Exploration of the Decentralization Method.

Hallett group

October 2019

We explore the so-called decentralization method that tries to adjust observed count matrices in comparative metagenomic analyses.

0.1 Introduction

The basic observation here is that current metagenomic studies are relativistic in nature. A sample is taken from two more more sites. Although the absolute number of organisms in two samples of the same size from the two sites might differ, most modern -omic technologies require a specific amount of starting material For example, a specific number of micrograms of cDNA will be required for next generation sequencing, or a specific number of micrograms of protein will be required for mass spectrometry. Therefore, biomaterial harvested from the sites is either amplified or distilled to this level, meaning all subsequent analyses are relativistic in nature. Specifically, each samples produces a raw number of observations and these numbers may differ across the sampling sites. However, they cannot be attributed to biological differences (eg the number of micro-organisms per microlitre); rather due to technological variance (eg differences in calibration of the machine or sample preparation).

Each site ina metagenomic study essentially provides a frequency vector across all organisms that can be used to compare the relative concentrations between organisms within that site. However it is problematic to compare the abundance of two organisms between sites (without some a priori information regarding the normalized absolute counts).

An issue that arises in several studies including our study of two sites of the Barbados reef is that a very small number of organisms are greatly increased at one of the two sites but the remainder of the organismal population does not change significantly. The frequency vector across all organisms between the two sites can suggest dramatic global changes. A common pattern is that components of the frequency vector for the site whose population has increased greatly for a handful of organisms is almost every where else smaller than the components of the frency vector of the second population. In other words, the first population's distribution has become centralized. We discuss a strategy for removing decentralizing these distributions in order to better detect true absolute differences between the populations at each site.

We will need the functions and datasets from the Barbados reef project later in our analysis.

```r
options(warn = -1)
setwd("~/repo/reefmicrobiome/experiments/exp5-decentralization")
library(xtable)
source("~/repo/reefmicrobiome/src/functions.R")
load("~/repo/reefmicrobiome/data/tree.latest")
tree[1,]
```

```
##   name tax_id parent    rank embl_code division_id  br_bel   br_may
## 1 root      1     NA no rank                     8 4708322 10181105
##   br_bel_frac br_may_frac Local.Freq.Bel Local.Freq.May Glob.Freq.Bel
## 1           0           0             NA             NA             1
##   Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
## 1             1         0        0       NA           NA root
```