**Pre 8/19/14 Meeting work**

Network analysis continued:

Now that I have an association network for soil genera from the

Wilson dataset using ccrepe and have attempted to visualize this network

with gephi, next objective is to apply basic network analysis methods to this network.

To do’s in bold.

Presentation:

1. slide going backwards: What is network analysis good for and what general questions can be answerered – shown below.
2. Specific points bolded below made into graphs/tables.
3. References to T Zhang, Wilson, and huttenhower

Python network analysis tool/data format?

Some general questions that I hope to answer:

What is the overall network structure? Does the connectivity follow a power-law?

Is this network a "small-world" network?

**data requirements: Degree distribution, mean path length, diameter, clustering coefficient, modularity index**

Are there "hubs" within this network? Do hubs behave differently than non-hubs?

**Data requirements: nodes sorted by degree, some cutoff criteria. Then identification of within and between module hubs. Degree vs abundance chart.**

Do phylogenetically more related groups interact differently than random groups?

**Data requirements: inter vs intra clade positive and negative association graph**

"Specialists" vs generalists: are organisms that are highly abundant in one ecosystem but not present in others

different (statistically, over all node/edges) than organisms that exist in most environments?"""

Association network properties:

Node connectivity in association network from Wilson paper follows a power law distribution: Proability of OTU with k connections ~ kY, Y =-.6. “Small-world” networks often have Y between 2-3.5, these types of networks have “hubs”, some nodes with much higher connectivity that act as centers for preferential network attachment.

Because the power law coefficient is low, this association network is more even or random than typical small world networks.

Next up: Modules/clusters within this network.

**Tuesday, August 19, 2014**

Hub identity: Given that some nodes are more connected (hubs) than others in this network, are there any commonalities between hubs? Are they generalists/specialists/etc?

Goals:

dataframe appended with specialist /abundance data

plots of degree vs abundance, degree vs specialist/generalist

results added to update docx.

Code commented and doc’ed somewhere

run home

work in 30 minute chunks

done:

responded to Mary Brand, signed up for data analysis class.

Thursday:

**Added degree vs abundance, degree vs # of occurrence.**

Inter va intra class associations? How to handle this:

I have an edge list with “k;p;c…” naming convention.

Filter this based on rules (positive/negative R>.9, q<.05)

For inter vs. intra class:

Function that takes an edge dataframe (or row?) and returns all unique classes + counts?

Function that takes an edge dataframe

**Monday, August 25, 2014**

Sampling stuff for tomorrow:

Bring sampling kit + completed forms

9:30 AM Kirie: Ask for James Kaminski + John Smoody

address: 3, contact #:

Thursday, September 4, 2014

Story up to now:

Used ccrepe on soil microbiome study

* In original paper, they used unifrac (define?) to demonstrate consistent differences between endo/rhizo/bulk soil microbiomes. They used RDA/ANOVA to show statistically significant differences between communities from different soil sites/strains.
* My approach was to use co-occurrence networks to identify potential interactions in soil dataset.

Procedure:  
Generate co-occurrence network in R-CCREPE implementation – point is to avoid compositional effects. Compositional effects strongest in environments with low alpha diversity. (I think one thing I’m missing in my work to date is a control – what does the network look like if I use Pearson and Pearson with shuffling?)

Analyze network generated by this procedure? Small-world etc.

Found power-scaling of connectivity

Find hubs – determine if hubs are overrepresented by some biological group/explanation

**Friday, September 5, 2014**

Wednesday, September 10

Things that are useful: argparse: add arg

Recursive factorial definition:

N = 5

REC(n) = n\*(rec(n-1))

**Monday, September 22, 2014**

To do:

Get code for sequence distance based methods working and make a flowchart type thing for 16s sequence analysis methods. (include results where applicable)

**Done:**

Figure out exactly what is needed for sending samples (including PO form) and compile for tomorrow

Finish sampling reports

Dr at 11/ **meeting/class -> tomorrow 12/1**

Class tomorrow (reading DLed, 1PM Tech F-395 (opposite A wing))

Neutral vs niche work:

PNAS 2010 Chrisholm: niche and neutral models predict asymptotically identical species abundance distribution curves at high diversity

Neutral model: zero-sum multinomial distribution at low diversity, log-series distribution at high diversity

Sampling plan:

1. Sequence current samples and perform preliminary screening

1a. Using unweighted and weighted Unifrac distances, calculate distance matrix for 6 samples. Are Egan samples close together?

2. Identify sequencing strategy going forward:

Goals/questions?:

Do niche or neutral models fit species abundance distribution?

Research question:

Can we quantify the relative impact of stochastic vs deterministic processes on community assembly by comparing communities with:

Egan: identical inoculum but different treatment regimes

Obrien: identical influent conditions AND treatment regimes

1. A. Preliminary check: compare Stickney A and B samples to Egan Taper/Spial samples. Compare Unifrac distances between Stickney A/B and Egan T/S samples and between Egan and Stickney replicates to gauge sampling biases/noise.

Data: 3 replicates from each set that has them: 3 tanks \* 3 replicates +1 basin \*1 replicate = 10 samples.

B. Fit neutral model parameters to a single WWTP community and see if these parameters can generate species abundance distributions for other plants. How will this depend on things like SRT etc?

1. Longitudinal study: Sampling from Egan and O’Brien (+ others?) on a weekly basis or biweekly basis will provide data for at least two projects.

2A. Unifrac distances between samples will be evaluated to gauge mean and variance of unifrac distancse for identical and non-identical operating conditions. If different aeration has no effect on unifrac distance, then community assembly would be consistent with neutral model.

Phylogenetic test of neutral model from Jeraldo et al. (PNAS 2010) or similar method could be used as a check for neutral vs niche assembly.

2B. Network analysis for community structure: weekly samples from 4 plants could be used to construct association networks. This goes back to **Huttenhower** paper. Network nodes can be same site or different site. Can identify whether same or different site have more edges/predictive ability. Same site different aeration tank same idea.

Data requirements? # sites, # samples, frequency?

2C. Linear models could be built to identify if OTUs explain variation in process conditions -> (Werner 2011 PNAS)

Used 1 year study, monthly samples, many plants

Can also summarize summer work -> start with Alex presentation and add newer stuff -> check phylogenetic difference vs association/inter vs intraclass connections. Show the network stuff and basic network analysis questions

Formalize this:

Beta diversity estimates