Uncovering metabolic cooperation from partial correlation graphs in Neo4j

CSE 544 Database Project

Janet Matsen (Chemical Engineering)

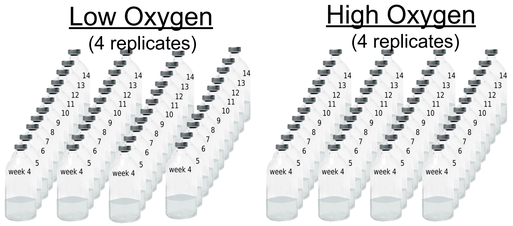
Background

How do bacteria cooperate to remediate methane? Which organisms tend to pair together in nature, and what metabolic roles do each play? Given enough samples of complex communities working together, it may be possible to infer such interactions from the partial correlation matrix of gene expression across organisms and samples.

Sediment was collected from Lake Washington using a sediment corer (Fig. 1a). The samples were then cultured in the lab with methane, and either high or low oxygen. Metagenomes were sampled for weeks 4-14. Sequences were mapped to isolate genomes, normalized, and the partial correlations for expression for every gene was computed.

Figure 1. Experimental setup & equipment

(a) Sediment corer used to gather the Lake Washington samples



(b) Diagram of the experimental setup: 8 independent sediment samples were collected from Lake Washington. 4 series were cultured in low oxygen conditions, and 4 in high oxygen conditions. Samples were serially transferred weekly, and metagenome RNA-seq data was collected for weeks 4-14 for each sample.

Methods

Graphs are a useful tool for describing how species interact [1]. Partial correlations are often used to determine the connections in a graph in which the edges represent influence of one node on another [2]. Once a graph is in hand, we can find community structure: groups of nodes where the nodes within a group are highly connected to each other but there are fewer connections between groups [3].

Graphs using different partial correlation cutoffs were generated using Cypher, Neo4j's query language. I wrote a small Java program that can be invoked from the command line to build graphs using different partial correlation cutoffs and determine connected components as computed by Neo4jSNA [4]. Python was used to analyze results.

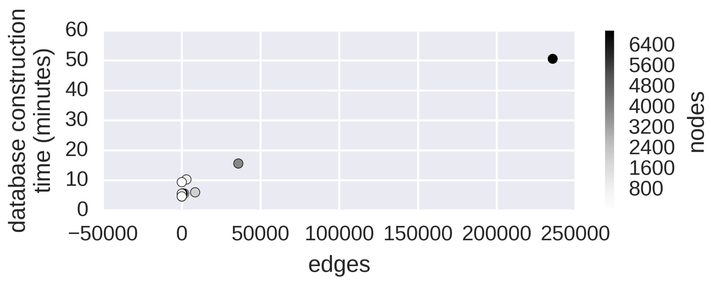
Cypher query for building databases

```
LOAD CSV WITH HEADERS FROM
    'file:%s'
AS line FIELDTERMINATOR '\t'
WITH toFloat(line.pcor) AS pcor,
     abs(toFloat(line.pcor)) AS abs_pcor,
     line.cross_species AS cross_species,
     line.association AS association,
     line.source_locus_tag AS source_locus_tag,
     line.target_locus_tag AS target_locus_tag,
     line.source_organism_name AS source_organism_name,
     line.target_organism_name AS target_organism_name,
     line.source_gene AS source_gene,
     line.target_gene AS target_gene,
     line.source_gene_product AS source_gene_product,
     line.target_gene_product AS target_gene_product
WHERE abs_pcor > %f
MERGE (g1:Gene {locus_tag:source_locus_tag,
                organism:source_organism_name,
                gene:source_gene,
                gene_product:source_gene_product})
MERGE (g2:Gene {locus_tag:target_locus_tag,
                organism:target_organism_name,
                gene:target_gene,
                gene_product:target_gene_product})
MERGE (g1) -[:X {pcor:pcor,
                 pcor_abs:abs_pcor,
                 cross_species:cross_species,
                 association:association}]-> (g2)
```

