Outline

- Correlations
- Overcoming Challenges
- Techniques
- Dealing with Composition and Sparse Data
- Ecological Relationships
- Non-linear Models
- Ensembles
- Discussion
Importance of Microbes

- Microbes are important in complex diseases
- Interactions are important for microbial sustainability
- Correlation is a power tool for hypothesis generation
Correlation Networks

- Correlation defines a relationships between features
- Applied over vector of OTU abundances across samples / time
- Correlation thresholds defines edges that are biological or biochemical relationships
Challenges of Correlation Network Analysis

- Computationally complex to create a large number of pair-wise comparisons
- Edge presence a result of sampling biases
- Sparsity and the definition of a true zero
- Relative abundance profiles create compositional datasets
- Non-linear relationships

Measures of Performance

• Sensitivity  $\frac{TP}{TP+FN}$

• Specificity  $\frac{TN}{TN+FP}$

• Precision  $\frac{TP}{TP+FP}$

• 91 simulated data tables with different relationships

Methods to Measure Correlation

- Bray - Curtis (distance between features)
- Pearson (linear)
- Spearman (rank relationships)
- CoNet
- Local Similarity Analysis (LSA)
- Maximal Information Coefficient (MIC)
- Random Matrix Theory (MENA)
- SparCC
Naive Correlation Methods

- Pearson and Spearman
  - Fisher z-transformations used to calculate p-value with Bonferroni multiple hypothesis test correction
- Bray - Curtis
  - Empirical approach and Bonferroni multiple hypothesis test correction
CoNet

- An ensemble of five correlation metrics
- 2000 positive and negative edges selected and subjected to ReBoot routine
- ReBoot is a permutation based bootstrapping of abundances to create a null distribution
Random Matrix Theory

- Pearson correlation coefficient (r-value) is calculated between all pairs of OTUS

- RMT calculates edge thresholds based on properties of data

Maximum Information Coefficient

- MIC is performed in R package minerva with default parameters
- Empirical approach for p-value calculation similar to Bray-Curtis
- Bonferroni multiple hypothesis test correction
Local Similarity Analysis

- Implemented eLSA program with default parameters
- P-value calculated by theoretical approximation (permutation approach)
- Multiple correction performed with q-values

SparCC

- SparCC run with default parameters and 500 bootstraps
- Pseudo p-values computed using a permutation process similar to CoNet

Ws are variances of Log-transformed abundances
• Lower number of species increases the affect of compositionally

• Normalization is necessary for correlations to prevent size biases
Normalization Techniques

- Rarefying
- MetagenomeSeq's cumulative sum scaling
- DESeq's Log-ratio-based variance stabilizing transformation

Rarefying Consistency

Histogram highlights overlap of ten interactions of rarefying
Control of False Positive Rate

- Two random model
  - A) OTU abundance drawn from 6 different zero-heavy distributions
  - B) Dirichlet distribution modeled on real data
False Positive Biases

- 100 OTU draws from each distribution
- Selective false positive tendencies in different models
- Parasitic relationships may go undetected in Chi-Squared or Uniform distributions
Ecological Relationships

- Developed linear models to model relationships
- Varied the type, strength, and number of OTUs with relationships
- Introduced compositions, sparsity, or both and explored performance

Adapted from Figure 1 Faust, Karoline, and Jeroen Raes. "Microbial interactions: from networks to models." Nature reviews. Microbiology 10.8 (2012): 538.
Performance of Ecological

- Tool performances improve with strengthening of relationship
- Mutual exclusion detection limited to a few tools
- Amensal and partial obligate syntrophic relationships not found
Multi-species relationships

- Detection profiles of three member relationships similar to two member

- SparCC and LSA correctly infer competitive three member relationship

- Relationships greater than three members likely too complex for current methods
ROC Performances

- Emphasis of Precision and Sensitivity, want to predict true interactions
- Performance improves for stronger ecological relationships (figure not shown)
Removing Rare OTUs

• TP to FP plots show 50% scarcity a dramatic improvement
Non-linear Lokta-Volterra Models

- Predator-Prey Relationships
- Two species models defined in a-f (Berry and Widder)
- Six species model described in g-h (Idema)


Performance of Non-Linear Models

- Average 10% drop in sensitivity compared to linear ecological relationships
- Two species - MIC, SparCC, LSA, CoNet, & Spearman
- Six Species - SparCC & Pearson
Time Series

- Significant edges fluctuate on sampling factors
- Frequency of sampling
- Number of samplings
Improving Precision with Ensembles

• Different edges on the same data (31.5% overlap on all pairwise edges on all datasets)

• CoNet (as a single entity)
Performance of Ensemble

- Improved precision at a cost to sensitivity
- Ensemble gains compared to LSA or MIC
- Greater robustness to increase sparsity
Discussion

• Disagreement between inferred networks between different tools

• Authors suggest thresholding tools for high precision edges

• Sequencing technology and normalization techniques affect networks composition

• RMT is pair with Pearson Correlation Coefficient and can be explored with Spearman Correlation
Discussion continued

- Sparsity dramatically affects performance
- Performance degraded with tables >50% 0’s
- An ensemble approach is suggested to deal with sparsity
- Tools robust to noise are likely to perform better on real world data sets
Correlation Flow-Chart

Implications in my Work

- Viral - bacterial relationships mimic Lotka-Volterra models
- Longitudinal modeling of Cystic Fibrosis dataset
Acknowledgements


- NAWG
- Strong Lab
- Lozupone Lab