ERSP Proposal

Sara Safa, Shayal Singh, Daniella Vo

Fall 2018

1 Research Context and Problem Statement

It is well known that microbiota, a community of bacteria, viruses, and other microorganisms, is responsible for different functions in many organisms and the environment. They have been shown to be critical for the health of many species, especially humans [4], but how much is actually known about these microorganisms? Studies have also shown the connection between certain microbes and diseases, which emphasizes the importance of studying microbes [1]. If there was a general rule that scientists know microbes would follow, it might make the process of identifying diseases easier. For example, if scientists know two microbes are related, they might be able to pinpoint diseased microbes. The number of possibilities for the use of microbes is endless, but these types of medical advancements are only feasible if relationships can be found between these microorganisms.

With all the publicly available data on microbiome, various studies have been conducted to discover more about these microbiota that inhabit most organisms, but a lot of the data is not being analyzed to the full extent. Currently, researchers use Pearson's correlation coefficient to determine whether there is a relationship[8]. This may work for certain data sets where a positive or negative linear relationship can be clearly distinguished, but not for non-linear graphs. In some cases, past studies disregard this dispersed data by saying there is not a relationship that can be determined, resulting in an analysis that is not thorough. There may be other types of relationships that cannot be identified through the methods of linear analysis currently in use.

We propose using Boolean analysis, a logical method to determine if a certain variable is dependent on another (if-then relationship), to study large amounts of microbiome data, comprehensively. Our research will be using Sahoo's methods on analyzing Boolean implications [6], which was used on gene expressions from human, mice, and fruit fly samples. Since his method of Boolean analysis has not previously been used on microbiome data, we will test if it is also an effective method to analyze microbe-microbe relationships.

There have been studies conducted using Boolean analysis to find the connection between oral microbiome and HIV-associated periodontists [5] and to

find a metabolic network of interactions in gut microbiome [7]. Although these studies focus on relationships within a certain microbiota community, specifically oral or gut microbiota, we are looking to use Boolean analysis to uncover universal rules between pairs of microbes that are applicable to any microbe community. For example, if we know there is a reoccurring Boolean relationship between two types of microbes, we will expect this relationship between these microbes to hold in any microbiome community, which makes this relationship a universal rule. Past studies lack universality and cannot, therefore, apply to the microbiomes of all living things.

While previous research resulted in correlations between microbiota communities, our project's goal is to formulate Boolean relationships that are more comprehensive, universal, and can be applied to every microbiome. No previous studies have focused on finding the fundamental rules between microbes in different biological systems. Our goal is to take advantage of all the publicly available microbiome data that we can use to determine Boolean implications between data samples. Then through Boolean analysis, we will try to find universal connections between the different microbe communities.

2 Proposed Solution

Our research will focus on collecting and analyzing publicly available microbiome data, which will hopefully be the most comprehensive study on this type of data. The first step in this process will be to obtain the microbe data, specifically bacterial 16S rRNA sequences in the form of FASTA and FASTQ files. Sources of these data sets will include samples from the Human Microbiome Project(HMP), Earth Microbiome Project(EMP), Genbank, and the European Bioinformatics Institution.

In order to process the thousands of raw data sequences available to us, we will be using command-line, data-processing applications such as QIIME, which is used for its visualization properties. QIIME will take in raw sequences and output the counts of each microbe in a frequency table. This will categorize the data into samples, features, and counts, where the counts represent how many samples contain a specific feature, and each feature represents a microbe. We will run these feature tables through a script that will normalize the data, which is important in order to scale the counts and make fair comparisons if the counts are higher for a certain sequence than another. Based on the frequency tables, the script will generate scatter plots on the Hegemon website, which is an existing web-based tool developed to visualize scatter plots [2].

