Spring 2013

Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data

Kumar Thurimella
University of Colorado Boulder

Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data

by

K. Kumar Thurimella

A thesis submitted to the

University of Colorado in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Department of Applied Mathematics

Defended April 10th, 2013

Dr. Rob Knight, Thesis Advisor

Department of Chemistry and Biochemistry

Dr. Michael Mozer

Department of Computer Science

Dr. Anne Dougherty

Department of Applied Mathematics

This thesis entitled:
Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data
written by K. Kumar Thurimella
has been approved for the Department of Applied Mathematics

_____

Rob Knight

_____

Professor Michael Mozer

_____

Senior Instructor Anne Dougherty

Date _____

Thurimella, K. Kumar (B.S., Applied Mathematics)

Using Rule Induction to Elucidate Co-Occurrence Patterns in Microbial Data

Thesis directed by Professor Rob Knight

Hundreds of studies have addressed whether the presence or absence of certain bacteria are linked with a particular phenotype. However, it is plausible that the causative agent (or the consequence) of a given phenotype is not a single type of microbe, but groups of them, perhaps in specific combinations. Rule Induction is a commonly used machine learning method to infer structure within observational data, and build rules to represent these structures. In this thesis I introduce the application of a method, Rule Induction, to infer co-occurrence patterns in microbial data.

First, I benchmark the methods within Rule Induction, to assess how rules are generated with regards to several parameters such as table density, support and confidence. I then subsample data over multiple iterations to understand the robustness of the rules being produced to verify due to sampling.

Next, I provide insight into different biological variables and examine their effect on rules produced. I compare 16S rRNA region, specifically V1-3 and V3-5 regions. I compare different sequencing technology, specifically 454 and Illumina. I finally compare time, specifically looking over a time frame of 400 days. Within all these comparisons I aim to understand the differences, but more importantly what is conserved when these samples are stratified by these variables in terms of the generated rules.

Finally, I explore Rule Induction using two microbial datasets, and compare the rules to already-known associations. The first dataset I interpret identifies a correlation between HIV and the Gut Microbiome. The second dataset distinguishes the Gut Microbiome over varying geographical locations. I link each of these rules

produced from each dataset with taxonomic information and consolidate those rules to give rise to the underlying structure within the biological data.

Dedication

To my family and friends.

Acknowledgements

First and foremost, I would like to thank Rob Knight for letting me be a part of his lab. He has been a fantastic mentor and someone who I look up to very much. His passion for research is contagious, and I have learned so much in my time here. I am very fascinated and intrigued by the research being done in this lab in addition to the microbiome field at large. I would like to thank Mike Mozer for being a great mentor as well, by giving me great advice specific to this project as well as general life advice. My final committee member, Anne Dougherty, has been nothing short of a phenomenal advisor. It is after talking to her my sophomore year that I switched my major to Applied Mathematics. She has consistently provided great support and I can go to her about anything like a friend.

I would like to thank Jose Clemente for his ideas, support and fabulous mentoring. I have learned a lot from Jose in terms of his perspective on being a researcher. Will Van Trueren has been nothing but a great friend and peer mentor in the lab and has always provided great insights to this research. I want to finally thank other members of this lab including: Yoshiki Baeza for his help understanding cluster computing, Cathy Lozupone for her HIV data and insight within co-occurrence and many others who were always there for support. I want to acknowledge the Gautam Dantas Lab at Washington University in St. Louis, specifically Kevin Forsberg and Mitch Pesesky. This past summer opened my eyes to exciting avenues of research, and that experience provided the direction that has guided my research to this day.

Without my close friends I wouldn't be where I am at today. Thanks to my roommates David Gillis and Thomas Lynn for helping me out this year. Thanks to Oriel Eisner for always being interested in (or putting up with) our late night talks, mostly regarding science. Many thanks to Sathish Subramanian for being a fantastic role model and providing help and support whenever I needed it, as well as late night life talks. I hope to be half the MD/PhD student he is, one day. Vivek Verma has been a great role model as well and made my time in St. Louis very memorable. Will Timbers has always been incredibly fun to talk with and has pushed me to pursue my passions. Andrew Fleming has shared a same passion for science since high school, and I appreciate all of his support over the years. Without Myke Samuels I don't know where I would be, and it was because of his help that I found my passion and interests. For that I am forever grateful.

