ANALYSIS OF THE SLO BAY MICROBIOME FROM A NETWORK PERSPECTIVE

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Viet Nguyen
COMMITTEE MEMBERSHIP

TITLE: Analysis of the SLO Bay Microbiome from a Network Perspective

AUTHOR: Viet Nguyen

DATE SUBMITTED: July 2021

COMMITTEE CHAIR: Theresa Migler, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Paul Anderson, Ph.D.
Professor of Computer Science
Microorganisms are key players in the ecosystem functioning. In this thesis, we developed a framework to preprocess raw microbiome data, build a correlation network, and analyze co-occurrence patterns between microbes. We then applied this framework to a marine microbiome dataset. The dataset used in this study comes from a year-long time-series to characterize the microbial communities in our coastal waters off the Cal Poly Pier. In analyzing this dataset, we were able to observe and confirm previously discovered patterns of interactions and generate hypotheses about new patterns. The analysis of co-occurrences between prokaryotic and eukaryotic taxa is relatively novel and can provide new insight into how marine microbial communities are structured and interact.
ACKNOWLEDGMENTS

Thanks to:

- My advisor, Professor Theresa Migler: you have given me lots of opportunity to apply graphs in practical ways and so much support throughout this project.

- Professor Alexis Pasulka: this project would not have been possible without your expertise and direction for it.

- My parents: you have done your best to allow me to pursue what I want. I didn’t choose to leave my family and friends to study abroad, but it turned out to be the most fulfilling path for me.
TABLE OF CONTENTS

LIST OF TABLES ............................................................... ix
LIST OF FIGURES ............................................................. x

CHAPTER

1 Introduction ............................................................... 1

2 Background ............................................................... 3
  2.1 Microbiome Data ..................................................... 4
  2.2 Gene Sequencing ..................................................... 4
  2.3 OTU Picking and Taxonomy Assignment ......................... 5
  2.4 Data format .......................................................... 6
  2.5 Correlation network ................................................ 7
  2.6 Correlation detection strategies .................................. 8
  2.7 Challenges in constructing robust networks .................... 9
    2.7.1 Computational expense ....................................... 9
    2.7.2 Compositionality bias ....................................... 9
    2.7.3 Data sparsity .................................................. 10
    2.7.4 Time series data .............................................. 11
    2.7.5 Parasitism .................................................... 11
  2.8 Interpreting networks .............................................. 11
    2.8.1 Module detection ............................................. 11
    2.8.2 Network statistics ........................................... 12

3 Related Works .......................................................... 14

4 Materials and Methods ................................................. 17
4.1 Sampling site ................................................. 18
4.2 Dataset ...................................................... 18
  4.2.1 Seasons ................................................. 20
  4.2.2 Data Filtration ........................................... 21
4.3 Correlation Detection ................................. 22
  4.3.1 Multiple testing correction ....................... 23
  4.3.2 Weak correlation filtration ....................... 24
  4.3.3 Correlation with Environmental Variables and Cross-Domain . 24
4.4 Graph Construction . ................................. 25
  4.4.1 Clustering .............................................. 26
  4.4.2 Transitivity ............................................ 27
4.5 Alpha diversity ........................................... 27
5 Results .......................................................... 28
  5.1 Environmental variables ............................ 29
  5.2 16S .......................................................... 30
    5.2.1 Community composition ....................... 30
    5.2.2 Alpha Diversity .................................. 32
    5.2.3 Network analysis ................................ 32
      5.2.3.1 16S Peak Upwelling ....................... 32
      5.2.3.2 16S Upwelling Relaxation .................. 34
      5.2.3.3 16S Year-Long Network .................... 34
  5.3 18S .......................................................... 36
    5.3.1 Community composition ....................... 36
    5.3.2 Alpha Diversity .................................. 37
    5.3.3 Network Analysis ................................ 38
5.4 Cross-Domain Network ................................................. 39
5.5 Discussion ............................................................. 39
  5.5.1 Choosing Taxonomy Level for Analysis ................. 39
  5.5.2 Importance of Diverse Analyses ......................... 41

BIBLIOGRAPHY ............................................................. 55

APPENDICES
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>A synthetic raw abundance matrix with 4 OTUs and 3 samples.</td>
</tr>
<tr>
<td>2.0</td>
<td>The relative abundance matrix produced by normalizing Table 2.0.</td>
</tr>
<tr>
<td>4.1</td>
<td>Number of OTUs resolved to each taxonomic level.</td>
</tr>
<tr>
<td>4.2</td>
<td>Number of sequences and OTUs in each network.</td>
</tr>
<tr>
<td>4.3</td>
<td>Number of positive and negative edges for each network.</td>
</tr>
<tr>
<td>5.1</td>
<td>Network metrics for most connected nodes in the year-long 16S network.</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>In this network there are covarying relationships between 2 and 3, 2 and 4, and 3 and 4.</td>
</tr>
<tr>
<td>4.1</td>
<td>Distribution of 16S library sizes</td>
</tr>
<tr>
<td>4.2</td>
<td>Distribution of 18S library sizes</td>
</tr>
<tr>
<td>5.1</td>
<td>Time series of water temperature</td>
</tr>
<tr>
<td>5.2</td>
<td>Time series of water salinity</td>
</tr>
<tr>
<td>5.3</td>
<td>Time series of Chlorophyll a</td>
</tr>
<tr>
<td>5.4</td>
<td>Relative abundance of prokaryotic OTUs. Only OTUs that have relative abundance of at least 5% and is present in at least 13 samples are shown.</td>
</tr>
<tr>
<td>5.5</td>
<td>Hill number of 16S samples</td>
</tr>
<tr>
<td>5.6</td>
<td>Correlation network of 16S taxa during the Peak Upwelling season. Green edges are positive correlations and red edges are negative correlations. Red nodes are top connected taxa and yellow nodes are top abundant taxa.</td>
</tr>
<tr>
<td>5.7</td>
<td>Correlation network of 16S taxa during the Upwelling Relaxation season. Green edges are positive correlations and red edges are negative correlations. Red nodes are top connected taxa and yellow nodes are top abundant taxa.</td>
</tr>
<tr>
<td>5.8</td>
<td>Correlation network of 16S taxa over the entire year. Network is divided into 3 clusters as shown. Green edges are positive correlations within cluster, red edges are negative correlations within cluster, blue edges are positive correlations across clusters, and yellow edges are negative correlations across clusters. Red nodes are top connected taxa and yellow nodes are top abundant taxa.</td>
</tr>
<tr>
<td>5.9</td>
<td>Normalized relative abundance of the taxa that are unique to each cluster in the year-long 16S network. Each taxon’s relative abundance values are normalized by their maximum value.</td>
</tr>
</tbody>
</table>
5.10 NMDS plot of 16S samples with significant \( p < .05 \) correlations of environmental factors.

5.11 Relative abundance of eukaryotic OTUs. Only OTUs that have relative abundance of at least 5% and is present in at least 13 samples are shown.

5.12 Relative abundance of diatoms and dinoflagellates.

5.13 Hill number of 18S samples.

5.14 Correlation network of 18S taxa over the entire year. Network is divided into 2 clusters as shown. Green edges are positive correlations within cluster, red edges are negative correlations within cluster, blue edges are positive correlations across clusters, and yellow edges are negative correlations across clusters. Red nodes are top connected taxa and yellow nodes are top abundant taxa.