In order to find Boolean relationships from the scatter plots, the StepMiner algorithm will be used to determine a threshold used to categorize a microbe's frequency as either high or low. If the counts fall below the generated threshold, there is a low frequency, and vice versa for high frequency. The StepMiner algorithm generates a threshold for each axis (i.e variable) of a scatter plot graph, which is used to divide the graph into four quadrants. Based on these

frequency levels, each data sample will be plotted in order to visualize the Boolean implication relationship between two different microbes.

A Boolean implication is a relationship between two variables. Each variable is placed on an axis. Boolean relationships have 2 general types: symmetric and asymmetric. If data points are concentrated in exactly two opposite quadrants, a symmetric (also known as linearly opposite or equivalent) relationship exists. If data points are concentrated in exactly three of the quadrants, an asymmetric relationship exists, which are the types of relationships we will be focusing on.

Consider the example where the data is concentrated in all the quadrants except the upper right quadrant as shown in Figure 1 below. We can label the horizontal axis as microbe species A and the vertical axis as microbe species B. The relationships in the graph is if A is high, then B is low for the following reasons: we know from the sparse upper right quadrant that microbe pairs will never exist when both A is high and B is high. So, the only possible relationship when A is high is if B is low. Therefore, the Boolean implication for this graph is if A high ⇨ B low.

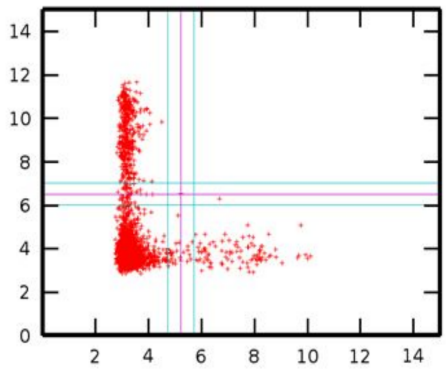

Figure 1: High ⇨ low relationship presented in [6].

From these plotted paired relationships, we will conduct Boolean analysis on all data samples to categorize these relationships as symmetric (opposite and equivalent) or asymmetric (low/low, low/high, high/low, high/high) using the previously stated method for identifying Boolean relationships [6] .

Using the discovered relationships, we will be able to compare relationships between microbe communities and identify trends that apply to these communities for all organisms. This method of Boolean analysis is comprehensive, and will provide more insight into the mostly unstudied relationships between the microbe communities.

In order to ensure our approach is successful, a series of trials will be conducted to determine which microbe communities display Boolean relationships.

Every microbe pair plotted may or may not show a meaningful, logical relationship, hence such testing and experimentation is vital. To ensure the relationships we discover are universal, we will look for Boolean relationships that are common across multiple microbe communities from different data sets. If we find a relationship that holds between all the communities we studied, then we can deduce that a fundamental rule is present across all microbe communities.

3 Evaluation and Implementation Plan

3.1 Analysis Plan

In order to validate each universal rule, we will keep updating the data with additional samples containing the microbes of interest. The new data samples will undergo Boolean analysis, which will be used to determine if the rules still hold. The more samples that follow each rule, the more confidence we have in the validity of the rule. On the other hand, if we establish a rule, and a future data set violates this rule, we can discard that rule because it is no longer universal.

To further confirm the universality of our rules, we plan to generate our own data in a wet lab by collecting samples. By repeating the Boolean analysis process, we will test if the samples follow the universal rules we found. If our samples also follow these rules, it further validates the universal rules we have previously established.

To our knowledge, this is the first study that looks for general rules between microbe communities. Although the rules will be universal for the data sets we studied, these data sets are not exhaustive because there may be some data sets unavailable to us, which may disprove our rules. Therefore, future studies are needed to verify our results. With further research, there will be a higher certainty that those microbes in any given community will follow such rules, which has potential applications in an array of different fields such as medicine and epidemiology.

