Interpretation and application of absolute abundance in Weighted UniFrac distance

Augustus Pendleton[1]* & Marian L. Schmidt[1]*

[1]Department of Microbiology, Cornell University, 123 Wing Dr, Ithaca, NY 14850, USA

Corresponding Authors: Augustus Pendleton: arp277@cornell.edu; Marian L. Schmidt: marschmi@cornell.edu

Author Contribution Statement: Both authors contributed equally to the manuscript.

Keywords: Microbial Ecology - Beta Diversity - Absolute Abundance - Bioinformatics

The UniFrac distance was first introduced by Lozupone & Knight in 2005, and has since become enormously popular as a measure of β-diversity within the field of microbial ecology [1]. A major draw of the UniFrac distance is that is considers phylogenetic information when estimating the distance between two communities. After first generating a phylogenetic tree representing species (or proxies like ASVs/OTUs) from all samples, the UniFrac distance computes the fraction of branch-lengths which is *shared* between communities, relative to the total branch length represented in the phylogenetic tree. UniFrac can be both *unweighted*, in which only the incidence of species is considered, or *weighted*, wherein the contribution a branch makes to the overall distance is weighted by the proportional abundance of taxa descended from that branch. The weighted UniFrac (U^W) is derived as:

$$U^R = \frac{\sum_{i=1}^{n} b_i |\frac{s_i^a}{A} - \frac{s_i^b}{B}|}{\sum_{i=1}^{n} b_i (\frac{s_i^a}{A} + \frac{s_i^b}{B})}$$

Where we weight the length of each branch, b_i, by the difference in the observed abundance of all species (s_i) descended from that branch in sample A or sample B (numerator), dividing by the sum of species in both samples (denominator). We normalize the observed abundances by dividing by the total observations in each sample (A and B), hence giving relative abundances. As such, we denote this distance as U^R, for "Relative Unifrac". Popular packages which calculate weighted Unifrac, including QIIME and the R packages `phyloseq` and `GUniFrac` run this normalization by default. However, if one wishes to use absolute abundances, weighted Unifrac can instead by written as:

$$U^A = \frac{\sum_{i=1}^{n} b_i |s_i^a - s_i^b|}{\sum_{i=1}^{n} b_i (s_i^a + s_i^b)}$$

We were curious as to the implications of using this "Absolute Unifrac" (U^A) compared to U^R, and the impact of absolute abundances between phylogenetic and non-phylogenetic distance measures like the Bray-Curtis Dissimilarity.

References

1. Lozupone C, Knight R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology* 2005;**71**:8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005