Microbiome, Metagenomics and High-Dimensional Compositional Data Analysis

Next-generation sequencing technologies allow 16S ribosomal RNA gene surveys or whole metagenome shotgun sequencing in order to characterize taxonomic and functional compositions of gut microbiomes. The outputs from such studies are short sequence reads derived from a mixture of genomes of different species in a given microbial community. We first present a brief overview of the statistical methods we used for 16S rRNA data analysis. We then introduce a multi-sample model-based method to quantify the bacterial compositions based on shotgun metagenomics data using species-specific marker genes. The resulting data are high-dimensional compositional data, which complicate many of the downstream analyses. We introduce the GLMs with linear constraint on regression parameters in order to identify the bacterial taxa that are associated clinical outcomes and a composition-adjusted thresholding procedure to estimate correlation network from compositional data. We demonstrate the methods using two on-going gut microbiome studies at the University of Pennsylvania.