Finally thanks to my brother and parents. Coming from a family of computer scientists certainly rubs off on me, and because of all their love and support I have been able to develop my own passions.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Humans are surrounded by microbial communities that live in, on, and around us. The microbiome is the full collection of these microbes, their genomes, and their environmental interactions [37, 27]. Scientists are only now starting to appreciate the complex interactions between microbial communities and larger organisms. These interactions have recently been suspected to play a large role in human nutrition and susceptibility to disease. In particular, recent studies have revealed that microbial communities have been associated with diseases such as diabetes, obesity, and various gastrointestinal diseases. After the huge push to study the microbiome, scientists consider our bodies to be a superorganism, with human and microbial cells working together in a beneficial manner [40].

Although geared towards mathematics and computer science, this thesis has its roots in microbiology. It is not essential to have a biology background to understand many of the concepts within this thesis, but I provide one here to ease the transition and to understand the fundamental, underlying goal.

Microbiology is the study of microscopic organisms [35] that are classified in and around environments. The importance of designating microbiology as a field is within the broad applicability to many areas such as biomedical research, medicine and various environmental applications. For example, let us examine the nitrogen cycle. The nitrogen cycle is critical to all life on earth via plants, and is wholly reliant

on microbes [20]. Nitrogen fixing bacteria, nitrifying bacteria as well as denitrifying bacteria, continually help collect nitrogen from the atmosphere for use in various plants and soil life. This process produces nitrogen by-products into the atmosphere through a well-known cycle. This cycle is critical to agricultural life present on Earth and is only available through various microbial interactions throughout the process.

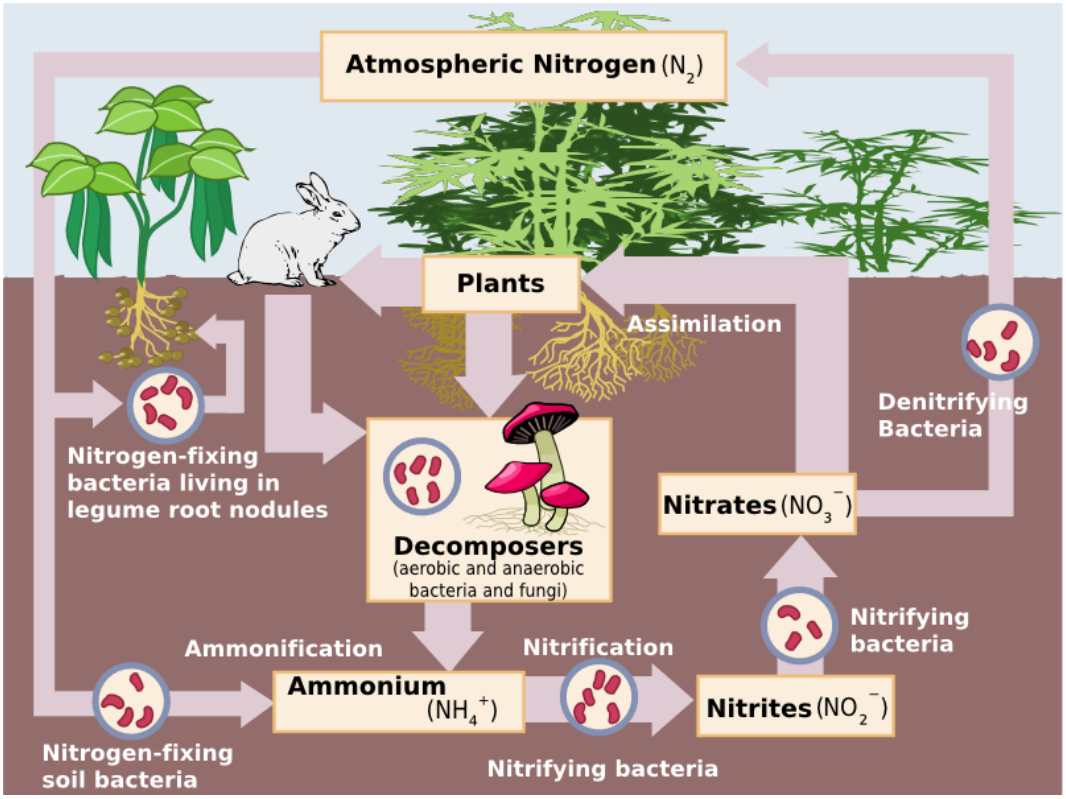

Figure 1.1: This figure depicts the complex interactions within the Nitrogen Cycle. There are various microbes associated with different functions that yield different products. Each product becomes a part of a cycle that is cruicial to growth of any plant form. Each interaction shows the importance of microbes to be able to produce useful chemical products that can be beneficial to life as a whole. Image reproduced from [2].

