Metagenomics Lab – Marine Niche Modeling

Goals:

- Predict marine microbial diversity patterns globally from relatively sparse metagenomics samples using GLMs and global environmental data
- Use model selection to determine which environmental variables are most associated with and predictive of bacterial diversity

Ideas to consider:
(i) confounding factors,
(ii) correlated variables,
(iii) overfitting, and
(iv) how to assess model performance.

Data:
The lab uses data (file=training.csv) from this paper

J. Ladau, T.J. Sharpton, M.M. Finucane, G. Jospin, S.W. Kembel, J. O'Dwyer, A.F. Koeppel, J.L. Green, K.S. Pollard (2013). *Global marine bacterial diversity peaks at high latitudes in winter*, ISME Journal, 7: 1669-1677.
http://www.nature.com/ismej/journal/v7/n9/full/ismej201337a.html

and global "rasters" of environmental covariates at 0.5-degree resolution (file= FormattedRasters.zip). Unzip FormattedRasters in the directory where you save training.csv. Run R with this as your working directory.

Code:
The file metagenomics.R includes code that explores the data, fits models for log(Richness) as a function of environmental covariates at the sampling locations, and then predicts this diversity measure globally using environmental covariates. There is code for making maps with the output and masking out predictions in areas that require extrapolating too far. Richness is average number of de novo OTUs over 1000 rarefactions to 4266 sequences per sample. The code then demonstrates different techniques for choosing environmental covariates to include in the model, quantifying their importance for predicting LogRichness, and evaluating model fit.

HOMEWORK

Modify the code for predicting diversity to now predict the probability of observing the taxonomic group that includes *Prochlorococcus* (outcome variable = ProchlorococcusOccur). Note that this is a binary outcome. You will need to change the call to glm() to use the appropriate error distribution and link function. You might also want to change the covariates: the covariates that are most predictive of *Prochlorococcus* occurrence may be different from the ones in the best model for diversity. Turn in your code and a map of where the color scale denotes the predicted probability of *Prochlorococcus* occurring at that location.

Biological Note: *Prochlorococcus* in this data set also includes the genus *Synechococcus*, which could not be distinguished from genus *Prochlorococcus* using the 16S data.