5.15 NMDS plot of 18S samples with significant \( p < .05 \) correlations of environmental factors.

5.16 Cross-domain correlations over the entire year. Nodes are grouped by their 5th level taxonomy. Red nodes are top connected 16S taxa, yellow nodes are top abundant 16S taxa, blue nodes are top connected 18S taxa, green nodes are top abundant 18S taxa, and violet are top connected and abundant 18S taxa.
Chapter 1

INTRODUCTION
Microorganisms have a central place in marine biogeochemical processes, which control climate and ecosystem functioning [37]. In addition to interacting with the abiotic environment, microorganisms participate in complex biological interactions, which can have positive, negative, or neutral effects on their populations [8]. Detecting and understanding these relationships will help elucidate their impact on climate and ecosystem functioning. Microbiome datasets usually consist of abundances of operational taxonomic units (OTU) over many samples, and so their pairwise relationships are naturally predicted based on the correlations between their abundance profiles, all of which form a correlation network. Correlation networks, however, are capable of more than just a way to visualize the interaction of OTUs. They can reveal modules of interconnected OTUs [6], detect relationship between communities and environmental conditions [6], quantify modularity of a community [4], and identify keystone OTUs (OTUs that have disproportionately large effects on their communities) [5].

A motivation of this works comes from the lack of a clearly defined framework for preprocessing, constructing, and analyzing microbiome correlation networks. While there are common stages that researchers follow for microbiome analysis, tools for each stage are not standardized in their data format, and the knowledge about such tools are not centralized. Additionally, the San Luis Obispo (SLO) Bay, the sampling site of this dataset, is a relatively understudied region; analysis from a correlation network perspective is novel there and can give new insights to the marine microbiome of the bay.
Chapter 2

BACKGROUND
2.1 Microbiome Data

In this thesis, microbiome data generated through 16S and 18S rRNA gene high-throughput sequencing of the V4 region. Sequences have been clustered into operational taxonomic units (OTUs). Coupled with the microbiome data, the environmental variables including water temperature, chlorophyll \textit{a}, oxygen, phosphate, nitrate, and silicic acid concentrations, and salinity. Each sample is a snapshot of the microbiome and its environment at a specific location at a specific time point. An OTU is a classification for a group of closely related organisms, often by similarity in gene sequences. For simplicity, OTUs can be thought of as ”species”, although their precise taxonomic ranks can vary widely.

2.2 Gene Sequencing

Gene sequencing is the first step in building microbiome datasets. Water samples are collected from the environment and microbes are filtered out based on size, usually keeping organisms larger than 0.2\,\textmu m [17]. The genetic materials are then sequenced to obtain sequences of nucleotides. The two common approaches include marker gene sequencing and shotgun sequencing.

\textit{Marker gene sequencing} relies on sequencing a specific gene in the microbes’ genomes. The chosen marker gene is one that all organisms must possess. Commonly, the 16S rRNA is chosen to sequence organisms in the Bacteria domain and 18S rRNA the Eukaryota domain. The entire gene is usually too long to be sequenced and analyzed efficiently, so the rRNAs are divided into regions and only specific regions are sequenced. These regions are characterized and chosen based on their conservativeness. A region that is too conserved, or in other words, one that does not change over lin-
eages because it is too crucial in biological functioning that any mutations will cause
termination of the organism, will not be very useful in distinguishing different groups
of organisms. Conversely, a highly variable region leads to extraneous discrimination
between similar organisms.

*Shotgun sequencing,* on the other hand, is applied on all genes and on as much genetic
materials as possible. While this is widely used in sequencing microbiomes in humans,
it is not suitable for marine microbiomes for reasons described in the next section.

### 2.3 OTU Picking and Taxonomy Assignment

After obtaining the sequences from gene sequencing, *OTU picking,* the process of
grouping similar sequences, is performed. There are three approaches to OTU pick-
ing: de novo clustering, closed reference, and open reference ([http://qiime.org/
tutorials/otu_picking.html](http://qiime.org/tutorials/otu_picking.html)).

*De novo clustering* uses clustering algorithms on OTUs’ pairwise edit distances. A
threshold similarity of 97% to 99% is usually used [17]. Because different genes cannot
be meaningfully compared (e.g. 16S gene cannot be compared to 18S gene), de novo
clustering is only suitable on sequences generated by marker gene sequencing. This
method is slow because computing edit distances is slow and the clustering algorithms
used are not parallelizable.

*Closed reference* OTU picking uses reference databases of previously discovered OTUs.
Sequences in the samples are compared against sequences in the database. While this
method is fast and is applicable to data from both sequencing methods, it excludes
sampled sequences not matching any reference sequences thus excluding novel organ-
isms, which are more common in marine microbiomes than well known ones such as
the human microbiome. Reference databases are built by previous studies using de novo clustering

*Open reference* is a hybrid of the above, where closed reference is performed first and de novo clustering is performed on the remaining sequences. This method is both fast on known OTUs, while also not excluding novel ones.

After OTU picking, OTUs’ are assigned taxonomy using reference databases similar to closed reference OTU picking.

### 2.4 Data format

Microbiome datasets usually come in raw sequence reads and are converted to relative abundance matrices, where each row represents an OTU and each column represents a sample. Samples are usually taken over time points or locations. The value of each cell represents the abundance of the OTU relative to the sample. This form of data is called compositional data, because each column sums to 1.

Datasets may also be in their original form of gene reads, but are usually normalized by dividing each cell by the total number of reads in its sample, i.e. the sum of its column. Tables 2.0 and 2.0 show an example of the transformation. This normalization is necessary because the number of gene reads differs by the amount of genetic material extracted from the community or the sequencing depth [9]. Although there exist more advanced normalization techniques, they have been shown to not perform well on microbiome data [43, 26] and have not seen wide adoption in recent years.
Table 2.0: A synthetic raw abundance matrix with 4 OTUs and 3 samples.

<table>
<thead>
<tr>
<th>OTU ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>6.0</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>5.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2.0: The relative abundance matrix produced by normalizing Table 2.0.

<table>
<thead>
<tr>
<th>OTU ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.3</td>
<td>.0625</td>
<td>.25</td>
</tr>
<tr>
<td>2</td>
<td>.2</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>3</td>
<td>.1</td>
<td>.375</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>.4</td>
<td>.3125</td>
<td>.25</td>
</tr>
</tbody>
</table>

2.5 Correlation network

A correlation network, or co-varying graph or co-occurrence network, $G(V,E)$, consists of a node set, $V$, whose elements can be pairwise connected by edges in an edge set, $E$. A node can be an OTU or an environmental factor (e.g. salinity, temperature). An edge between two OTUs or between an OTU and an environmental factor represents a covarying relationship between them. Edges can represent a positive or negative correlation and can have correlation coefficients as their weights. Figure 2.1 is a synthetic network generated from the abundance data in table 2.0.

Figure 2.1: In this network there are covarying relationships between 2 and 3, 2 and 4, and 3 and 4.
Hypergraphs, generalizations of networks, can have more than two nodes that are joined by one edge. Hypergraphs allow for a natural expression of complex relationships involving many OTUs. Even though there are techniques such as associate rule mining and multiple regression to construct hypergraphs from abundance data [8], hypergraphs have rarely been used in practice.

2.6 Correlation detection strategies

The choice of correlation detection strategy is crucial in building correlation networks, because different strategies vary widely in the numbers and types of significant edges, and thus in precision and sensitivity, for the same data [43].

