test implies that the microbial composition of the investigated human stool is not homogenous on the macroscopic scale.

We use a series of artificially constructed mixtures of samples to study the performance of the method in lower contrast settings and show that it performs as well as PERMANOVA.

2. Methods

2.1. Triangulation test for 3×3 samples

Assume that we are given three biological samples A, B and C, and for each biological sample $S = A$, B or C we are given three replicates, S_1, S_2 and S_3. We thus have nine samples in total: A_1, A_2, A_3, B_1, B_2, B_3, C_1, C_2, C_3. Moreover, let M be a specific chain of analysis steps that provides taxonomic profiles for all samples, and a dissimilarity measure.

The null hypothesis and the alternative hypothesis of interest are stated as follows:

- H_0: the nine samples $A_1, \ldots, C_3$ are all drawn independently according to the same distribution, versus
- H_a: For each of the three biological samples $S = A$, B or C, the three replicates S_1, S_2, S_3 are drawn from the same distribution for S, but the distributions for each of the biological samples are not all identical. Moreover, replicates within the same biological sample expected to be more similar to each other than any ones that lie in different biological samples.

To address these hypotheses, we define the *triangulation test* as a simple non-parametric significance test for H_0. It is based on the concept of a *random triangulation* of the set of samples.

For ease of exposition, we first describe this for the special setting used in our practical study, namely three biological samples, each represented by three replicates. Below, we will then present a general version of this test that allows for any number of biological samples, each with its own arbitrary number of replicates.

We define a *triangulation* of the set of nine replicates $A_1, A_2, \ldots, C_3$ to be a graph $G = (V, E)$ with node set $V = \{A_1, A_2, \ldots, C_3\}$ and edge set E consisting of undirected edges that form node-disjoint 3-cycles that each involve replicates from exactly two different biological samples. We desire that any such triangulation contains as many triangles as possible, i.e., three triangles as in Fig. 1.

We assume that an analysis of the taxonomic content (or functional content, if desired) of each replicate has been performed using M and we have obtained a dissimilarity measure $d(S_i, T_j)$ between any two replicates $S_i, T_j \in V$.

The triangulation test is performed in two steps. In the first step, we randomly choose a single triangulation that involves all replicates. In Fig. 1 we show one such choice for the case of three biological samples, each represented by three replicates. As required, each depicted triangle contains two replicates from the same biological sample, and a third replicate from a different biological sample.

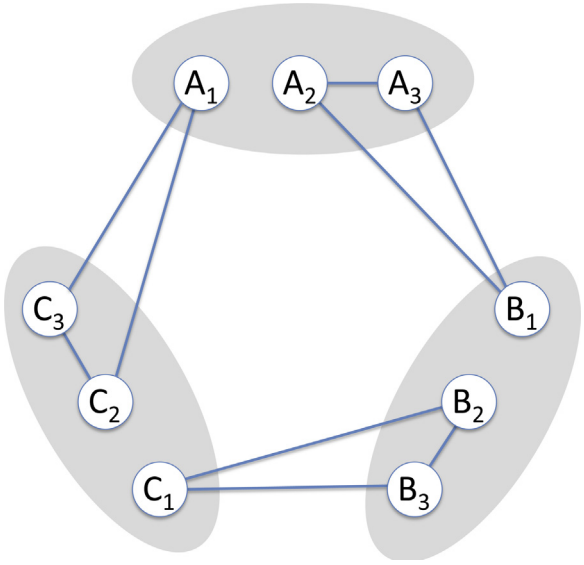

Fig. 1. Triangulation test. The nine replicates $A_1, \ldots, C_3$ (represented by labeled discs) come from three biological samples (indicated by gray ovals). Lines connecting discs represent a triangulation of the data. The null hypothesis is rejected, if for each and every triangle, the distance between replicates from the same sample is less than the distance from either of those replicates to a replicate from a different sample.

In the second step, we then ask the following question: For each triangle in the chosen triangulation, is the dissimilarity between the two replicates that are contained in the same biological sample less than the dissimilarity of either of them to the third replicate in the triangle? If the answer is yes for all triangles, then reject H_0 with a significance level that is less than the 0.05.

To see this, note that under the null hypothesis H_0, all three pairs of replicates in a triangle have equal probability of exhibiting the strictly smallest dissimilarity value, 1/3 (or less, in the case of a tie). Thus, the probability that it is always the pair of replicates contained in the same biological sample that has the smallest value in all three triangles is at most $(1/3)^3 = (1/27) < 0.05$. Hence, the probability of rejecting the null hypothesis when it is true (i.e., the type 1 error) is less than 0.05.

Whether the triangle test is able to reject H_0 depends not only on the samples, but also on the details of the analysis method M, we would like to emphasize. For example, if the samples are analyzed using a low resolution method M that can only detect the general presence or absence of bacteria, say, and if bacteria are detected in all samples, then the dissimilarities will be constant, in which case the test will fail to reject the null hypothesis. However, it is of course possible that the use of a higher resolution analysis method will lead to rejection of the null hypothesis on the same data.

We would also like to emphasize that application of the test only involves the use of one randomly chosen triangulation. The test does not require that one looks at multiple triangulations. In particular, this makes the test very easy to apply "by hand".

2.2. Triangulation test in the general setting

We now describe how the simple triangulation significance test described above extends to the general setting where we have m bins (i.e., biological samples) of arbitrary sizes $r_1, \ldots, r_m$. Let $V = \{S_i(j) \mid j = 1, \ldots, r_i\}$ be the set of replicates present in bin i, for $i = 1, \ldots, m$. (Above we studied the special case of $m = 3$ and $r_i = 3$ for $i = 1, \ldots, 3$.) The total number of replicates is $N = \sum_{i=1}^{m} r_i$. Again,