Masters Project Report

Emily Palmer

Spring 2021

Abstract

Application of a recent method for microbiome association studies to larger datasets provided computational challenges. Several aspects of the method to more general data were developed for use beyond the case study provided in the initial paper.

Contents

1 Introduction

1.1 About microbiome data

One technique for microbiome analysis is sequencing the 16S ribosomal rRNA gene. This gene is a useful marker for

Microbiome data generated from 16S ribosomal RNA (rRna) sequencing present many obsticals to analysis using standard statistical methods. There are several aspects common to microbiome data that violate the assumptions of common statistical tests.

1.1.1 Issues with microbiome data

1. sparcity
2.

1.2 Generalized Estimating Equations

[1]

2 About the GEE method ()

3 Extensions to the method

3.1 ”Missing data”

3.2 Generating the taxa correlation from microbiome data

3.3 How to deal with unlabeled taxa

3.4 Application to American gut data

3.4.1 About the American gut dataset

3.4.2 Filtering the American gut dataset

3.4.3 Computational challenges

4 Conclusion

References

[1] Bo Chen and Wei Xu. “Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures”. In: *PLoS computational biology* 16.9 (2020), e1008108.