Project Proposal

Cassandra Sperow

2023-10-25

Classification of Algal Clade using Bacterial Abundance

What is the relationship between bacterial abundance and algal clade? Which bacteria strains present the most impact in defining algal clade?

Project Category

Application of machine learning for classification

Team Details

Self

Introduction/Motivation

Research into the specific characteristics of coral microbiomes is nascent. A relevant field from which to draw parallels is the human microbiome research area. This area has predominantly focused on disease prediction and other health outcomes or characteristics based on the presence or absence of bacteria or other environmental factors. Within approximately the past decade, the application of machine learning techniques to explore or predict healthy vs. diseased individuals based on bacteria found in the individual have been explored in multiple studies (Boutin et al 2021, Namkung 2020). One main outcome of these studies is to increase the overall health of an individual or to estimate risk for certain diseases.

As corals are known to harbor and support at least a quarter of the ocean's marine life, coral health is an important and appropriate area of research that promises to increase understanding of protecting ecosystems around the world. Bourne et al (2016) identified that coral microbiomes are central to a reef's resilience. They can buffer, exacerbate or modify environmental effects for corals. Coral microbiomes are also believed to be specific to species.

Van Oppen and Blackall (2019) showed that heat as an environmental stress on corals was not associated with bleaching when the coral microbiome was identified to be stable. This means that when certain bacteria strains are present in the coral's algae during elevated temperature exposure, the bacteria act as regulators of the coral's ability to maintain its feeding cycle and stay healthy. A variety of different interactions occur between a coral's algae and bacteria, creating a symbiotic relationship that regulates many complex processes within coral bodies (Morris et al 2019). As ocean temperatures are expected to rise due to climate change, heat and other environmental factors are expected to increase the risk of coral bleaching and the destruction of reefs (Cressey 2016).

This project aims to use similarly applied machine learning methods to explore the relationship between a coral sample's bacterial attributes and algal clade A with the goal of understanding which bacteria present the most impact in defining algal clade A. Despite not having the data available as labeled in terms of healthy

vs. bleached coral, describing the relationship and possible predictors provide insight into the descriptive characteristics of which bacteria may most associate with algae clade A. Understanding the underlying structure of how bacteria are associated with algal clade is one step in a multi-step process of ultimately deciphering how corals behave or cope when exposed to certain stressors.

Methodology:

The proposed methodology is to use binary classification technique(s), specifically logistic classification, to model the coral sample algae clade as predicted by bacteria abundance counts. Other possible predictors include region, reef, and species. The response variable is a one-dimensional vector of algal clade C as a binary outcome of "1", while non-C observations will be labeled as "0". Each coral sample was previously profiled for its algal clade and bacterial abundance counts in previous work by Buitrago-López et al (2023).

Counts-based data from bacterial and algae genetic sequencing is known to be linearly dependent and naturally violates many statistical assumptions of normality. Data transformations are often used in the modeling process to account for the compositional nature of sequencing data. The recommended data transformation for these data is the log-centered log ratio, outlined in Godichon-Baggioni et al (2018).

The dataset is expected to be compositional and high-dimensional; therefore lasso or ridge regression in combination with logistic classification may be used, as well as Principal Components Analysis.

Intended Experiments:

Observations of coral samples may be split into two species: species "P" is *Pocillopora verrucosa*; species "S" is *Stylophora pistillata*. It is possible that splitting the data by species will help in decreasing any noise from either species; however, in the entire dataset of non-zero algae sequencing counts of ITS2 algae gene types (n = 778), there are **n = 628 observations for clade A** and **n = 139 observations of clade C** with less than 1-5 observations for clades B, D, and G, which will not be used such that the binary classification will be clade A or C. The remaining observations of non-clade C observations (n = 639) will be randomly sampled to balance the number of observations for each class label for binary classification.

For the scope of this project and given the number of observations, it may be necessary to subset 100, 500, or 1000 most abundant bacterial strains out of 5,889. This will assist is dimensionality reduction, as well as help the overall interpretability and computational expense of the modeling process. During the genetic sequencing process in Buitrago-López et al (2023), the bacteria are ranked in terms of the most abundant across all samples; however, as clade C is the second-most abundant clade in the dataset, Principal Components Analysis (PCA) will be used to create uncorrelated variables for modeling before subsetting any columns.

Several diagnostic and evaluation metrics for the logistic classification model will include:

1. ROC Plot(s)
2. Confusion Matrix
3. Accuracy Rate
4. P-value(s) of possible predictor bacteria or other variables
5. Interpretability of which predictor variables contribute the most to the classification result

If time allows, Random Forests, or Gaussian Mixture Models may also be explored.

Prior Research:

Boutin et al (2021) outlines a logistic modeling approach for predicting childhood asthma based on gut microbiome bacteria. The expected outcome of improving overall childhood health is parallel to the case of

improving coral overall health. The binary modeling approach is also parallel in terms of having two labels as outcomes. A third similarity of this use case is the nature of the bacterial data in the individual's microbiome, as these data would also be based on abundance counts to establish presence within the individual. This research takes into account the bacterial abundance counts as possible predictors as well as environmental factors, such as rural/non-rural area, exposure to mold, older sibling, and breastfeeding status, among others.

Uddin et al (2019) reviews a large number of research articles exploring machine learning methods applied to microbiome datasets. The article summarizes the success or non-success of each method as well as the advantages and disadvantages of each in terms of their application for predicting disease outcomes. Performance metrics were also compared when multiple methods were used on the same datasets in order to understand which algorithms seem the most applicable depending on the dataset.

References:

Boutin, R. C., Sbihi, H., McLaughlin, R. J., Hahn, A. S., Konwar, K. M., Loo, R. S., ... & Finlay, B. B. (2021). Composition and associations of the infant gut fungal microbiota with environmental factors and childhood allergic outcomes. MBio, 12(3), 10-1128. doi.org/10.1128/mBio.03396-20

Buitrago-López, C., Cárdenas, A., Hume, B. C. C., Gosselin, T., Staubach, F., Aranda, M., Barshis, D. J., Sawall, Y., & Voolstra, C. R. (2023). Disparate population and holobiont structure of pocilloporid corals across the Red Sea gradient demonstrate species-specific evolutionary trajectories. Molecular Ecology, 32, 2151–2173. https://doi.org/10.1111/mec.16871

Cressey, D. Coral crisis: Great Barrier Reef bleaching is "the worst we've ever seen". Nature (2016). https://doi.org/10.1038/nature.2016.19747

Godichon-Baggioni, A., Maugis-Rabusseau, C., & Rau, A. (2019). Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. Journal of Applied Statistics, 46(1), 47-65. https://doi.org/10.1080/02664763.2018.1454894

Morris, L. A., Voolstra, C. R., Quigley, K. M., Bourne, D. G., & Bay, L. K. (2019). Nutrient availability and metabolism affect the stability of coral–Symbiodiniaceae symbioses. Trends in microbiology, 27(8), 678-689. https://doi.org/10.1016/j.tim.2019.03.004

Namkung, J. (2020). Machine learning methods for microbiome studies. Journal of Microbiology, 58, 206-216. DOI 10.1007/s12275-020-0066-8

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16. doi.org/10.1186/s12911-019-1004-8

van Oppen, M. J., & Blackall, L. L. (2019). Coral microbiome dynamics, functions and design in a changing world. Nature Reviews Microbiology, 17(9), 557-567. https://doi.org/10.1038/s41579-019-0223-4