Inferring microbiome networks using graphical models

Jared Samilow, Shibu Yooseph
jared_samilow@brown.edu , shibu.yooseph@ucf.edu

Introduction/Abstract

Microbes account for more than half of cells in the human body, but biologists have only just begun to catalogue them and explore their impact on health and diseases. The goal of this project is to assess the correlations between microbes—when the presence of one species predicts the presence of another. Our data come from microbiome samples from different parts of the human body. The samples were sequenced on their 16S ribosomal marker and microbiologists recorded how many times a particular species (a taxon) was observed. This leads to an $n \times p$ matrix M of n samples of p taxa which, as the title suggests, is modeled as a graph G, with the taxa the nodes of G, and the relationships between them the undirected edges.

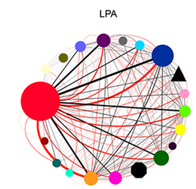


Fig. 1 — A mathematical representation of a microbe network.[10]

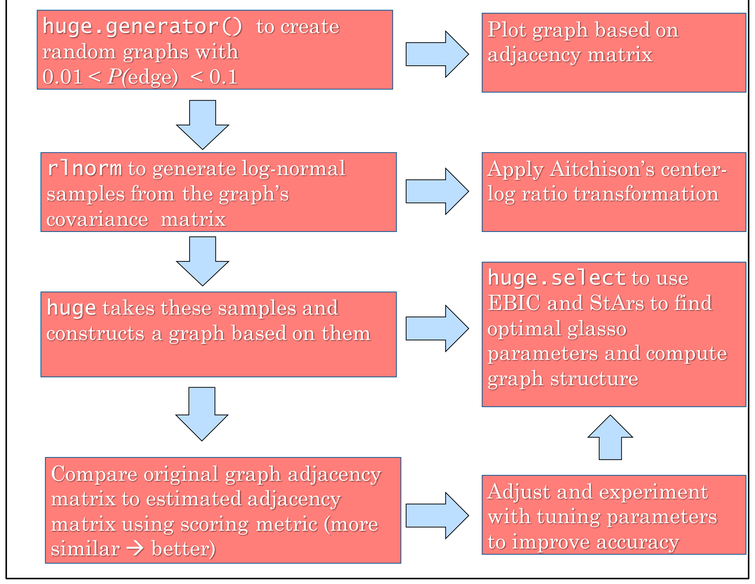


Fig. 2 — A hand-colored electron micrograph of a microbe cluster from the human gut (Martin Oeggerli)

We work primarily in R, using a Graphical Gaussian Model (GGM) to estimate the inverse covariance matrix of G, along with the Extended Bayesian Information Criterion (EBIC) and StArs models to train the estimation techniques employed by the GGM.

Method

How we build simulated data for microbe networks:



Method (cont'd)

We use both real-world data sets from the National Institutes of Health *and* generate simulated data as a basis for comparison

- Matrices contains compositional data, so columns aren't independent
  - Use Aitchison's centered log-ratio transformation to eliminate bias
- Use R packages `base`, `glasso`, `huge`, `compositions` and `rlnorm` to simulate and estimate the data.
- Once technique works on simulated data, can apply it to real biological data.

Results

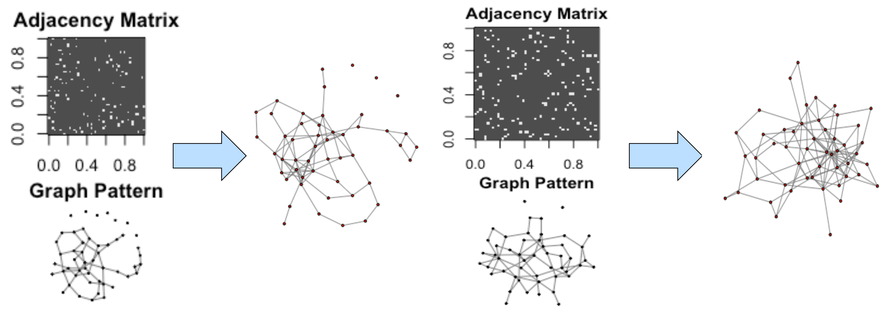


Fig. 3 — On the left is the "true" graph we generated. On the right is the approximation given by our methodology using EBIC. The approximation is good, but not yet perfect.

