Computational Analysis of Microbial Community Datasets

Shree Madan

Abstract

Microbial communities are an important cornerstone of all ecosystems. By studying their interactions and the way they respond to environmental changes we can better understand which factors contribute to the ability of the community to function. Previous experiments were conducted to gather data on how microbial community composition changes over time of a nutrient perturbation. After collecting this data, the microbial community composition was determined by sequencing of the 16SsrDNA gene. Being able to analyze these datasets of community composition, is where we are able to gain the most understanding of the significance of the community. There are main ways to look at these communities but the main three are the statistical approach, network analysis, and linear systems modeling. Each of these provides a different lens to look at datasets which can be used to give context to the various trends in the data. Being able to understand how to use each of these methods and determine which is the most useful for any given dataset is an important part of this research.

Introduction

Natural microbial communities, such as those from soil and aquatic environments, as well as the microbiome associated with humans, are complex systems to use to answer questions about microbial ecology (Bao 19). We are working with phototrophic community enrichments that degrade cellulose and fix nitrogen that was enriched from marine and freshwater environments as a model system. This system can provide insights into how renewable resources such as cellulose, N2, and light can be efficiently converted into cell biomass and other products. We have previously performed nutrient perturbation experiments on the community enrichments to also understand how microbial communities respond.

Network analysis enables hypotheses about the way and kinds of relationships that can be formed as communities are given different nutrient sources. For example, by employing network analyses to a large soil microbial dataset generated by pyrosequencing, such as in a paper called “Molecular ecological network analyses” by Deng et al, the process of exploring the complex set of data is more feasible and interesting than unseen patterns emerged, including non-random association, deterministic processes at different taxonomic levels and unexpected relationships between community members. This analysis gives insight into community dynamics and relationships which can help to predict the effects of climate change on ecosystem functioning or how the gut microbiome can be altered to improve human health.

Creating a multi-faceted approach to community analysis allows for a better understanding of these computational methods from a mathematical as well as biological perspective. Advancing this pipeline enables for better implementation of these skills. A future

aspect of this work would be being able to apply these methods to other microbial communities to create a predictive model for these interactions and the tenacity of microbial communities to survive environmental changes and help answer questions about how the composition of microbial communities relates to how they function.

Methods

Nutrient Perturbation Experiments (Previous Work)

The nutrient perturbation experiments consisted of shifting a freshwater phototrophic community that was growing on cellulose as a sole carbon source to media that contained either glucose or malate as the sole carbon source. Replicates of communities maintained on cellulose as a sole carbon source served as a control. Carbon source perturbations were performed by transferring communities in triplicate at a 5%v/v ratio to a fresh medium of the same composition with either 10mM Glucose, 10mM Malate, or amorphous crystalline cellulose as the sole carbon source. It was transferred twice to cellulose, with two samples taken on the 3rd, 5th, 7th, 10th, and 20th days after transfer. For the glucose and malate perturbations, the first transfer was to the perturbation medium, and the second and third transfers were to the cellulose medium. Samples from three biological replicates were taken on the 3rd, 5th, 7th, and 10th days after transfer. 1ml samples of communities were removed over time and each culture was sampled once.

Community genomic DNA was extracted and the v4 region of the 16SrRNA gene was amplified at MRDNA. Once these samples were sequenced we received data on the community composition and number of operational taxonomic units (OTU). Each OTU represented an individual species. Questions that can be asked about this data from these tractable communities may provide insight into long-standing questions in microbial ecology: What factors influence community composition? How do microbial communities recover from perturbation? This summer I continued the analysis phase for this project learning methods to gain insight into these motivating questions.

Initial Set-Up

There were a few basic aspects of the pipeline that had been put together but a more thorough understanding of different methods was required to make further progress on the computation side. Starting with a large and intensive literature search allowed for more calculation techniques to be researched and gave many more options to choose from when it came to compiling it all. As the pipeline was developed, this background research also gave more perspective on the choices being made such as choosing which metrics and visualization methods were the most useful for later biological analysis. This is one of the many instances where the computational methods needed to be adjusted to maximize the significance of the finals results.

Learning to program in Python was a huge aspect of this research. Most of the pre-existing pipeline was put together using different Python packages so it made the most sense to continue to use Python rather than switch to another programming language such as R. Python made the most sense for the programming for a few different reasons. Python already has a powerful network analysis package called "Networkx". This package has most of the tools needed for a network analysis program built in so rather than building the network up from scratch, it's possible to manipulate various aspects of the network with relative simplicity. This ability meant that many different arrangements could be tested quickly to see which provided the most useful results. Additionally, Python had pre-existing packages for other computational methods such as "scikit-bio" for testing things like alpha and beta diversity was made more straightforward. The initial literature search made it clear the best approach was doing different statistical analyses, combining data to draw conclusions, and then visualizing it in a way that some hypotheses could be drawn about the biological relationships. This is also the aspect of network analysis that could exist as a predictive measure for other communities and their ability to respond or adapt to different nutrient perturbations.

