Network testing for microbiome data

Julia Fukuyama and Susan Holmes, Department of Statistics, Stanford University

Overview

Network testing

- Introduced by Friedman and Rafsky as a multivariate generalization of the Wald-Wolfowitz runs test.
- Idea: make a graph on all the points, and count the number of edges between samples of the same group. Compute a p-value by permutation.
- Can use any distance, and any graph-building method.

Properties

- Non-parametric
- Sensitive to local structure
- Robust to data corruption

Implementation

- Github package phyloseqGraphTest performs these tests for phyloseq objects
- Can use any distance from phyloseq
- Multiple network-building methods and visualizations

Power simulations

- For both distances, k-nearest neighbors graph and the minimum spanning tree have the best power
- The distance thresholded graph performs worst, possibly because it has unconnected nodes
- Different distances perform similarly
- Recommend k-nn graph or MST

Jaccard

Bray–Curtis
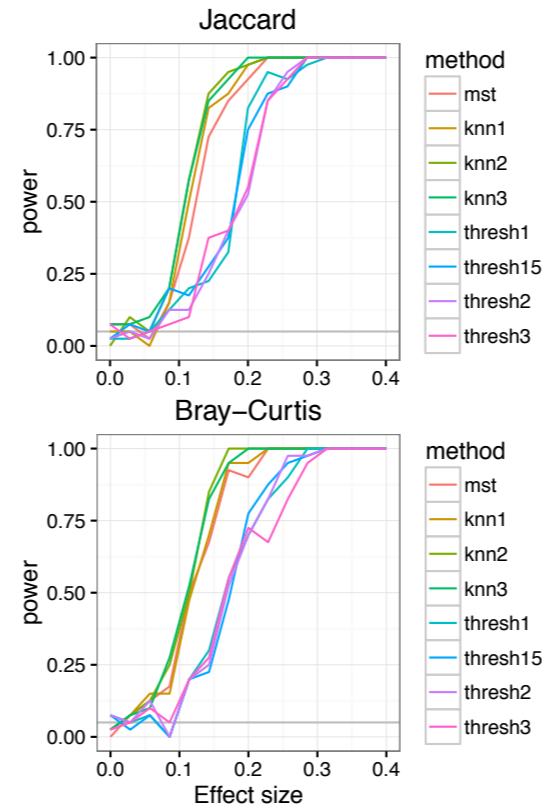
Network tests more powerful than adonis

Jaccard
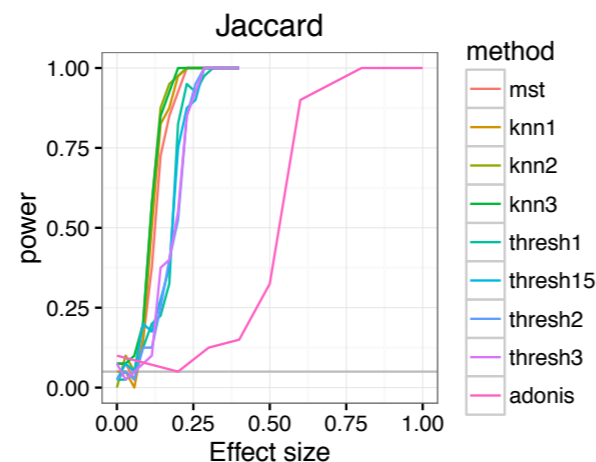
- Power of adonis (pink line) vs. some network tests
- Adonis has substantially less power in this simulation
- Possible reason: Network tests using more local information

Example network

- Network created is the minimum spanning tree
- Nodes colored by M/F, what we are testing
- Many more "pure" edges than expected by chance, so reject null
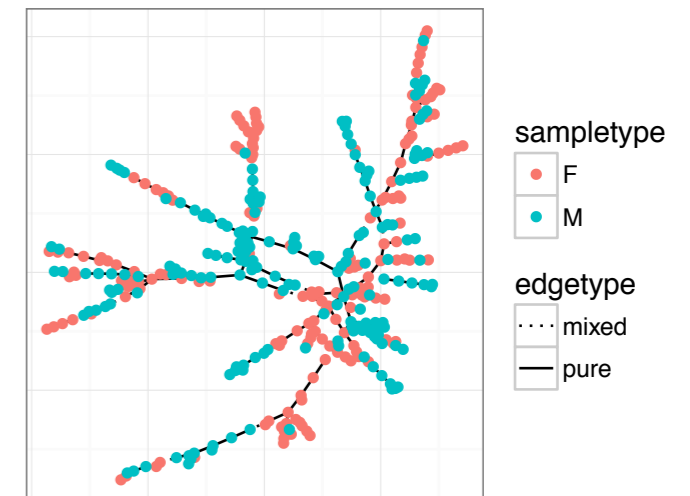
References

Friedman, Jerome H., and Lawrence C. Rafsky. "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests." The Annals of Statistics (1979): 697-717.

Schilling, Mark F. "Multivariate two-sample tests based on nearest neighbors." Journal of the American Statistical Association 81.395 (1986): 799-806.