Microbiome and Data Integration

Statistical and Logistical Challenges

Mateen Shaikh[1] (PDF) Joseph Beyene[1]

[1]McMaster University

January 24, 2014

1. Overview
   - Data types in the project
2. Microbiomes
   - Common Methodology
   - Proposed Ideas
3. Data Integration
   - CCA
4. Members
   - Microbiome
   - Data Integration (and a little microbiome)

- Several data types:
  - Clinical
  - Dietary
  - Gene Expression
  - Methylation
  - Microbiome
- Most of the data are dependent on each other
- We should expect many of the variables are noise

Overview

- Whole populations of organisms live on and in individuals
- Our focus will be on gut microbiota
- Relationships are expected with diet and with health

- The data will be from sequencing 16S ribosomal RNA from stool samples
- Microecology provides a good set of tools to work with
- Most focus is on the diversity, quantity, and ratio of populations
- Largely exploratory

- Good approximations to the populations are made into OTU tables
- Given an idea of the gut microbiota, we analyze what the population is like
- Several methods to discern relationships between the populations and other variables
- Most methods involve dimension reduction to identify features by eye

- To get better ideas, we may identify subpopulations unique to some individuals
- Combinations of subpopulations can be modelled through finite mixture models
- Lots of literature on inferences using mixture models
- Need to account for overdispersion
  ...maybe using mixed/random effects
- Seek sparse solutions—considerable variable selection

- Canonical Correlation Analyis (CCA) identifies a maximum linear relationship between two data sets
- CCA is not intrinsically designed for the idea of 'response' variables
- We are more interested in many instances where one small set of variables in one data set are related to another small set of variables in another data set, which are also related to another set of small variables in another data set, and so on
- This is far less common in the literature, probably because of complexity

Mike Surette's team (Surette Lab)

- Mike Surette, Jen Stearns, Fiona Whelan

- Focused on microbiome data

Joseph Beyene's team (SIGMA lab)

- Joseph Beyene, Mateen Shaikh
- Additional methodology for microbiome analysis
- Data integration