The Report Committee for Kiona Natasha Pilles
Certifies that this is the approved version of the following report:

Unweighted Unifrac Is A Theoretically Better Measure For Dietary and Cardiometabolic Data

APPROVED BY

SUPERVISING COMMITTEE:

Dan Powers, Supervisor

Jaimie Davis

Unweighted Unifrac Is A Theoretically Better Measure For Dietary and Cardiometabolic Data

by

Kiona Natasha Pilles

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2018

Dedication

Dedicated to myself for surviving this long off of coffee and ramen noodles.

Acknowledgements

I would like to acknowledge the following contributors who reviewed and approved this final manuscript:

Dan Powers for his constant patience and guidance and for being such a steady and fair person in this academic atmosphere. Also to Jaimie Davis for her continuous support from conception to design of this study, all the way to publication.

Abstract

Unweighted Unifrac Is A Theoretically Better Measure For Dietary and Cardiometabolic Data

Kiona Natasha Pilles, MSStat

The University of Texas at Austin, 2018

Supervisor: Dan Powers

The objective of this report was to compare unweighted and weighted Unifrac statistical methods and decide which method is best for analyzing dietary and cardiometabolic data. The Freshmen Health Study (n=77), a study on exclusively Hispanic college students collected anthropometric, dietary, cardiometabolic, and microbiome data. Weighted and unweighted Unifrac were used to analyze differences in the microbiome between groups of dietary and cardiometabolic variables. The results showed that unweighted Unifrac was the only significant measure and is statistically better for analyzing this type of data because of the sampling method of selecting only unique sequences to each community and analyzing their similarities. This is important for detecting subtle changes in different groups because diets are composed of different compositions that can influence the gut microbiome in small amounts.

Table of Contents

Chapter 1: Introduction To Microbiome Analysis

Many microbiome analyses are performed using a system called QIIME, a bioinformatics pipeline that takes raw DNA sequencing data and generates analyses and visualizations.[1] This pipeline takes 16s amplicon sequencing of bacterial RNA, groups the sequences that are 97% similar together, and classifies them into an operational taxonomic unit (OTU).[2] The OTUs are then used to analyze microbiome diversity.

UniFrac is a technique that measures the distance between microbial communities, counting the absence, presence, and abundance of OTUs and was devised by Catherine Lozupone and Rob Knight.[3] There are many different ways to measure the similarity or dissimilarity between predefined groups, but two commonly used measurements are unweighted and weighted Unifrac analysis. Unweighted Unifrac is "the *distance* between community *A* and community *B* and is defined as the fraction of branches of the phylogenetic tree that lead to members of community *A* or community *B*, but not both."[4] The equation for unweighted Unifrac is: $\beta = (A_i - c) + (B_i - c)$ where A_i is the number of OTUs that descend from branch *i* in community *A* that is *unique* to community *A*, and B_i is the number of OTUs the descend from branch *i* in community *B* that is *unique* to community *B*, and *c* is the number of common or shared taxas.[3]

Weighted Unifrac is defined as the *dissimilarity* between two communities where length between communities is weighted according to abundance in community *A* in proportion to the total, compared to the abundance in community *B* in proportion to the total.[4] The equation for weighted Unifrac is: $\sum_{i=1}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$ where A_i and B_i are the

number of OTUs that descend from branch i in communities A and B, respectively, A_T is the overall abundance of OTUs in community A and B_T is the overall abundance of OTUs in community B, n is the total number of branches in the tree, and b_i is the length of branch i.[4,5]

While both unweighted and weighted Unifrac are used to analyze differences in microbiome communities, there has been no consistency in dietary studies as to which is significant. In some cases, only the weighted Unifrac is significant[6] and others only the unweighted[7–9]. This poses the question as to whether there is a theoretically better measure for comparing microbiome communities between subjects. In the case of diet and clinical biomarkers, when an individual's daily consumption of macro and micro-nutrients vary so widely from day to day and thereby their clinical measures, does one measure make more sense than the other? In the Freshmen Health Study, a study of an exclusive freshmen college Hispanic population, anthropometrics, blood lipids, diet, and microbiome were analyzed in students. After analysis of diet, cardiometabolic risk factors and the gut microbiome in this population, results were consistently significant *only* in the unweighted measures. This paper aims to explore why only the unweighted Unifrac was significant and what are the differences in measurement between unweighted Unifrac and weighted Unifrac analysis in diet and clinical biomarker data in relation to the microbiome.

Chapter 2: Comparing Measurements

Comparing unweighted versus weighted Unifrac, it is clear to see that these methods analyze the results for different purpose. When two communities have the same taxa, the unweighted Unifrac difference would be zero and the communities would be deemed not significantly different from each other making unweighted Unifrac useful for finding distinctly different OTUs between two communities and clearly separating them. This equation is excellent for finding differences between microbiome groups[3] as the sampling method only takes into account species that are unique to each individual community and assigns it a beta-diversity number. However, this equation is limited because the equation would generate insignificant results if there are similar species, but in different abundances,[10] making it susceptible to "noise" by presenting shallow differences.[3,4] Although microbial groups with the same types of gut bacteria would indicate no significant difference, the microbiome could still be significantly different in composition.