In the case of human health, microbes also play an important role. Microbial pathogenicity has been suspected for long periods of time. Different pathogenicity factors were originally labeled by using biochemical approaches or through various screening measures. However, with the recent advent of cheaper, more efficient se-

quencing technology we are now able to understand microbes at a whole-genome level. Whole microbial populations might be pathogenic, versus individual virulence factors [15]. With this new technology we can understand microbial communities, not just individual pathogens, which can yield insight into the interconnectedness of the populations.

With more and more information being produced from sequencing, we can better define the microbiome. Recent studies have linked autoimmune disease with the microbiome [17] and have shown a linkage of certain microbes to obesity. Given that the microbiome is very environmentally malleable, there has been a large focus towards understanding microbes and their relation to non-genetic disease. The understanding is that the microbiome is changed in the gut microbiota composition and that is correlated directly with obesity [29].

1.1 Co-Occurrence Relationships within the Microbiome

It is very rare to link a phenotype to just a single microbe [18]; more likely, a group of microbes acting with one another are involved. In several ecological situations the nature of co-occurrence occurs in the forms of symbiosis, commensalism, parasitism, amensalism and synnecrosis. For example, the colonization of Clostridium difficile in the gut is thought to be related to an unbalanced relation between intestinal taxa, which leads to an overabundance of Clostridium difficile [18, 39]. These correlations have been studied within the context of the microbiome, but the full nature of these interactions is unclear. The ability to characterize all these interactions would yield valuable insight into microbial effects on health.

In terms of health and the host-microbiome, the concern is how specific microbes and combinations of microbes consistently different within various hosts. The future of this area is that if groups of microbes are always precursors to a related disease, then there is a need to identify these various groups. The ultimate vision is to screen

diseases based on specific changes in microbial communities. With that knowledge, more advanced therapeutics can be designed and used to alleviate various microbially-related diseases. Using co-occurrence can have far ranging applications to various ecological awareness and as a pre-diagnositic tool to help illustrate disease.

1.2 Introduction to Rule Induction

Rule Induction stems from a field combining machine learning and data mining. The basis of rule induction stems from the idea of being able to find useful co-occurrence patterns in large databases [6]. Being able to sample frequent itemset within a database was quite a complicated task and there was no way to interpret that data [41]. These methods were developed to first look at supermarkets and see the types of products consumers were buying together. For instance, customers would frequently buy tomato, onion and burger buns. Using Rule Induction, a predictive association can be made claiming that if one buys tomato, onion and burger buns we predict they also buy hamburgers. These types of associations are mined and using the predictive behavior we can provide a thorough analysis to market baskets. From those mined associations we have information on the types of products people buy frequently together [9].

Due to the high dimensionality of the data being mined many steps must take place before creating such associations [4].In dealing with this type of data it is useful to classify the data initially. In terms of market basket analysis, data is normally grouped into transactions and items. Every transaction contains a certain number of items that were bought. Using several measure to further classify and create associations, the terms of support and confidence were introduced [5]. Support directly correlates to the ratio of which an item(s) was observed in all transactions and the confidence is how to define our predictive behavior. A more detailed explanation is provided in Chapter 2. With this information in had one can use these associa-

tions and discover frequent itemsets using two different algorithms, the Apriori and Eclat algorithms [3, 36]. The Apriori is a depth search algorithm while the Eclat is a breadth search algorithm. Each of these algorithms have their advantages to different types of data and work efficiently to mine associations and build predictive rules. For the purposes of this work we will only focus on the Apriori algorithm, which is further detailed in Appendix A and Chapter 2.

Figure 1.2: This figure depicts the initial use of Rule Induction. Many of the associations built were used for product placement and understanding of market basket data. For example, a rule from this picture is if someone buys bell peppers, they also buy cheese. Image reproduced from [1].

1.3 Novelty of Rule Induction

Various methods have been applied to grasping co-occurrence and many have worked with great success. Previous methods include using estimated linear Pearson correlations between microbe components after a log-transform, generalized boosted linear models combined with multiple similarity measures, and Bray-Curtis distances [19, 18, 32]. Many of these methods utilize different similarity measures and account for compositional data. While most other methods predict co-occurring microbes, a

large majority of them are limited to looking at pairwise co-occurrence [32]. The novelty within Rule Induction, however, is the idea of building co-occurrence patterns to a very high order. Rule Induction build rules that range from pairwise co-occurrence to co-occurences of a higher order. With the aid of this method, we are able to identify clusters and understand co-occurrence at a broader level.

