Graphical representation of microbial community subpopulations using penalized Kendall's distance

Deepak Nag Ayyala, Shili Lin

Department of Statistics, The Ohio State Univerity, Columbus, OH

When metagenomic count data are recorded by classifying gene sequences, there are several methods to visualize the data to uncover hidden community substructures. One such method is to construct a multidimensional scaling model using a measure of dissimilarity between observations. UniFrac is one such measure of distance that is very commonly used for metagenomic data. In this work, we construct a multidimensional scaling model to represent the data on a 3-dimensional coordinate system based on a novel penalized Kendall's τ-distance to characterize dissimilarity between observations. We applied the proposed procedure to a human microbial community dataset composed of over 800 observations from multiple habitats and body sites. The constructed scaling model exhibits several features in the data set that are not seen in a UniFrac-based model. Clustering based on the model reveals several physiological similarities between the observations within each of the clusters.

Keywords: Metagenomics, Multidimensional scaling, Kendall's distance.