On the other hand, weighted Unifrac is useful for suppressing shallow differences by normalizing the data and can detect differences in OTU abundances making it useful for finding differences in bacterial count.[4,10] When two communities have the same OTU to abundance ratio, the weighted difference would be zero and the communities would be deemed not significantly different. Therefore, this equation is excellent for finding differences between microbiome composition as the sampling method takes into account species and count in each individual community and allows overlapping of OTU between the communities. However, this equation is limited because abundance count can drown

out small branches.[4] Knowing the strengths and limitations of each equation, one must decide which is better when analyzing dietary and clinical biomarker data.

In the case of the Freshmen Health Study, all of the relevant dietary and cardiometabolic variables were indicated to have significance exclusively in the unweighted model and not the weighted model. For example: saturated fat was grouped into tertiles and also grouped by dietary recommendations. Both were analyzed separately. Both times, the unweighted model was significant and the weighted model was not ($p=0.007^*$ vs. 0.604 for tertiles, $p=0.014^*$ vs. 0.684 for dietary recommendations). When microbiome biodiversity was further analyzed there was indeed a Shannon biodiversity index of increased diversity of the microbiome in subjects who met saturated fat recommendations compared to those who exceeded recommendations (5.21 ± 0.90 vs. 4.92 ± 0.52; $p=0.01$). The analysis was able to further specify which bacteria were contributing to the significant difference. This was the same case with body fat ($p=0.023^*$ vs. 0.152), insulin ($p=0.048^*$ vs. 0.406), and low-density lipoprotein ($p=0.020$ vs. 0.699). Each time the variables that achieved significance only achieved significance in the unweighted model but not in the weighted model.

It is clear from the Freshmen Health Study that the unweighted model was able to detect relevant significant differences in microbiome diversity and composition in dietary and cardiometabolic data. However, the question remains as to why this happens.

Chapter 3: Dietary and Clinical Biomarker Data

Grouping diet and analyzing the differences between groups is not like grouping demographics. For example, an analysis of age produces no variations in regards to time, while a carbohydrate can be quantified into either grams per day or a percentage of daily calories consumed. In addition, carbohydrates can be assessed by quality as they contain unhealthy components, such as total and added sugars, and healthy components, such as dietary fiber. These differences have the potential to influence or be influenced by the microbiome and should be reflected in its composition. In this case, the unweighted Unifrac is theoretically better since it is sensitive to the small changes in dietary data. Because cardiometabolic measures are influenced by diet, unweighted Unifrac also makes more sense, given types of carbohydrates influence measures such as blood sugar in different ways according to its glycemic index. These differences should all be reflected in the gut microbiome since it is part of the digestion process.

This logic would explain why in the Freshmen Health Study, unweighted UniFrac would detect differences in data and weighted UniFrac did not. In addition, detecting specific species of bacteria that are present or not present is important for future dietary interventions. Currently, there are probiotics with different types of bacteria that aid in strengthening the gut microbiome. By determining distinct differences between groups and providing the type of bacteria that is distinctly associated with certain nutrients or cardiometabolic markers, researchers can make better decisions on the composition of probiotics to guide a healthy gut microbiome and disease prevention.

Chapter 4: Conclusion

In conclusion, both unweighted Unifrac and weighted Unifrac answer different research questions. In the case of diet and cardiometabolic factor, the unweighted Unifrac theoretically makes more sense to use since it describes dissimilarity, not similarity, between two microbial communities. This measure can also detect subtle changes in distinctly different groups, which is important in dietary data where different food groups have different compositions and can influence the gut microbiome even in small amounts. This is due to the sampling of the unweighted UniFrac only taking into account branches that occur in one community and not the other, and assigning these differences a diversity number. The weighted UniFrac includes shared branches and is susceptible to drowning of significance since the denominator includes total abundance and any significance can be driven by a particular species that exists in large numbers. Therefore, while both measures should be analyzed and current practices suggest both measures should be significant in order to confirm “real” significance, the significance of only the unweighted UniFrac gives valuable information for potential dietary interventions.

References

1. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335-336. doi:10.1038/nmeth.f.303.QIIME.
2. Nguyen N, Warnow T, Pop M, White B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. 2016;(March). doi:10.1038/npjbio.
3. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac□: an effective distance metric for microbial community comparison. *ISME J*. 2010;5(2):169-172. doi:10.1038/ismej.2010.133.
4. Fukuyama J, Mcmurdie PJ, Dethlefsen LES, Holmes S. Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Pacific Symp Biocomput*. 2015:213-224.
5. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and Qualitative NL Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities □. *Appl Environ Microbiol*. 2007;73(5):1576-1585. doi:10.1128/AEM.01996-06.
6. Hoffmann C, Dollive S, Grunberg S, et al. Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS One*. 2013;8(6). doi:10.1371/journal.pone.0066019.
7. Claesson J, Jeffery B, Conde S, et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature*. 2012;488(7410):178. doi:10.1038/nature11319.
8. Pilles KN, Van Der Pol WJ, Morrow CD, Asigbee FM, Davis JN. Altered composition of fecal microbiome associated with adiposity and metabolic parameters in Hispanic college students. *Rev ACJN*. 2017.
9. Pilles KN, Van Der Pol WJ, Morrow CD, Asigbee FM, Bray MS, Davis JN. Saturated Fat Intake Correlates With Altered Composition of Fecal Microbiome in Hispanic College Students. *Rev ACJN*. 2017.
10. Chang Q, Luan Y, Sun F. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*. 2011;12(118). https://doi.org/10.1186/1471-2105-12-118.