Some of the popular correlation detection strategies for microbial networks in recent years include Sparse Correlations for Compositional data (SparCC) [9], local similarity analysis (LSA) [32], and Spearman’s rank correlation coefficient (SRCC) [38].

SparCC [9] is specifically designed to work with compositional and sparse data. It achieves these goals by using log-ratio transformation data and by assuming that true interactions are few, i.e. most OTUs are not correlated with each other. It estimates P-value and q-value with bootstrapping. Chafee et al. used SparCC to detect correlations between free-living marine bacterioplankton [6].

LSA is designed to work with time series data and detect any kind of monotonic, time-lagged relationships [32]. It achieves these goals by using ranking variables and a time delay parameter (shifting abundance profiles). Berdjeb et al. constructed their network of protists with LSA [4]; Needham et al. used LSA on their dataset of marine phytoplankton, bacteria, and myoviruses [30].

SRCC is designed to detect any kind of monotonic correlation [38]. It is defined as the Pearson correlation coefficient between rank variables. Jones et al., Cram et al.,
2.7 Challenges in constructing robust networks

Researchers have to make many choices that can result in different networks. Taking into consideration the challenges below helps build networks that are more ecologically plausible.

2.7.1 Computational expense

The rise of high-throughput sequencing techniques has allowed the collection of large datasets with over 16,000 OTUs to be affordable [27, 14]. The number of possible pairwise interactions grows quadratically \( \binom{n(n-1)}{2} \) interactions for n OTUs). Researchers have been paying more attention to this problem and developed more efficient algorithms [46].

2.7.2 Compositionality bias

Because of the nature of compositional data, correlation measures can produce misleading results [23, 9]. It is easy to see the problem with an example. Consider a community consisting of 3 OTUs: A, B, and C. If A’s raw abundance increases independently while B and C’s raw abundances remain constant, because all relative abundances must sum to 1, A’s relative abundance will increase while B and C’s relative abundances will decrease. This artificially creates negative correlations with A and positive correlation between B and C.

Experimentally using human microbiome data, Friedman et al. showed that standard

and Trombetta et al. used SRCC to detect correlations between marine microorganisms [14, 7, 40].
correlation analysis produces 3 false species-species interactions for each true interaction (precision=.25) and misses 60% of true interactions (recall=.4) [9]. They also showed that the severity of compositional effects diminish as community diversity increases and correlation density (the amount of correlations between OTUs) decreases. Rank-based techniques such as SRCC are less affected by this problem, and SparCC was developed specifically to tackle this problem and has been shown to maintain high precision [9, 43]. Weiss et al. recommend using SparCC if the inverse Simpson index (effective number of OTUs) of the dataset is less than 13 [43].

### 2.7.3 Data sparsity

OTUs with many having zero counts in many samples require some attention. A zero count has ambiguous interpretation: either the OTU is truly absent from the sample, or more likely that it is present in a small amount [8]. Some tools such as SparCC do not work with zero counts and require additive (Laplace) smoothing. Weiss et al. showed that the ratio of true positives to false positives drops significantly for OTUs absent in more than 50% of samples [43]. They recommend filtering out such OTUs entirely although such filtering will destroy network structure. This is because a low false positive count is more important so that true interactions can be studied further.

In practice, many studies filter out rare OTUs, with minimum prevalence ranging from 10% to 40% of the number of samples [6, 30, 14, 7]. Many studies also filter out low abundance OTUs by using thresholds between 0.05% and 1% or by selecting an arbitrary number of most abundant OTUs [30, 14, 7, 4]. However, compositionality bias should be considered when filtering out OTUs, because lower diversity means higher compositionality bias [9].
2.7.4 Time series data

Unless compositionality bias is a problem, it is recommended to use LSA for time series data to detect time-lagged relationships, in which the change in population of one OTU affects at some later time point [32, 43].

2.7.5 Parasitism

It is easy to incorrectly detect parasitism as a positive correlation because while parasites may drive their host population down, their own population will simultaneously diminish as well; and when the host population grows, the parasite population will grow accordingly. Fortunately, the 3 correlation detection tools discussed above have shown to be able to detect this type of relationship as mutual exclusion [43].

2.8 Interpreting networks

Correlation networks not only provide a formalized representation of microbes’ interactions but can also afford researchers unique ways to analyze these interactions.

2.8.1 Module detection

A cluster or module is a group of vertices within a network that interact positively and densely with each other while interacting negatively and/or sparsely with other vertices. Ecologically, a cluster of taxa can be similar functionally (having overlapping niches) or phylogenetically (being related in their evolutionary) [8]. Clusters thus are very interesting when studying microbiome data. Modules are groups of highly interconnected OTUs. Studies have shown that modules consist of organisms with
similar seasonal patterns [6]. However, it may also be the case that no patterns can be found [4].

Despite not originating from network science, hierarchical agglomerative clustering has been used widely to detect modules [6, 4, 14]. Since this method is dissimilarity-based, a transformation such as \( (1 - \text{correlation}) \) is required. Although this type of transformation does ensure that more negative correlations map to greater dissimilarities, it does not represent mutual exclusion explicitly. It is also reasonable to not consider negative correlations during module detection [7]. Modules may be determined qualitatively by visually inspecting the network [7]. However, this method is impractical with large datasets.

LambdaCC [41] is a framework that generalizes multiple graph clustering objectives including modularity, sparsest cut, and cluster deletion using a parameter \( \lambda \). It was shown to work on collaboration networks and social networks.

*manta* [33] is the only algorithm we found that was experimented on biological networks, is designed to consider negative edge weights, and differentiates between weak and strong cluster assignments.

### 2.8.2 Network statistics

Across tools, multiple studies have reported that detected correlations are overwhelmingly positive [6, 4, 43, 30, 40, 21]. The majority of the studies above find that this domination is plausible and have hypotheses to explain it [21]. However, Weiss et al. showed that all correlation detection strategies they benchmarked, which include the 3 discussed in this review, have poor performance on relationships other than mutualism (win-win) and commensalism (win-neutral) [43].

The degree of a node is the number of nodes connected to it. Studies have reported
that correlation networks’ degree distributions follow the power law degree distribution [4]. However, not all biological networks follow the power law distribution [14, 19].

Nodes with high degrees can be interpreted as keystone OTUs because their removals can impact the community structure greatly [8].

Average path length (APL) is the mean of the lengths of the shortest paths between all pairs of nodes. A small APL in microbial networks is interpreted to respond quickly to perturbations [8]. Studies have shown that microbial networks have small APLs [4, 14, 8].

Modularity and clustering coefficient are heuristics to measure how well the network can be divided into highly connected modules. Studies have reported high numbers or significant differences from random network models in both measures [4, 14].
Chapter 3

RELATED WORKS
Many studies have applied correlation networks to study microbial communities, both in marine ecosystems and freshwater ecosystems. They have found modules of microbes that either have direct relationships (e.g. trophic) or share similar environmental preferences.

Chafee et al. studied the bacterioplankton community in a temperate coastal marine environment[6]. Using weekly sampled data over 2 years, they identified modules of co-varying microbes. The modules’ abundances appear to be driven by oligotrophic and eutrophic conditions, which are controlled by seasonal phytoplankton blooms.