Community Composition analysis during nutrient perturbation

The beginning layer of analysis was about community composition. Although these are not directly related to network analysis, they provide an important foundation for common analysis that characterizes microbial communities. Starting with the absolute abundances of each OTU in every sample and converting those into relative percentages in each community helps compare the way samples changed over time. The nutrient perturbation experiment was performed in triplicate and this enabled comparison of the triplicate samples against one another. The initial step was to graph every individual sample and pick out any samples which stood out within the triplicates. This was one way of checking whether the biological replicates were consistent because there were certain samples where the OTUs were unusually different from their sister samples. These are often samples that have been incorrectly sequenced by the company and hence need to be removed from the dataset to ensure accurate data. Taking the average of all the triplicate values to determine average percentages for each OTU and then graphing them over time gave context to the overall story of each OTU over the time course data. There were a few other ways this data was graphed so that it could be considered from different angles including graphing the differences from the original community during the same time.

Alpha Diversity Analysis

The second form of analysis here is alpha diversity metrics. Alpha diversity statistics are one way to calculate the species diversity of different community samples and come in the forms of many different indices. A basic literature search informed the decision to use the following indices: ACE, CHAO1, Shannon, and Simpson, as they seemed to be the most useful. Using the

sci-kit bio package to calculate these values also streamlined the process because it allowed the existing data to be simply plugged in and automatically gave back the values needed.

Beta Diversity

Beta diversity was another statistical method to understand the difference between samples. A cut-off was defined where OTUs with less than 5 counts (approximately 0.005% of the community) across all samples were left out of the analysis as these represented an extremely small percentage of the community composition and may not be significant. Bray-Curtis dissimilarity, calculated by dividing shared abundance by total abundance, is a measure of the distance between populations. A distance matrix was created by calculating the distance between all pairs of the samples. The technique used for visual exploratory analysis was Principal Coordinate Analysis (PCoA) was used for visual exploratory analysis. PCoA was performed on a subsection of 62 samples that were growing on cellulose, giving axes that explained 34.21% and 13.13% variance. The main axes of variation picked up on the pattern of stationary phase effects seen in later transfers.

Network Analysis

Network analysis was the final piece of analysis and focused on visualizing the relationships between different OTUs. We closely followed the methods of the Deng et al paper to explore the correlation patterns and dependencies of OTUs. The 86 samples were split into 8 groups, corresponding to the transfer and carbon source series. Each group had 10-12 samples, with the exception of the second cellulose transfer group which only had 6 samples. For correlation analysis, the relative abundance counts for each OTU were normalized by subtracting the mean across all samples in the group and dividing by variance across all samples in the group.

For each group, a similarity matrix was constructed by taking the absolute value of the Person correlation between each pair of samples. A correlation threshold to separate random fluctuations from meaningful signals was found using Random Matrix Theory. To form the adjacency matrix, values greater than the correlation threshold were set to 1, else they were set to 0. The Python module empiricalRMT was used to compare the spacing of unfolded eigenvalues of the adjacency matrix to a Gaussian Orthogonal distribution (characteristic of random noise) and Poisson distribution (characteristic of signal). Networks were constructed from the adjacency matrix formed with a correlation threshold of 0.76. In these networks, each node represented an OTU, each edge represented a correlation >= 0.76 between a pair of nodes. The greedy modularity algorithm was used to group nodes into modules while maximizing the network's modularity score. A module represents a set of nodes with a shared correlation pattern.

