Poster Number: R-042 **Presenter:** Yong Wang **E.mail:** ywang002@fiu.edu

Ecoinformatics Tools: Unsupervised Clustering and Classification of Microbial Communities Using ALH Profile Data from 16S rRNA

Y. Wang[1], C. Yang[1], D. Mills[1], K. Mathee[1], P. Gillevet[2], J. Entry[3], K. Jayachandran[1], G. Narasimhan[1]

[1]Florida International University, Miami, FL, [2]George Mason University, Manassas, VA, [3]USDA Agricultural Research Service, Northwest Irrigation and Soils Research Laboratory, Kimberly, ID.

❑ 1. Abstract

The identification, characterization, and comparison of microbial communities are important goals in microbial ecology. Analysis of amplicon length heterogeneity (ALH) profiles of 16S rRNA genes has been used as a fingerprinting method to rapidly compare complex microbial communities and to study their diversity (Mills, *et al.*, 2003). In this study, we consider the combination of the ALH profiles obtained from different hypervariable regions within the 16S rRNA gene as a way to discriminate one community from another. An array of "unsupervised clustering" tools has been developed. These tools take combined ALH profiles from a collection of samples of microbial communities, and cluster the data into groups. To evaluate the clustering results from known samples, we introduce the concept of *Robust Entropy* that helps alleviate the effect of a limited number of outliers on the overall clustering results.. The two data sets that were studied included ALH profiles of the bacterial communities from (a) Idaho agricultural soils and (b) sediments from the Chesapeake Bay marsh regions. Our results indicate that the groupings obtained by our clustering tools are consistent with *a priori* information about the data. Our tools provide efficient and reliable methods for comparative analyses of microbial communities.

❑ 2. Introduction

- 16S rRNA was chosen as the gene target because of its well-known conservation across all the prokaryotes.
- 16S rRNA has 9 variable regions (V1 through V9). ALH profiles of the microbial communities were obtained by using Polymerase Chain Reaction (PCR) (Suzuki, M., M. S. Rappe, et al. 1998) with different primers that target the individual variable regions.
- A *combined ALH profile* was obtained by concatenating the ALH profiles obtained from using different pairs of primers on the same sample.
- A variety of clustering tools were applied to the combined ALH profiles.

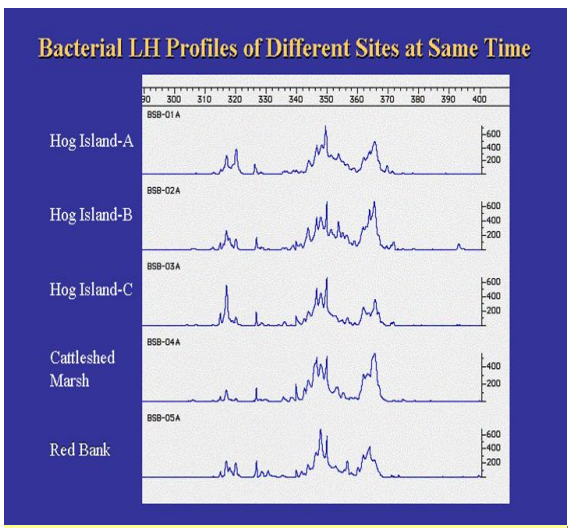

Figure 1. ALH Profiles from V1-V2 region of 16S rRNA for soil samples from five different Chesapeake Bay marsh sites.

❑ 3. Clustering Software Packages

We have built several tools to perform unsupervised clustering of combined ALH profiles:

- Hierarchical Clustering (Hastie, *et al.*, 2001) :
  - Greedy, neighbor-joining clustering method
  - Output depicted as dendrograms
  - Two distance functions used: *Euclidean Distance* and *Pearson Correlation Coefficient*
  - Prone to errors because of early greedy choices
- K-Means Clustering (Hastie, *et al.*, 2001) :
  - Has the ability to recover from early mistakes
  - Needs the specification of the number of clusters (i.e., K)
  - As K changes, the cluster membership can change dramatically
- Self-Organism Maps (SOM) (Kohonen, 1995) :
  - Similar to K-Means; Needs to specify the architecture of the map
  - Provides information about potential "neighborhood" relationships between clusters
- X-Means Clustering (Pelleg, *et al.*, 2000) :
  - Variant of K-Means, where K needs not be precisely specified
  - Chooses the number of clusters, K, based on model selection, and performs clustering
- Adaptive Self-Organism Maps (ASOM):
  - New variant of SOM, where the architecture needs not be precisely specified
  - Can find the optimal clusters and number of clusters based on model selection

❑ 4. Evaluation of Clustering Results

Three evaluation techniques were employed:

- *Entropy* is a measure of the information content or "uncertainty" of a group. It requires *a priori* grouping information for its computation. The existence of outliers can make the Entropy value become very large. Therefore, it cannot adequately measure the quality of the actual clustering results.
- *Robust Entropy* is similar to Entropy, but is relatively insensitive to noise in the data. It is the Entropy calculated after eliminating select outliers.
- *Bayesian Information Criterion* (also known as the Schwartz criterion) is a log-likelihood measure. It is used as a criterion for model selection.