Cram et al. studied a marine bacterioplankton community at five layers of different depths: 5m, the deep chlorophyll maximum layer (between 5m and 40m), 150m, 500m, and 890m (just above the seafloor)[7]. Within the 2 uppermost layers and the deepest layer, they found two modules of microbes and environmental parameters where correlations are positive within each module and negative across modules indicating periodic shifts between two environmental conditions. Investigating time-lagged interactions between-depth, two-thirds of the correlations characterize a pattern of shallower population changes affecting deeper populations.

Jones et al. studied one year of monthly sampled data of bacteria and microbial eukaryotes at two locations within a low-salinity lake[14]. This is one of the few works in our literature review where both domains are studied and cross-domain correlation analysis is performed. The authors found correlation patterns that reflect the seasonality of the lake and disturbance events, including a bloom of the ecologically disruptive P. parvum.

While numerous works have indicated that environmental factors drive microbial communities’ composition, short-term (daily to weekly) variability of marine protist communities studied by Berdjeb et al.[4] cannot be fully explained by changes in their
environment, but also by interspecific relationships, stochastic processes, mixing with adjacent water, and advection.

While correlation network analysis on microbiome data is not novel, the field is lacking a defined framework for approaching network analysis. Furthermore, the impact of difference methodological choices on the analysis and interpretation of microbiome networks has not been thoroughly investigated. Weiss et al. benchmarked different correlation detection strategies, a crucial choice in constructing a robust network, and make recommendations for their usage[43]. They also noted that while some techniques are better than others, each technique has their own caveat to be cautious of.
Chapter 4

MATERIALS AND METHODS
4.1 Sampling site

Samples were collected as part of the Cal Poly Pier Microbial Oceanography Time-series, established by the Pasulka Lab in 2017. The Cal Poly Pier is located in the San Luis Obispo (SLO) Bay, which is a semi-enclosed embayment along the central coast of California and an understudied part of the California Current System [2].

4.2 Dataset

Our dataset consists of sequence counts of 28 samples collected from the Cal Poly Pier from January 2018 to December 18 with the goal of capturing the seasonal dynamics of microbial communities. Following amplicon sequencing of both the prokaryotic and eukaryotic communities via 16S rRNA and 18S rRNA and OTU clustering, the dataset consists of 63,814 16S prokaryotic OTUs and 9,983 18S eukaryotic OTUs in 28 samples. The time gaps between samples range from one to five weeks. It is important to note that the prokaryotic dataset was reduced to 27 samples due to sequencing issues with one sample.

The taxonomic hierarchy of prokaryotic OTUs includes seven levels: domain, kingdom, phylum, class, order, family, and genus. Prokaryotic OTUs are divided into two domains: Archaea and Bacteria. Within each domain Archaea and Bacteria, there is only one kingdom: Archaeabacteria and Eubacteria, respectively. Therefore, the kingdom level is omitted from our data and thus the taxonomic depth shown in the data is only six. Some OTUs are identified down to the most granular taxonomic level, genus, while some can only be identified to higher levels. Table 4.1 shows the number of OTUs that are resolved to each level, 6 being the most granular and lowest. Most
of the 16S OTUs (75%) are identified to the this level. This table, as well as most
data manipulation steps, are done using the pandas library in Python [31, 44].

The taxonomic hierarchy of eukaryotic OTUs includes nine levels. All eukaryotic
microbes are in domain Eukaryota and kingdom Protista. Therefore, similar to the
prokaryotic dataset, the kingdom level is omitted from and the taxonomic depth
shown in the data is only eight. In contrast to prokaryotic taxonomic ranks, eukaryotic
taxonomic ranks are more messy, as some ranks also have a sub-rank. For example,
the fifth level taxonomic rank may be the of one OTU may be its family, but may be
the suborder of another OTU. Table 4.1 shows the number of OTUs that are resolved
to each level, 8 being the most resolved. Most of the 18S OTUs are identified to this
level.

<table>
<thead>
<tr>
<th>Level</th>
<th>Number of prokaryotic OTUs</th>
<th>Number of eukaryotic OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5 (0.01%)</td>
<td>2 (0.02%)</td>
</tr>
<tr>
<td>2</td>
<td>176 (0.28%)</td>
<td>2 (0.02%)</td>
</tr>
<tr>
<td>3</td>
<td>618 (0.97%)</td>
<td>10 (0.10%)</td>
</tr>
<tr>
<td>4</td>
<td>1038 (1.63%)</td>
<td>160 (1.60%)</td>
</tr>
<tr>
<td>5</td>
<td>13851 (21.71%)</td>
<td>228 (2.28%)</td>
</tr>
<tr>
<td>6</td>
<td>48126 (75.42%)</td>
<td>265 (2.65%)</td>
</tr>
<tr>
<td>7</td>
<td>NA</td>
<td>1058 (10.60%)</td>
</tr>
<tr>
<td>8</td>
<td>NA</td>
<td>8258 (82.72%)</td>
</tr>
<tr>
<td>Total</td>
<td>63814 (100%)</td>
<td>9983 (100%)</td>
</tr>
</tbody>
</table>

Table 4.1: Number of OTUs resolved to each taxonomic level

The numbers of sequences in samples vary widely. This variation does not reflect
the true richness of the microbiome but is rather an artifact of the sampling and
sequencing processes. The distributions of these sequence counts, or library sizes, are
shown in figures 4.1 and 4.2.

In addition to the OTU counts, each sample has corresponding metadata: the tem-
perature (in degree Celsius), salinity (gram of salt per kilogram of water), and Chloro-
phyll a concentration (microgram of Chlorophyll a per liter of water) of the water at the time and location the sample was taken.

### 4.2.1 Seasons

The coastal marine environment of central California is regulated by winds driving upwelling, a process where nutrient-rich water deep in the ocean is brought to replace the nutrient-depleted water at the coast [11]. This results in seasonal nutrient conditions that drive OTUs’ abundances, as observed in previous works [6]. Peak Upwelling represents the period when this process takes place most aggressively, when Upwelling Relaxation represents the period when winds that drive upwelling dwindle down or reverse.
In our dataset, there are three seasons: Winter Storms (spaning from December to February), Peak Upwelling (April-May), and Upwelling Relaxation (July-September). All of the other months are transitions between the seasons. We analyze each dataset over multiple time period: Peak Upwelling season, Upwelling Relaxation season, and the entire year.

4.2.2 Data Filtration

Having many OTUs can be both an advantage and disadvantage. On one hand, it increases the diversity of the dataset, thus lowers the compositionality effect when calculating correlations between OTUs. On the other hand, most of the OTUs are too rare to be significant ecologically, and too sparse (not appearing in many samples)
to calculate correlations accurately for (see section 2.7.3). Additionally, the number of hypothesis tests done when calculating correlations grows quadratically with the number of OTUs, which leads to the multiple testing problem.

Following previous works as described in section 2.7.3, we filter OTUs as follows:

- For seasonal networks (16S Peak Upwelling, 16S Upwelling Relaxation, 18S Peak Upwelling, and 18S Upwelling Relaxation), we retain OTUs that have relative abundance of at least .1% in at least one sample and are present in at least three samples, or half of the number of samples in each season as recommended by Weiss et al.[43].

- For year-long networks (16S, 18S, and cross-domain), we retain OTUs that have relative abundance of at least 1% in at least one sample and are present in at least three samples.