3.2 Timeline

Fall Quarter 2018
Week 7:

- Finish part 3 of Project Proposal- Implementation and Evaluation
- Run one data sample through QIIME 2
- Refine project proposal (Writing Hub)

Week 8:

- Fine tune proposal

- Download all aligned data

Week 9:

- Perfect proposal-Due next week
- Run aligned samples through QIIME 2

Week 10:

- Finish Research Proposal - Due
- Work on presentation

Winter Quarter 2019:
Weeks 1-2

- Organize data into tables/matrices(feature tables) using QIIME 2
- Categorize data into samples, features, and counts

Weeks 3-4

- Normalize data
- Scale data to make fair comparisons

Weeks 5-6

- Find proper analyzing software to analyze tables
- Analyze tables for possible relationships

Weeks 7-8

- Plot microbe pairs to see if a relationship exists
- Conduct multiple trials with different gene pairs

Weeks 8-10

- Get familiarized with Professor Sahoo's tools for generating Boolean implication graphs
- Make Boolean implication graphs for relationships that were found

Spring Quarter 2019:

- Compare these results to others
- Find similarities and trends to find rules that are universal throughout all microbiome communities
- Make poster for presentation

References

[1] Clemente, Jose C and Ursell, Luke K and Parfrey, Laura Wegener and Knight, Rob. *The impact of the gut microbiota on human health: an integrative view.* Cell, 148(6), 2012.

[2] Dalerba, Piero and Kalisky, Tomer and Sahoo, Debashis and Rajendran, Pradeep S and Rothenberg, Michael E and Leyrat, Anne A and Sim, Sopheak and Okamoto, Jennifer and Johnston, Darius M and Qian, Dalong and others. *Single-cell dissection of transcriptional heterogeneity in human colon tumors.* Nature Publishing Group, 29, 2011.

[3] Hall, Andrew Brantley and Tolonen, Andrew C and Xavier, Ramnik J. *Human genetic variation and the gut microbiome in disease.* Nature Publishing Group, 11, 2017.

[4] Niv Zmora, Jotham Suez, and Eran Elinav. *You are what you eat: diet, health and the gut microbiota.* Nature Reviews Gastroenterology & Hepatology, 1, 2018.

[5] Noguera-Julian, Marc and Guillén, Yolanda and Peterson, Jessica and Reznik, David and Harris, Erica V and Joseph, Sandeep J and Rivera, Javier and Kannanganat, Sunil and Amara, Rama and Nguyen, Minh Ly and others. *Oral microbiome in HIV-associated periodontitis.* Wolters Kluwer Health, 12, 2017.

[6] Sahoo, Debashis and Dill, David L and Gentles, Andrew J and Tibshirani, Robert and Plevritis, Sylvia K. *Boolean implication networks derived from large scale, whole genome microarray datasets.* BioMed Central, 10, 2008.

[7] Steinway, Steven N and Biggs, Matthew B and Loughran Jr, Thomas P and Papin, Jason A and Albert, Reka. *Inference of network dynamics and metabolic interactions in the gut microbiome.* Public Library of Science, 6, 2015.

[8] Zuber, Verena and Strimmer, Korbinian. *Gene ranking and biomarker discovery under correlation.* Bioinformatics, 25(20), 2009.

4 Changes Made In This Revision 12/7

- Moved real-world example from Evaluation Section to Research Context/Problem Section
- Fixed statement about how only 2 out of 6 relationships have been studied from the Problem Section
- Elaborated on universality in the Problem and Solution sections
- Explained why data is not being used to the full extent in the Problem Section
- Introduced Sahoo's method in Problem Section, and explained how we will use his method on the microbiome
- Explained QIIME process more in the Solution Section
- Made a smoother transition for normalizing data in the Problem Section
- Reworded explanation of Boolean implications in the Problem Section
- Added a diagram to help explain Boolean relationships in the Solution section
- Elaborated on how a rule can be validated in the Evaluation Plan, and explained that any rules we found are only universal for the data sets we found, but other data might contradict our results
- Added how we can use wet lab to validate our rules in the Evaluation Section