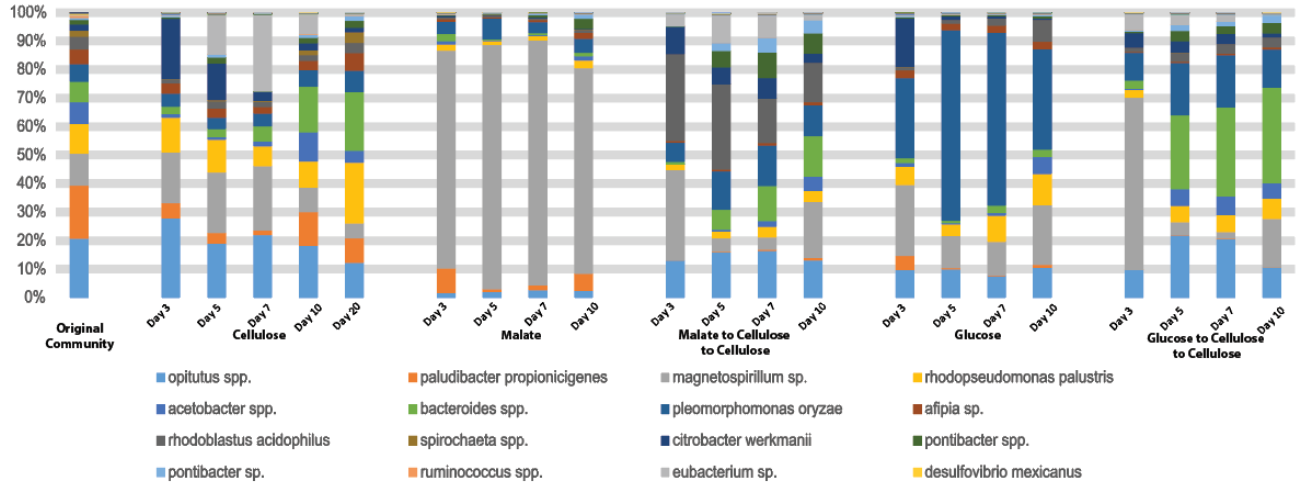
To analyze and compare the differences in the networks, a set of topological indexes were calculated for the overall network structure and for node-level attributes. Node-level attributes summarize a node's connection with other nodes, its module, and give a sense of its relative role and impact within the community. The network was visualized with a fruchterman reingold layout. The correlations between primary eigenvectors of each module (eigengenes) were calculated and visualized to get a sense of broad correlation patterns and the hierarchy of modules and samples.

Based on all the different analysis values, a set of biologically significant members were identified and their relationships were analyzed in more detail.

Results

Changes in microbial community composition in response to nutrient perturbation

When the communities were shifted to a new carbon source, we see new species start to take over the community because they are able to better degrade the new carbon source (Fig.1). However, when we shift back to cellulose, which was the original carbon source, we see that the community actually starts to return to the original composition and networks (Fig. 1).



The return to the original community composition following the perturbation happened gradually. PCOA analysis showed that communities became more similar to the original transfer over time after returning to the original cellulose nutrient condition (Fig. 3).