<table>
<thead>
<tr>
<th></th>
<th>Number of sequences</th>
<th>Number of OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S Peak Upwelling</td>
<td>466,498 (78.45%)</td>
<td>227 (0.78%)</td>
</tr>
<tr>
<td>16S Upwelling Relaxation</td>
<td>295,254 (76.79%)</td>
<td>235 (1.16%)</td>
</tr>
<tr>
<td>18S Peak Upwelling</td>
<td>1,256,105 (58.30%)</td>
<td>75 (0.11%)</td>
</tr>
<tr>
<td>18S Year-Long</td>
<td>363,184 (87.49%)</td>
<td>214 (3.56%)</td>
</tr>
<tr>
<td>18S Upwelling Relaxation</td>
<td>191,426 (87.47%)</td>
<td>115 (2.85%)</td>
</tr>
<tr>
<td>18S Year-Long</td>
<td>1,019,242 (76.95%)</td>
<td>98 (0.98%)</td>
</tr>
</tbody>
</table>

Table 4.2: Number of sequences and OTUs in each network

4.3 Correlation Detection

We choose to use SparCC [9] to detect correlation between OTUs based on their abundances, because it is appropriate for our dataset for many reasons. First, the dataset is unevenly sampled, i.e. the temporal gaps between samples are not equal, and Local Similarity Analysis (LSA), a good metric for time series correlation, assumes the
dataset is evenly sampled. Our dataset is sparse in the sense that there are many
zero sequence counts (90% of the 16S and 85% of the 18S abundance tables’ entries
are zeros), and SparCC can deal with this by adding small pseudocounts (equivalent
to Laplace smoothing). SparCC has also been shown to perform well on a diverse set
of datasets[43]. Specifically, we use FastSpar (github.com/scwatts/FastSpar) [42],
an faster implementation of SparCC with an unbiased P-value estimator.

The parameters that were used to run FastSpar on all subsets of our dataset is as
follows: 50 iterations, 10 exclusion iterations, correlation strength exclusion threshold
of 0.1, and random number generator seed of 1. The p-values are calculated using
bootstrapping with 1,000 random permutations, with 5 iterations per permutation.

One drawback with correlation detection measures in general is that the correlations
they detect do not indicate causal relationships, nor are there directions. This again
emphasizes the hypothesis-generating nature of the analysis, and significant or inter-
esting relationships have to be tested ecologically. A correlation between two OTUs
may be caused by a relationship with a third OTU or with other environmental
variables.

4.3.1 Multiple testing correction

For the year-long networks, we correct for multiple testing using the Benjamini-
Hochberg procedure [3] using an alpha level of 0.05. For the seasonal networks,
because of the limited number of samples, after correcting for multiple testing, all of
the correlations are not significant. Because this is a hypothesis generating study and
we do not conclude biological relationships without further biological experiments,
we think that it is appropriate to accept a higher rate of false discovery. Specifically,
we use an uncorrected alpha level of .001 for the seasonal networks.
4.3.2 Weak correlation filtration

SparCC correlations range from -1 to 1, where negative values indicate adverse relationships such as parasitism or predation, and positive values indicate mutualism. The absolute value of a correlation represents how strong the relationship is. For the seasonal networks, all correlations have absolute values greater than 0.8, so we do not filter out any weak correlations.

<table>
<thead>
<tr>
<th>Network</th>
<th>Positive edges</th>
<th>Negative edges</th>
<th>Edges</th>
<th>Nodes</th>
<th>Transitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S Peak Upwelling</td>
<td>64 (85.33%)</td>
<td>11 (14.67%)</td>
<td>75</td>
<td>75</td>
<td>0.52</td>
</tr>
<tr>
<td>16S Upwelling Relaxation</td>
<td>2 (18.18%)</td>
<td>9 (81.82%)</td>
<td>11</td>
<td>59</td>
<td>0.00</td>
</tr>
<tr>
<td>16S Year-Long</td>
<td>146 (55.09%)</td>
<td>119 (44.91%)</td>
<td>265</td>
<td>54</td>
<td>0.51</td>
</tr>
<tr>
<td>18S Peak Upwelling</td>
<td>5 (50.00%)</td>
<td>5 (50.00%)</td>
<td>10</td>
<td>81</td>
<td>0.00</td>
</tr>
<tr>
<td>18S Upwelling Relaxation</td>
<td>0 (0.00%)</td>
<td>1 (100.00%)</td>
<td>1</td>
<td>14</td>
<td>0.00</td>
</tr>
<tr>
<td>18S Year-Long</td>
<td>121 (64.71%)</td>
<td>66 (35.29%)</td>
<td>187</td>
<td>50</td>
<td>0.44</td>
</tr>
<tr>
<td>Cross-Domain</td>
<td>132 (52.38%)</td>
<td>120 (47.62%)</td>
<td>252</td>
<td>77</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4.3: Number of positive and negative edges for each network

4.3.3 Correlation with Environmental Variables and Cross-Domain

Aside from correlations between taxa, it can also be interesting to look at correlations between taxa and environmental variables. Since environmental variables are not compositional with the sequence counts, it is not appropriate to use SparCC. Following Cram et al.[7], who used LSA (which in turn uses Spearman’s correlation underneath), we use Spearman’s correlation because our samples are not evenly spaced. Environmental variables are not transformed before correlation detection, but OTUs are normalized by the total sequence count of each sample.

Following Mikhailov et al. [28] and Jones et al. [14], we also use Spearman’s correlation to calculate correlations between prokaryotic OTUs and eukaryotic OTUs. It is not logical to concatenate the 16S sequence counts and 18S sequence counts and cal-
culate SparCC correlations because prokaryotic and eukaryotic OTUs are sequenced separately. We thus normalize the two dataset separately (again, by the total sequence counts of each sample), and use the relative abundance vectors to calculate Spearman’s correlations (only between each prokaryotic and eukaryotic OTUs).

4.4 Graph Construction

We model the community as undirected, weighted graphs using the NetworkX library for Python [13]. Each graph represents the covariance relationships between taxa in a period of time. In each graph $G = (V, E)$, nodes represent taxa, and an edge between two taxa represents a correlation between them and has weight between -1 and 1, indicating relationships as explained in section 4.3.2. Edges are undirected because it is not plausible to detect causal relationships and directions with just abundance dynamics, and more experimentation is necessary to detect those.

Initially, all nodes in the graph are OTUs, as used in correlation detection. For effective visualization, we combine nodes based on their taxonomy and correlations, up to level 5 of the taxonomy. Level 5 roughly corresponds to the family level in most OTU’s taxonomies, although this may not be true for all OTUs. For each level from the OTU level to the 5th level, we group nodes by their taxonomy up to that level. For each such group, we check if (1) there are any negative correlations among the OTUs in the group and (2) if there are conflicting correlations with OTUs outside the group (e.g. OTU A within the group correlates positively with OTU C outside the group, but OTU B within the group correlates negatively with OTU C). We combine the group into a node if both conditions are false. If we do combine, in NetworkX, it is done by creating a new ”combined” node and creating edges between it and neighbors of the group where the weights of the new edges are the average of the
weights of the edges between the external OTUs and the nodes in the group. For example, if the correlations between nodes A, B within the group and node C outside the group are 0.7, 0.9, then the correlation between the "combined" node with C is \((0.7 + 0.9)/2 = 0.8\). Note that we may combine even when not all nodes in the group correlate with an external OTU, so this may overemphasize some patterns and should be kept in mind for analysis.

