Inferring Homeostatic Structure in Microbial Systems

Jordan Ferreras, 2020

The microorganisms inhabiting the human body are integral to our health. Understanding how these organisms interact with one another and their environment is believed to hold valuable medical implications. To understand these interactions, researchers study microbiomes: the combined genetic material of the microorganisms in a particular environment. Although major advances in technology allows us to gather tremendous amounts of data, the statistical analyses used to study metagenomic data has not evolved as quickly. The goal of this summer research was to understand the state of the literature and explore challenges associated with analyzing ecological count data using a mixture model. Mixture models are probabilistic models used for data which derives from multiple subpopulations, where there is no explicit way to assign a datum to a subpopulation or class.

Microbiome data is typically represented as a matrix of frequencies where rows represent samples and columns represent species, or taxa. The frequencies are commonly represented as either counts or proportions, where proportions across a sample sum to 1. The data exhibits complexity as samples have different sizes, matrices are sparse (mostly zeros), and the communities we sample from are diverse, leading to a skew towards rare taxa. Lots of data is publicly available on the web thanks to large-scale research projects such as the NIH Human Microbiome Project (HMP) and European Bioinformatics Institute's Tara Oceans Voyage data. Much of the beginning of the summer was spent gathering datasets from sources such as these. To help decide whether or not a dataset was worth lookt.inni1nsM-93(H H)4(uman)]TJETq0.26rxture models are