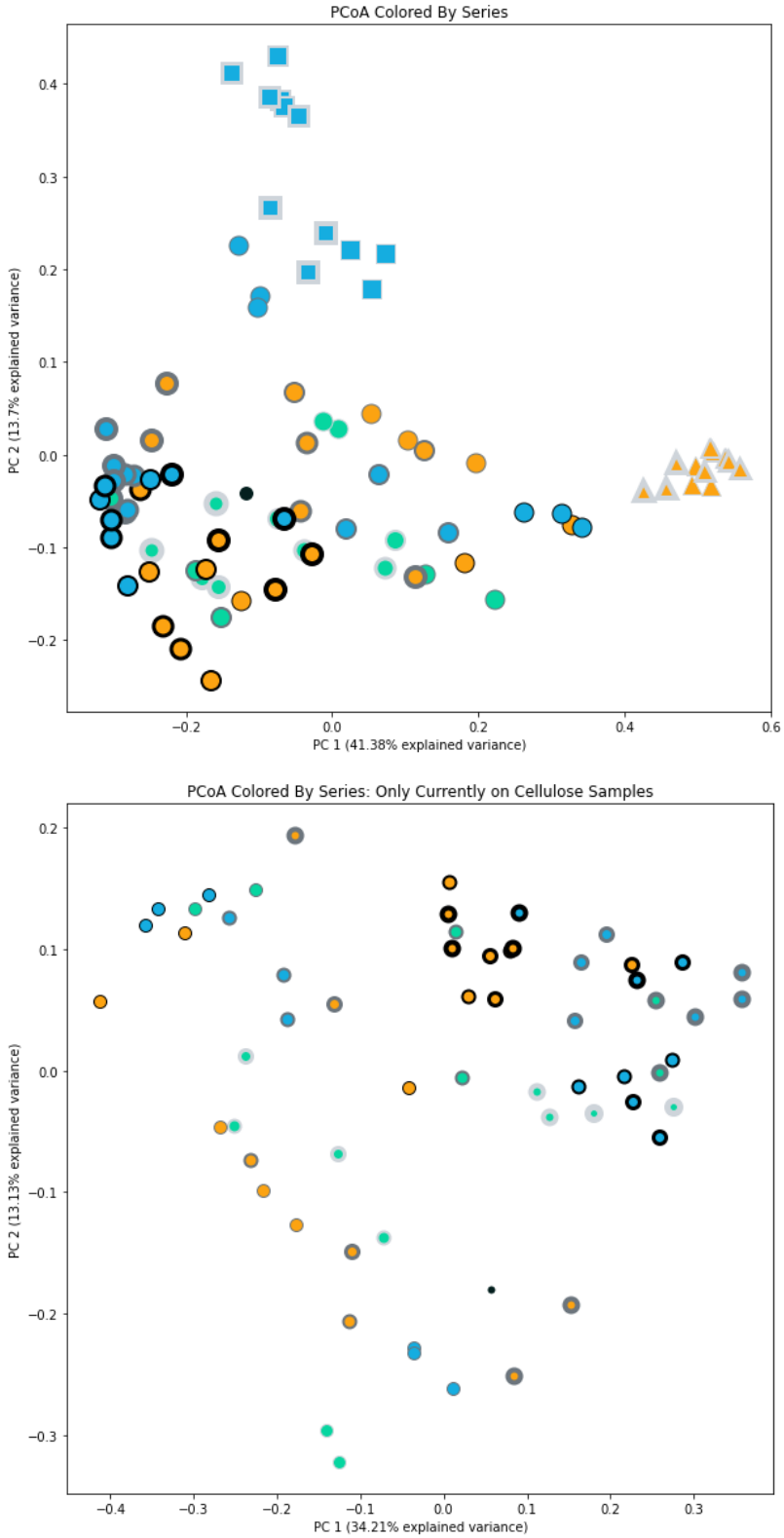


Figure 3: PCoA analysis of microbial communities undergoing nutrient perturbation. orange: malate grown current/previous, green: cellulose grown, blue: glucose grown current/previous Triangles: malate as a current C-source, Squares: glucose as a current C-source,

Circles: cellulose as a current C-source, border thickness: increasing days of culture growth 3-20 days

Examination of potential network interactions between community members

The communities were able to return to cellulose degradation following nutrient perturbations, but not all the networks remained the same. Community members showed a variety of dynamics during nutrient perturbations. Conclusions about three members of the community were drawn by putting together various statistical metrics and using those to understand the larger picture of their potential roles in the community. One example is *Eubacterium* sp., which seems important in the cellulose communities as it has many connections and high eigenvector centrality. These values decrease when transferred to glucose and malate but return to higher values when the community is transferred back to cellulose. This supports the idea that Eubacterium sp. is an important bacteria for the cellulose-degrading communities. The species is a potential cellulose degrader based on homology (Flint, 2012). By being able to understand the different interactions this bacterium had in the community over the different parts of the time course gave insight into its overall role in the community.

***Magnetospirillum* sp.**	1C		1M		2M		3M	
Eigenvector centrality	0.00124306731190604 (66th)		0.0196069651278316 (19th)		0.0161739172862933 (41st)		1.48804540653487E-17 (93r)	
Module/Degrees	1	7	2	7	0	4	4	1
Who it's connected to	paludibacter propionicigenes, bacteroides spp., pleomorphomonas oryzae, pontibacter sp., inquilinus spp., pseudochrobactrum kiredjianiae, steroidobacter spp.		rhodopseudomonas palustris, acetobacter spp., afipia sp., rhodoblastus acidophilus, spirochaeta spp., rhodocyclus tenuis, rhodoplanes elegans		pontibacter spp., pontibacter sp., pleomorphomonas sp., fibrobacter spp.		escherichia spp.	

***Pleomorphomonas* oryzae**	1C		1G		2G		3G	
Eigenvector centrality	0.00456514736685744 (47th)		6.56073829432077E-15 (65th)		8.47118344404463E-06 (98th)		0.0488889484169298 (38th)	
Module/Degrees	1	9	7	1	2	4	0	5
Who it's connected to	magnetospirillum sp., bacteroides spp., afipia sp., spirochaeta spp., pleomorphomonas sp., inquilinus spp., pseudomonas veronii, streptococcus salivarius		magnetospirillum sp.		ancalomicrobium spp., acinetobacter junii, pseudomonas veronii, streptococcus sanguinis		magnetospirillum sp., rhodocyclus tenuis, pleomorphomonas spp., pleomorphomonas sp., bradyrhizobium sp.	