### 4.4.1 Clustering

To cluster the networks, we use *manta* ([github.com/ramellose/manta](https://github.com/ramellose/manta)) [33]. The graphs constructed with SparCC correlations are exported to the GML format. We run the *manta* script with the following configuration: minimum cluster size of 0, number of clusters from 2 to 10, and at least 90% of the edges must be used if the graph is not balanced. Because *manta* has a randomized component that may output different number of clusters and minimally different cluster assignments, we run the process 20 times for each graph and choose the majority result.

We only ran *manta* on the two year-long networks (16S and 18S) but not the seasonal networks (16S PU, 16S UR, 18S PU, 18S UR) because *manta* does not work with networks that contain low numbers of edges. We thus cluster the seasonal networks visually. The networks are imported into Cytoscape [35], and nodes are first arranged using the "Prefuse force directed layout", where nodes that share positive edges are put close together, while nodes that share negative edges are pushed away from each other. Some initial clusters are formed by this layout, but we further examine each cluster and if a node is not too well connected with the rest of the cluster, we move it out to its own cluster.
4.4.2 Transitivity

A triplet is three connected nodes, either by two edges (open triplet) or three edges (closed triplet). Transitivity is calculated as follows using NetworkX:

\[
\frac{\text{number of closed triplets}}{\text{number of all triplets}}
\]

Note that a clique of three nodes counts as three closed triplets.

We want to measure how significantly different the transitivity of these microbiome networks are from random networks created using Erdős–Rényi with the same number of nodes and edges. We do this by creating a distribution of 1000 such random networks’ transitivity, and calculate portion of the distribution that are greater than the corresponding real network’s transitivity.

4.5 Alpha diversity

Alpha diversity is how diverse each sample is (as oppose to beta diversity, which measures diversity between samples) [45]. We measures alpha diversity using the effective number of species, or Hill number of order 1 defined as

\[
\exp\left(\sum -p_i \ln(p_i)\right)
\]

where \( p_i \) is the relative abundance of OTU \( i \). The Hill number (of order 1) ranges from 1 to the number of OTUs in the sample, and indicates the number of equally abundant OTUs needed to achieve the same Hill number as the sample.
Chapter 5

RESULTS
5.1 Environmental variables

Time series of the sampling site’s temperature, salinity, and Chlorophyll a are shown in figures 5.1, 5.2, and 5.3, respectively. Temperature from January to June is generally lower than the rest of the year. There is little change in salinity. Chlorophyll a level experiences a major spike from August to October.

Figure 5.1: Time series of water temperature
5.2 16S

5.2.1 Community composition

The composition of the most abundant 16S OTUs is shown in figure 5.4. These 14 OTUs are the ones that exceed 5% relative abundance in any given sample, and account for as much as 45% of a sample’s sequence counts. The most dominant OTU in each sample accounts for as much as 22% of the sample’s sequence counts. The composition of these most abundant OTUs varies seasonally, and it is rarely the case that one OTU dominates for more than 3 consecutive samples (1.5 months) at a time. However, Fuhrman et al. have noted that microbial systems can have
substantial changes in as little as a few hours [10], and in our dataset these patterns cannot be observed. Rather, changes that occur at the weakly to monthly timescales reflect microbial communities’ responses to variation in their environment including temperature upwelling [10].

Dominance is exchanged between a few OTUs: 83481, 434162, 205859, 126851, 338877, 286696, 411113. OTU 83481, the NAC11-7 lineage of the Roseobacter clade, dominates during the end of the Peak Upwelling season and the transision period to Upwelling Relaxation. During Peak Upwelling, the 14 OTUs amount to 37.7% of the community, the highest of the year on average. This number drops quickly to 23.2% during Upwelling Relaxation, the lowest of the year.
5.2.2 Alpha Diversity

The effective number of species (Hill number) over samples is shown in figure 5.5. This number can take values from 1, when one OTU accounts for 100% of the community’s abundance, to the number of OTUs in the dataset, 63,814, when all OTUs are equally abundant. The 16S community’s diversity varies seasonally, and is the highest in the Winter Storms and Upwelling Relaxation periods and lowest in the Peak Upwelling season. This makes a lot of sense when we look at figure 5.4 that during Peak Upwelling, a few OTUs account for a majority of the community’s abundance.

5.2.3 Network analysis

As mentioned in section 4.2.1, we built and analyzed three 16S networks: 16S Peak Upwelling, 16S Upwelling Relaxation, and 16S Year-Long. We discovered patterns in seasonal networks that are not present in the year-long network.

5.2.3.1 16S Peak Upwelling

The correlation network of prokaryotic taxa during Peak Upwelling season is shown in figure 5.6. As shown in table 4.3, the edges are primarily positive, which is consistent with the literature [6, 4, 43, 30, 40, 21]. The most connected nodes are the NS9 marine group, Polaribacter, and Tenacibaculum, each correlating with nine to ten other nodes. The transitivity is 0.52, meaning that there is a 52% probability that node B and C are connected if they are both connected with another node A. This indicates a well structured network and is significantly different from random networks of the same size (p-value=0.00).
To analyze microbial networks, we first look at the five most abundant (colored yellow) and five most connected (colored red) taxa. The most abundant taxa are identified by the average relative abundance over the samples. There are no taxa that are both highly abundant and well connected, which is a notable pattern. This validates the accuracy of SparCC, because highly abundant taxa are likely to have spurious negative correlations [43].

OTUs of the NS9 marine group are the most connected, despite being low in abundance at only 0.13% over the season. It has negative correlations with OTUs of the Polaribacter and Ulvibacter genera. It was observed previously by Seo et al. [34] that the NS9 marine group tend to populate coastal environments (which is where the sampling site is) while Polaribacter and Ulvibacter reside in further out in the ocean. However this spatial variability does not explain the negative correlations observed over time at the same sampling site. Furthermore, abundances of Polaribacter and Ulvibacter at our coastal sampling site are equal to and greater than that of NS9, respectively.

OTUs of the Polaribacter and Tenacibaculum covary positively and share 75% of their correlations with other taxa. Habib et al. previously found that some isolates of Tenacibaculum tend to phylogenetically cluster with a strain of Polaribacter [12]. Because the Tenacibaculum genus is known to include several fish pathogens, Habib et al. also speculated that pathogenic lineages are also in the Polaribacter genus, but a conclusion would require more analysis on complete genome data [12].

As mentioned in 4.4.1, the manta clustering algorithm could not cluster the network because of the low number of edges, thus we cluster the network manually based on some heuristics. The most significant cluster is the almost clique-like Rhodobacteraceae group. The NS9 marine group and the sister clades Polaribacter and Tenacibac-
ulum seem to form to opposing clusters, but the ecological significance of this division is unclear.

5.2.3.2 16S Upwelling Relaxation

The correlation network of prokaryotes during Upwelling Relaxation season is shown in figure 5.7. The edges, as shown in table 4.3, are mainly negative. This contradicts previous literature, which observed that edges in microbiome correlation networks were overwhelmingly positive [6, 4, 43, 30, 40, 21]. The negative edges could be an artifact of the compositional effect, but is unlikely because of the relatively high alpha diversity during the season, which is much greater than the minimum effective number of species of 13 recommended by Weiss et al. [43]. The transitivity is 0, because there are no connected triplets.

