COMPSCIX-415.1 Final Project Idea

Benjamin Cole

March 31, 2018

Short Bio

I am a biologist interested in studying how plants interact with their environment using quantitative methods. I graduated with a B.S. in bioinformatics and molecular biology from the Rennsselaer Polytechnic Institute in 2006, and finished a Ph.D in biology at the University of California, San Diego in 2011. I have been a postdoc and (more recently) a project scientist at the Lawrence Berkeley National Laboratory since 2013. I am interested in learning how to apply Data Science to my own research, which is focused in three primary areas:

- Understanding the genetic basis of how plant-associated microbial communities are formed, and how they impact plant physiology.
- Understanding the complex molecular responses of Sorghum to drought stress using whole-transcriptome profiling
- Characterizing cell-type specific gene expression patterns in model organisms, and how those patterns change in response to external stimuli.

Github account

My github repositories can be found at https://github.com/b-coli/

Final project idea

With advances in sequencing technology, it is now feasible to cheaply sample the phylogenetic distribution of microbial organisms (especially bacteria) in a host of environments by sequencing the 16S ribosomal RNA molecule common to most prokaryotic organisms. These sequencing projects typically generate counts tables of Operational Taxonomic Units (OTU tables, containing counts of observations of a particular 16S rRNA sequence), and have associated metadata, including the location of the sample, the type of environment the sample was obtained from, in addition to other metrics. The Earth Microbiome Project has made publically available OTU tables from thousands of such samples, and repositories such as the Short Read Archive (at NCBI) is continually growing with such information. I would like to use the distribution (at the Class or Order level) of microbes to predict either the type of environment (host, terrestrial, marine, etc.) or another feature contained in the metadata, provided there is sufficient depth for the collected samples. This type of model could be useful in a range of applications, including sample misclassification detection, forensic investigation, and determining soil health in agricultural settings.