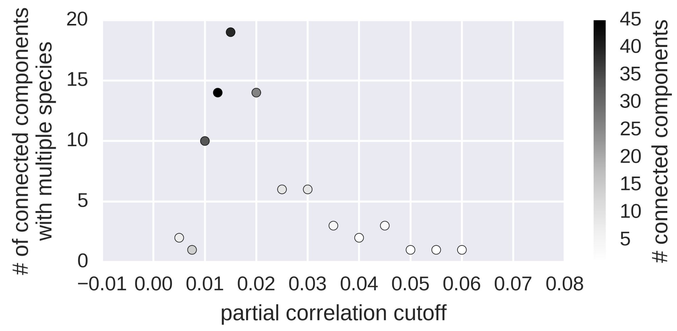
Substitute string in Java, compile in JAR that also prints build statistics, run via python's subprocess module

Results

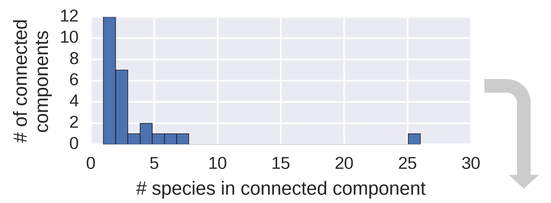
Neo4j database build time scaled with # edges



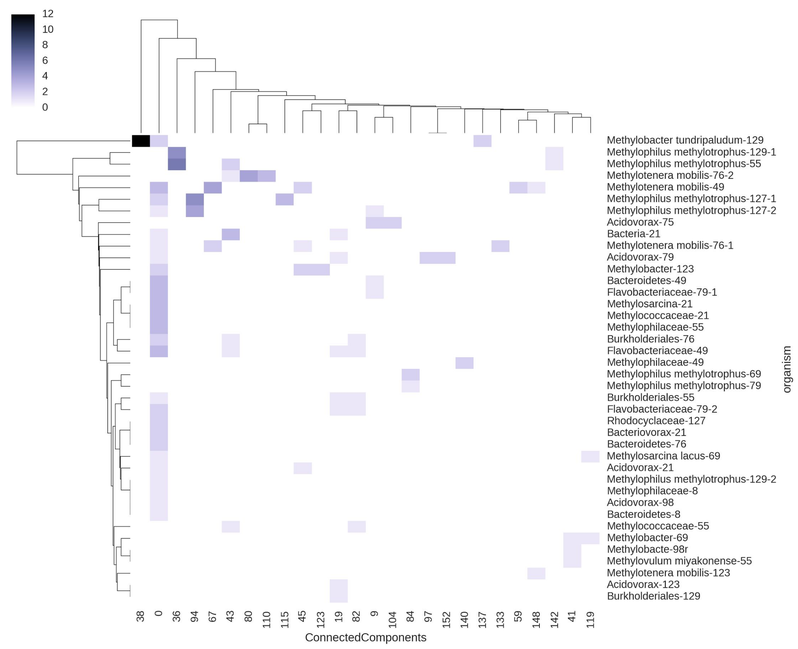
The number of multi-species connected components peaks for partial correlations >0.0125



Species count for each connected component (cutoff = 0.02)



number of genes for each organism across connected components (cutoff = 0.02)



Next Steps

- Use Cypher to investigate sub-graphs
  - `MATCH (n) where n.ConnectedComponents=333 RETURN n`
- How do interesting connected components connect to the rest of the graph?
- Query for particular motifs, e.g.
  - `MATCH (n)<- [e {cross_species:'True'}] -> (m) WHERE n.gene_product =~'.*transport.*' AND e.pcor_abs > 0.045 RETURN n, m`

References

1. Ana I Borthagaray, Matıas Arim, and Pablo A Marquet. Inferring species roles in meta-community structure from species co-occurrence networks. Proceedings of the Royal Society of London B: Biological Sciences, 281(1792):20141425, 2014.
2. Alfred Hero and Bala Rajaratnam. Hub discovery in partial correlation graphs. IEEE
3. Transactions on Information Theory, 58(9):6064–6078, 2012.
4. https://github.com/besil/Neo4jSNA

Learn more at https://github.com/JanetMatsen/Neo4j_meta4

Seattle, WA, December 2016