Similar to the Peak Upwelling season, there is no overlap between the top five most connected nodes and the top five most abundant nodes, which are notably are isolated.

The genera Polaribacter and Tenacibaculum share no common patterns of correlations that are observed in the Peak Upwelling season.

5.2.3.3 16S Year-Long Network

Correlations between prokaryotes over the year are shown in figure 5.8. There are 10% more positive edges than negative edges. Polaribacter and the NS9 marine group still have the most correlations similar to the Peak Upwelling network, but Synechococcus, which was not very connected during Peak Upwelling, has taken Tenacibaculum's place. Tenacibaculum does not appear in the year-long network because it does not
meet the higher minimum abundance requirement. The transitivity is 0.51, similar to that of the Peak Upwelling network (p-value=0.00).

*manta* is able to put the nodes into 3 clusters, where the majority are intra-cluster positive edges (green) and inter-cluster negative edges (orange), while there are few inter-cluster positive ones (blue) and intra-cluster negative ones (red). The relationships between the three clusters are structurally unbalanced, because the three clusters have negative relationships with each other. This goes against the assumption "the enemy of my enemy is my friend" that *manta* makes.

The most connected nodes are distributed among the three clusters. While it may be tempting to interpret them as keystone taxa, there are many caveats as discussed by Berry and Widder[5]. They identified four predictors of a keystone: high degree, low betweenness centrality, high closeness centrality, and high transitivity [5]. These metrics are shown in table 5.1. We see that none of these nodes meet all of the above requirements, thus further experiments are needed before conclusion.

<table>
<thead>
<tr>
<th>Node</th>
<th>Degree rank</th>
<th>Betweeness centrality rank</th>
<th>Closeness centrality rank</th>
<th>Transitivity rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS9 Marine Group</td>
<td>1st</td>
<td>7th</td>
<td>1st</td>
<td>52nd</td>
</tr>
<tr>
<td>Polaribacter</td>
<td>1st</td>
<td>5th</td>
<td>3rd</td>
<td>51st</td>
</tr>
<tr>
<td>Synechococcus</td>
<td>5th</td>
<td>16th</td>
<td>4th</td>
<td>44th</td>
</tr>
<tr>
<td>Sulfitobacter</td>
<td>3rd</td>
<td>13th</td>
<td>9th</td>
<td>40th</td>
</tr>
<tr>
<td>denovo59236 (Rhodobacteraceae)</td>
<td>4th</td>
<td>11th</td>
<td>40th</td>
<td>41st</td>
</tr>
</tbody>
</table>

Table 5.1: Network metrics for most connected nodes in the year-long 16S network

In figure 5.9, we plot the population dynamics of taxa that are unique to each cluster and exceed 5% relative abundance in any given sample. As expected, we see taxa in the same cluster having similar patterns of relative abundance. Taxa in the left cluster tend to peak in abundance from July until November, taxa in the upper right cluster tend to peak from March to May, while those in the lower right cluster do not have a clear pattern. This pattern corresponds very well with the dynamics of
water temperature, where it is significantly hotter from July to November than the rest of the year (figure 5.1). From the network point of view, temperature correlates positively with taxa in the left cluster and negatively with those in the upper right cluster, while having mixed correlations with the lower right cluster (figure 5.8). In the NMDS plot (figure 5.10), temperature correlates positively with samples from September to November and negatively with most samples from March to June.

Water salinity has been identified as a major environmental determinant of microbial community composition [24]. Salinity positively correlates with the population of many OTUs in the left cluster, established above to be one that grows from July to November (figure 5.8). However, salinity is relatively low during this period, so whether there is a relationship between the population dynamics of the entire left cluster with salinity is inconclusive. In the NMDS plot (figure 5.10), salinity almost points in the opposite direction of temperature and Chlorophyll a, correlating negatively with samples from September to November and contradicting the pattern shown in the correlation network (figure 5.8).

5.3 18S

5.3.1 Community composition

The composition of the most abundant 18S OTUs is shown in figure 5.11. These 19 OTUs are the ones that exceed 5% relative abundance in any given sample, and account for as much as 93% of a sample’s sequence counts. The most dominant OTU in each sample accounts for as much as 93% of the sample’s sequence counts. Although the composition of these most abundant OTUs does vary seasonally, dissimilar to the prokaryotic community, OTUs of the phylum Dinoflagellata, including Clade 10 and 11 of Dino Group II (5513), , Gyrodinium (21749), Clade 30 of Dino Group II
(25261), Tripos fusus (16339), Tripos concilians (28426), and Margalefidinium fulvescens (4910) dominate the eukaryotic community throughout the year with the occasional interruption of diatoms, specifically Minidiscus trioculatus (17845).

A study done by Barth et al. [2] showed that the in abundance of the eukaryotic community in the San Luis Obispo Bay (the same location our data were collected from) is dominated by diatoms (Bacillariophyceae class, or level 4 of Bacillariophyta in our data) during the spring upwelling season and by dinoflagellates (Dinoflagellata superclass) during the fall relaxation season. The pattern of phytoplankton seasonality is less clear with molecular data relative to the microscopy data used in Barth et al. [2] as shown in figure 5.12. Diatoms are only dominant at the beginning of February (which is not during upwelling) and the beginning of May (which is during upwelling).

5.3.2 Alpha Diversity

The effective number of species (Hill number) over samples is shown in figure 5.13. This number can take values from 1, when one OTU accounts for 100% of the community’s abundance, to the number of OTUs in the dataset, 9,983, when all OTUs are equally abundant. The 18S community’s diversity varies seasonally, and is the highest at the end of the Winter Storms season and during the March Transition period and lowest at the end of the Upwelling Relaxation season and during the Winter Transition period. During these low diversity periods, the community is dominated by one dinoflagellate that accounts for as much as 93% community’s abundance.
5.3.3 Network Analysis

While we built the seasonal 18S networks, each season was dominated by a few nodes and there is little dynamics. Thus those networks are unstructured, indicated by a transitivity of 0 (table 4.3). Such a low transitivity is unusual for microbiome covariance networks [8]. We therefore only analyze the year-long network, shown in figure 5.14. The transitivity is 0.44 and is significantly different from random networks of the same size (p-value=0.00).

The clusters in this network are not as clear-cut as the ones in the 16S network, having multiple inter-cluster positive correlations and intra-cluster negative correlations. This is also exhibited by the high number of weak assignments by *manta*. Most of the nodes are dinoflagellates, and they also account for the most abundant nodes, as expected from the community composition analysis above. The only two diatom nodes are in the right cluster (Radial-centric-basal-Coscinodiscophyceae and Polar-centric-Medlrophyceae). While the two nodes correlate positively, Medlrophyceae is significantly more connected, and correlates negatively with dinoflagellates. Barth et al. observed a rotation in dominance between diatoms and dinoflagellates at the same sampling site [2], which explains this negative relationship.

The negative relationships within the right cluster are particularly interesting because they are associated with parasitic dinoflagellates, which are an important group that are often missed by microscopy studies.

One of the most connected taxa, Cryptomonadales, is positively correlated with three out of four taxa in the Ciliophora phylum (ciliates). A common genus among the Cryptomonadales lineages is *Teleaulax*, which is known to be a plastid donor to ciliates, [16] which may explain these positive relationships.
The two clusters look to correspond to the upwelling relaxation (left cluster) and peak upwelling relaxation (right cluster), based on their correlations with temperature and temperature variation over the months (figures 5.1 and 5.15).