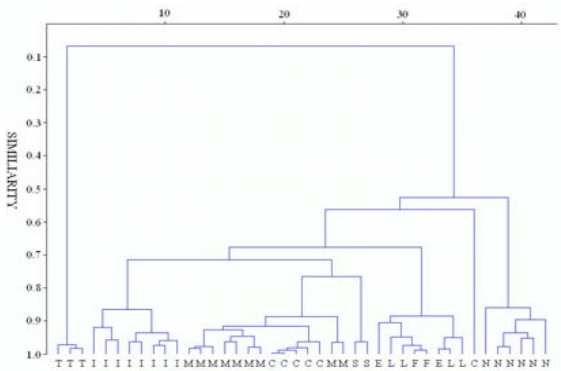

Figure 2. Hierarchical Clustering (Pearson Correlation Coefficient) Result of Idaho soil data. (C: conservation tillage E: composted manure F: commercial fertilizer I: irrigated pasture L: control, no fertilizer M: moldboard plow N: natural sage brush S: solid green manure T: composted soil)

❑ 5. Experiments and Results

The data sets that were analyzed using these unsupervised clustering tools include:

1. Combined ALH profiles from soil samples collected from Idaho (representing nine different soil management types)
2. Combined ALH profiles from soil samples collected from Virginia's Chesapeake Bay (representing nine different locations and seven different times)
3. synthetic data

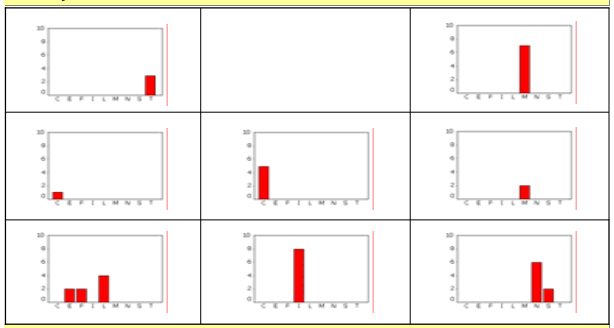

Figure 3: Adaptive SOM (ASOM) Results of Idaho soil data.

❑ 6. Conclusions and Discussion

- Combined ALH Profile of 16S rRNA genes
  - Combined ALH profiles of 16S rRNA genes can be used to characterize the microbial communities and compare them.
  - ALH profile of regions V1 and V1 plus V2 contain more useful information than region V3.
- Result Evaluation
  - It is difficult to ascertain the validity of inferences drawn from the output of unsupervised community clustering, especially for unknown data.
  - The measures of entropy and robust entropy are only meaningful for the same data and the same architecture.
  - The BIC was successfully used to make model selection. However, it takes everything into consideration, and is not very robust in the presence of outliers.
- Comparison of Unsupervised Clustering Tools
  - K-Means and X-Means tend to be influenced greatly by different initializations.
  - SOM and ASOM are more stable than K-Means and X-Means.
  - X-Means and ASOM make greedy decisions, so early decisions can greatly affect the results.
  - Overall, our results show that ASOM is the best tool among these unsupervised tools.

References and Acknowledgements

- Kohonen, T. (1995). Self-Organizing Maps. Springer Series in information sciences.
- Mills, D. K., P. M. Gillevet, et al. (2003). A Comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum contaminated soils. J. Microbiol. Meth. Vol. 54 p.57-74.
- Pelleg, D. and A. Moore (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings of the 17th International Confence on Machine Learning. p.727-734.
- Suzuki, M., M. S. Rappe, et al. (1998). Kinetic Bias in Estimates of Coastal Picoplankton Community Structure Obtained by Measurements of Small-Subunit rRNA Gene PCR Amplicon Length Heterogeneity. Applied and Environmental Microbiology Vol. 64 p. 4522-4529.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

Travel for Y. Wang is supported by Graduate Student Association and College of Arts & Sciences, FIU.