Fig. 4 — Same as Fig. 3, but for a different simulated graph.

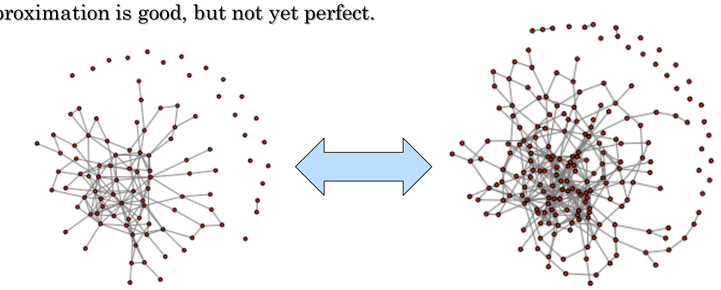


Fig. 5 (above) — Graphs we generated from real-world NIH data-set with $n = 39$ using StArs. The graph on the left represents the 100 most frequent taxa; the one on the right the 200 most frequent taxa.

While the graphs look complex and connected, both are quite sparse. The graph on the left could have a maximum 4,950 edges; the one on the right a maximum of 19,900 edges!

Fig 6.(**right**)—Accuracy scores for an average of 11 random graphs.

% correct edges	89.1
% false positives	10.5
% false negatives	0.4
# real edges – # estimated edges	2.0

Future Plans

So far, our simulated graphs have all been random (i.e. a fixed probability of an edge). But there are also "band" and "cluster" graphs that should be simulated and explored. Additionally, we want to see how our approximation for the real-world data-set compares with the estimates given by SparCC and SPIEC-EASI, two other methods for inferring microbiome networks described by 16S ribosomal data.

There is potential for minor improvement through trial-and-error. For example, instead of using a centered-log ratio transformation, we could instead use an additive or isometric log-ration transformation.

References

1. The Human Microbiome Project Consortium, *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-221.
2. The Human Microbiome Project Consortium, *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207-14.
3. Jumpstart Consortium Human Microbiome Project Data Generation Working, G., *Evaluation of 16S rDNA-based community profiling for human microbiome research.* PLoS One, 2012. **7**(6): p. e39315.
4. Layeghifard, M., D.M. Hwang, and D.S. Guttman, *Disentangling Interactions in the Microbiome: A Network Perspective.* Trends Microbiol, 2017. **25**(3): p. 217-228.
5. Bühlmann, P., M. Kalisch, and L. Meier, *High-Dimensional Statistics with a View Toward Applications in Biology*. Annual Review of Statistics and Its Application, 2014. **1**(1): p. 255-278.
6. Friedman, J., T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics, 2008. **9**(3): p. 432-41.
7. R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2008.
8. Friedman J, Alm EJ (2012) Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol 8(9): e1002687.
9. Milici, et al., *Co-occurrence Analysis of Microbial Taxa in the Atlantic Ocean Reveals High Connectivity in the Free-Living Bacterioplankton.* Frontiers in Microbiology, 2014. http://journal.frontiersin.org/article/10.3389/fmicb.2016.00649/full
10. National Institutes of Health, *NIH Human Microbiome Project defines normal bacterial makeup of the body,* 2012. https://www.nih.gov/news-events/news-releases/nih-human-microbiome-project-defines-normal-bacterial-makeup-body
11. Zhao, T. Li, X., et al., *High-Dimensional Undirected Graph Estimation*, 2015. https://cran.r-project.org/web/packages/huge/huge.pdf
12. Aitchison, J., *The Statistical Analysis of Compositional Data.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 44, no. 2, 1982, pp. 139–177.

Acknowledgements & Disclosures

No disclosures. The authors thank the National Science Foundation (grant no. 1461121) for supporting this REU program at the Center for Research in Computer Vision.