Eubacterium sp.	1C	1G	2G	3G	1M	2M	3M
Eigenvector centrality	0.262407652385771 (3rd)	0.15857240892855 (16th)	0.285486898154699 (3rd)	0.286616026707805 (4th)	N/A	0.000019524595317839 (74th)	0.29972747344208 (2nd)
Module/Degrees	0 17	0 9	0 15	3 Degrees 13	N/A N/A	Module 1 Degrees 23	Module 0 Degrees 14
Who it's connected to	caulobacter spp., phaeospirillum sp., rhodopseudomonas pangongensis, opitutus sp., ochrobactrum anthropi, nitrospirillum azospirillum amazonense, brevundimonas sp., escherichia vulneris, brevundimonas spp., rhizobium straminoryzae, rhodothermus spp., rhizobium sp., enterobacter sp., gemmatimonas spp., oligotropha carboxidovorans, pseudomonas sp., tumebacillus sp.	paludibacter propionicigenes, spirochaeta spp., citrobacter werkmanii, ruminococcus spp., desulfovibrio mexicanus, clostridium sp., citrobacter spp., acinetobacter junii, rhizobium straminoryzae	rhodopseudomonas palustris, caulobacter spp., phaeospirillum sp., rhodopseudomonas pangongensis, opitutus sp., ochrobactrum anthropi, nitrospirillum azospirillum amazonense, brevundimonas sp., escherichia vulneris, cytophaga spp., rhizobium straminoryzae, rhodothermus spp., rhodovibrio sodomensis, chelatococcus spp., methyloligella halotolerans	rhodopseudomonas palustris, bacteroides spp., citrobacter werkmanii, pontibacter sp., rhodoplanes elegans, caulobacter spp., barnesiella viscericola, citrobacter spp., fibrobacter spp., ochrobactrum anthropi, nitrospirillum azospirillum amazonense, brevundimonas sp., bacteroides fragilis	N/A	caulobacter spp., phaeospirillum sp., rhodopseudomonas pangongensis, ochrobactrum anthropi, nitrospirillum azospirillum amazonense, brevundimonas sp., rhizobium petrolearium, escherichia vulneris, brevundimonas spp., methylocystis heyerii, cytophaga spp., rhizobium straminoryzae, rhodothermus spp., rhodovibrio sodomensis, enterobacter sp., gemmatimonas spp., chelatococcus spp., afipia broomeae, thioalkalivibrio spp., methyloligella halotolerans, spirosoma spp., pseudomonas sp., tumebacillus sp.	caulobacter spp., rhodopseudomonas pangongensis, opitutus sp., ochrobactrum anthropi, nitrospirillum azospirillum amazonense, brevundimonas sp., escherichia vulneris, brevundimonas spp., cytophaga spp., rhizobium straminoryzae, rhodothermus spp., rhodovibrio sodomensis, afipia broomeae, methylocella silvestris

Table 1: Community members that changed in abundance and their network dynamics over time of perturbation

Discussion and Future Work

One key takeaway from this work is that there are many methods for computational analyses for describing microbial communities. Even in just the network analysis aspect, there were many methods that could have been following but following the Deng paper was the most beneficial because their dataset was most similar to ours. Beyond network analysis, looking at alpha and beta diversity as well as community composition was necessary to create a complete picture of the community to understand the biological connections. With the right combination of analysis, connections between biological relationships and the corresponding statistics were clear-cut and distinct.

This research has many directions it could go in. One way is to identify a new community to further evaluate the efficacy of our pipeline. Another community has been established that could be used to do further testing on the pipeline. This community is a novel plastic degrading community that is both anaerobic and phototropic and is tractable. This combination of characteristics makes it an excellent candidate for further study. By going through the time course with different nutrient perturbations (as done in the previous freshwater community), the resulting dataset could be analyzed using similar methods. Determining whether similar patterns will occur in different communities could aid in furthering the understanding of how microbial communities respond to perturbation.

Acknowledgments

I'd like to acknowledge the funding I received through the Clare Boothe Luce Research Scholars Program at Olin College which is funded by the Clare Boothe Luce Program of the Henry Luce Foundation. Additionally, I'd like to thank my research advisor, Jean Huang, for all the mentorship and guidance she has given me this summer. I'd also like to acknowledge the other researchers who have worked on this project, both past and present, who have not only

created this opportunity by doing the initial research but also have supported me throughout this project.

References

Bao, Y., Dolfing, J., Wang, B., Chen, R., Huang, M., Li, Z., ... & Feng, Y. (2019). Bacterial communities involved directly or indirectly in the anaerobic degradation of cellulose. Biology and Fertility of Soils, 55(3), 201-211.

Deng, Y., Jiang, YH., Yang, Y. et al. Molecular ecological network analyses. BMC Bioinformatics 13, 113 (2012). https://doi.org/10.1186/1471-2105-13-113

Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., & Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. Gut microbes, 3(4), 289-306.