5.4 Cross-Domain Network

Microbial prokaryotes and eukaryotes also interact with each other, and previous studies have looked at this [14, 28]. A subset of cross-domain covariances are shown in figure 5.16. The network’s transitivity is 0 not because there is a lack of structure, but because by definition, the network is bipartite (every edge must connect a prokaryote and a eukaryote) and thus there are no closed triplets.

Rhodobacteraceae has been noted to be a major vitamin supplier for eukaryotes, including dinoflagellates [36]. In our networks, several OTUs of the phylum Dinoflagellata correlate densely and positively with OTUs of the family Rhodobacteraceae, displaying this relationship. One OTU of the Ciliophora phylum (ciliates) also covaries with Rhodobacteraceae, and while we have not found previous literature on the relationship between these two taxa, Lee and Kugrans [18] have compared ciliates with dinoflagellates and found similarities between the two phyla’s feeding apparatuses.

5.5 Discussion

5.5.1 Choosing Taxonomy Level for Analysis

Choosing the taxonomy level to construct the network at is tricky because different patterns emerge at different levels. Typically, it is desirable to differentiate organisms
based on their ecological niche, but currently there is no consensus on the sequence similarity level (common values are 97%, 98%, and 99%) necessary to differentiate them [10]. Even more challenging is that the differentiating threshold varies for different lineages, such that two organisms whose sequences are 97% similar can be ecologically identical while another pair of organisms that have 98% sequence similarity are ecologically distinct.

Studies we have looked at build their networks at the OTU level with sequence similarity ranging from 97% to 99% [4, 30, 14, 7, 6, 28]. In our study, we have attempted to calculate correlations after collapsing OTUs to the genus and family levels. Despite finding significant correlations between the taxa at those levels, we found little interesting patterns in the network. Subsequently we also found that there is a lack of discussion on whether building networks at higher taxonomic levels was valid or would reveal patterns unobservable otherwise. Thus we follow the literature and built our networks at the OTU level.

Calculating correlations at the OTU level can be challenging for practical and statistical reasons. As mentioned in section 2.7.1, the rise of high-throughput sequencing has allowed sequence analysis of large number of OTUs, particularly as many as 63,814 distinct OTUs in our 16S dataset. Pairwise correlations detection necessitates quadratic algorithms, making network construction for such large datasets impractical. Downstream tasks such as module detection and analysis are also impractical. From a statistics perspective, each pairwise correlation is a hypothesis test and thus requires correction for multiple testing. With so many tests, the corrected \( \alpha \) with Bonferroni correction would be effectively 0 and p-values corrected by the less conservative Benjamini-Hochberg procedure [3] are too large to be rejected.

In this study, we still detect correlations at the OTU level, but to address the computation and statistics concern, we filtered out OTUs that are not abundant and
prevalent enough to have large effects on the communities. To make networks easier to be analyzed visually, we combine nodes with the same taxonomy and similar interactions afterwards. However, combining nodes iteratively without a stopping point can lead to nodes that only have one or two taxonomy levels (in our case, Archaea was one such node in the 16S network) and thus are meaningless for analysis. We therefore only combine nodes up to the Family level.

5.5.2 Importance of Diverse Analyses

In section 5.2.3.3, we used correlation network and NMDS to analyze the effect of Salinity on community dynamics. While the correlation network (figure 5.8) showed that Salinity correlates positively with OTUs that are abundant in the Upwelling Relaxation and Winter Transition periods, the NMDS plot (figure 5.10) showed the opposite. We cannot explain this contradiction, but it shows that analyzing the dataset using different techniques is possible and important. Conversely, we are able to show that temperature correlates positively with the same communities using different techniques, strengthening the analysis.
Figure 5.4: Relative abundance of prokaryotic OTUs. Only OTUs that have relative abundance of at least 5% and is present in at least 13 samples are shown.
Figure 5.5: Hill number of 16S samples
Figure 5.6: Correlation network of 16S taxa during the Peak Upwelling season. Green edges are positive correlations and red edges are negative correlations. Red nodes are top connected taxa and yellow nodes are top abundant taxa.
Figure 5.7: Correlation network of 16S taxa during the Upwelling Relaxation season. Green edges are positive correlations and red edges are negative correlations. Red nodes are top connected taxa and yellow nodes are top abundant taxa.
Figure 5.8: Correlation network of 16S taxa over the entire year. Network is divided into 3 clusters as shown. Green edges are positive correlations within cluster, red edges are negative correlations within cluster, blue edges are positive correlations across clusters, and yellow edges are negative correlations across clusters. Red nodes are top connected taxa and yellow nodes are top abundant taxa.
Figure 5.9: Normalized relative abundance of the taxa that are unique to each cluster in the year-long 16S network. Each taxon’s relative abundance values are normalized by their maximum value.
Figure 5.10: NMDS plot of 16S samples with significant (p < .05) correlations of environmental factors
Figure 5.11: Relative abundance of eukaryotic OTUs. Only OTUs that have relative abundance of at least 5% and is present in at least 13 samples are shown.
Figure 5.12: Relative abundance of diatoms and dinoflagellates
Figure 5.13: Hill number of 18S samples
Figure 5.14: Correlation network of 18S taxa over the entire year. Network is divided into 2 clusters as shown. Green edges are positive correlations within cluster, red edges are negative correlations within cluster, blue edges are positive correlations across clusters, and yellow edges are negative correlations across clusters. Red nodes are top connected taxa and yellow nodes are top abundant taxa.
Figure 5.15: NMDS plot of 18S samples with significant (p < .05) correlations of environmental factors
Figure 5.16: Cross-domain correlations over the entire year. Nodes are grouped by their 5th level taxonomy. Red nodes are top connected 16S taxa, yellow nodes are top abundant 16S taxa, blue nodes are top connected 18S taxa, green nodes are top abundant 18S taxa, and violet are top connected and abundant 18S taxa.
BIBLIOGRAPHY


[20] Lima-Mendez, Gipsi and Faust, Karoline and Henry, Nicolas and Decelle, Johan and Colin, Sébastien and Carcillo, Fabrizio and Chaffron, Samuel and Ignacio-Espinosa, J. Cesar and Roux, Simon and Vincent, Flora and Bittner, Lucie and Darzi, Youssef and Wang, Jun and Audic, Stéphane and Berline, Léo and Bontempi, Gianluca and Cabello, Ana M. and Coppola, Laurent and Cornejo-Castillo, Francisco M. and d’Ovidio, Francesco and De Meester, Luc and Ferrera, Isabel and Garet-Delmas, Marie-José and Guidi, Lionel and Lara, Elena and Pesant, Stéphane and Royo-Llonch, Marta and Salazar, Guillem and Sánchez, Pablo and Sebastian, Marta and Souffreau, Caroline and Dimier, Céline and Picheral, Marc and Searson, Sarah and Kandels-Lewis, Stefanie and , and Gorsky, Gabriel and Not, Fabrice and Ogata, Hiroyuki and Speich, Sabrina and Stemmann, Lars and Weissenbach, Jean and Wincker, Patrick and Acinas, Silvia G. and Sunagawa, Shinichi and Bork, Peer and Sullivan, Matthew B. and Karsenti, Eric and Bowler, Chris and de Vargas, Colomban and Raes, Jeroen.