Chapter 2

Understanding Rule Induction

Being able to determine different co-occurrence structures of high order (i.e. beyond pairwise interactions) is crucial in understanding an intricate network amongst different microbes. Many approaches have been taken to understand this complex problem. Here we present a novel approach to microbial co-occurrence, Rule Induction, and explain the underpinnings of Rule Induction.

2.1 Rule Induction

Given a microbial dataset, observances are defined as the number of OTU's (Operational Taxonomic Units) present in a given sample. We label an OTU as a group of microbes where, within each group, all 16S rRNA gene sequences are within some threshold of pairwise sequence identitiy. This is a way that we can classify and label microbes in a given dataset. Microbial datasets are defined as OTU Tables where various OTU compositions are detailed per sample in a study.

Rule Induction is defined as a way to extract formal rules from a set of observations. More generally speaking, let set $O = \{O_1, O_2, \ldots, O_n\}$ be defined as all the OTU's present in a dataset and let $S = \{S_1, S_2, \ldots, S_n\}$ represent all the samples that are present in the study dataset. Each sample is unique and contains some subset of

items from set O. We can build rules from this set and define them mathematically:

$$A, B \subseteq O \ \textit{where} \ A \cap B = \emptyset \ \textit{with} \ |B| = 1 \ \textit{and} \ A \rightarrow B$$

In other words A and B are *unique* subsets of our OTU's present where B can only be one OTU. The goal is to build rules where A implies B [7, 5].

In order to select and mine rules of value, we must employ several interest measures to validate our findings. We only consider 3 of the most popular interest measures when building and consolidating those rules [11]. The first is support, $supp(A)$, which is defined as the proportion of samples that contain that given set of OTU's. The next is confidence, $conf(A \rightarrow B)$, which can be inferred as the probability, $P(B|A)$ of finding OTU set B within the sample under the conditions that OTU set A is also contained in that sample. Mathematically speaking these probabilities are carried out under

$$conf(A \rightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

Finally our last interest measure is defined as lift, which measures how many more times A **and** B occur together than if they were statistically independent. This is represented as:

$$lift(A \rightarrow B) = \frac{supp(A \cup B)}{supp(A) * supp(B)} = \frac{conf(A \rightarrow B)}{supp(B)} = \frac{conf(B \rightarrow A)}{supp(A)}$$

Rules are then built by utilizing different threshold values for support and confidence. With the support and confidence in hand, the next task is to collect *frequent* OTU sets using the given support. When finding the combinations of every possible OTU set, assuming that $\{|A| > 0 | A \subseteq O\}$, the possible OTU sets can be very large, more precisely $2^n - 1$ where n is the size of O. Many different approaches can be

taken, but with regards to this thesis the Apriori algorithm was used. The motivation behind the use of the Apriori algorithm (pseudocode provided in Appendix A) is that it takes advantage of the Downward Closure Lemma, as formally defined in Appendix A. The idea being that for any given support of an OTU set with cardinality n doesn't meet the support threshold set, then no combinations of cardinality greater than n can be made and will thus be pruned out. After mining all frequent itemsets of the model, rules are then built based on the confidence defined, and associations are then created.

Rule Induction employs the use of set theory and builds associations, based on various statistical measures, which provide insight into co-occurrence patterns within microbial datasets.

2.2 Benchmarking Rule Induction

Throughout this experiment we make use of several packages within the R and Python and related libraries. The decision to use these packages was primarily based on the open-source nature of the tools. The implementation of Rule Induction was carried out through an R package *arules* [23]. Using the built-in fucntion *apriori* we carried out several benchmarks to test the robustness of the methods. This package only considers presence and absence in a given dataset regardless of the abundance.

To begin we created several synthetic datasets without any prior assumptions about the correlations set. We started with 20 samples with a total composition of 100 OTU's (20 by 100 OTU Table). We then created subsets of that OTU table with a varying density between 0.5 - 0.9, with a step size of 0.01. For example with a 50% dense table this means that of all 2000 elements in the matrix, 1000 of those elements are non-zero (present for our binary purposes). Note that because only presence or absence is taken into account, the table counts were kept at either a 1 (signifying presence) or 0 (signifying absence). From there each of those tables at varying densities

had each of their corresponding elements shuffled, at 50 iterations. Note that when we define shuffling we generate all possible permutations of the elements in each row and select at random one of those permutations using a Pseudorandom number generator (20! permutations in our case). After each of those fifty tables was generated for each density between 0.5-0.9 (step size 0.01) the number of rules being output was collected for each synthetic table at a confidence and support set at 0.9. Figure 2.2 below summarizes those results.

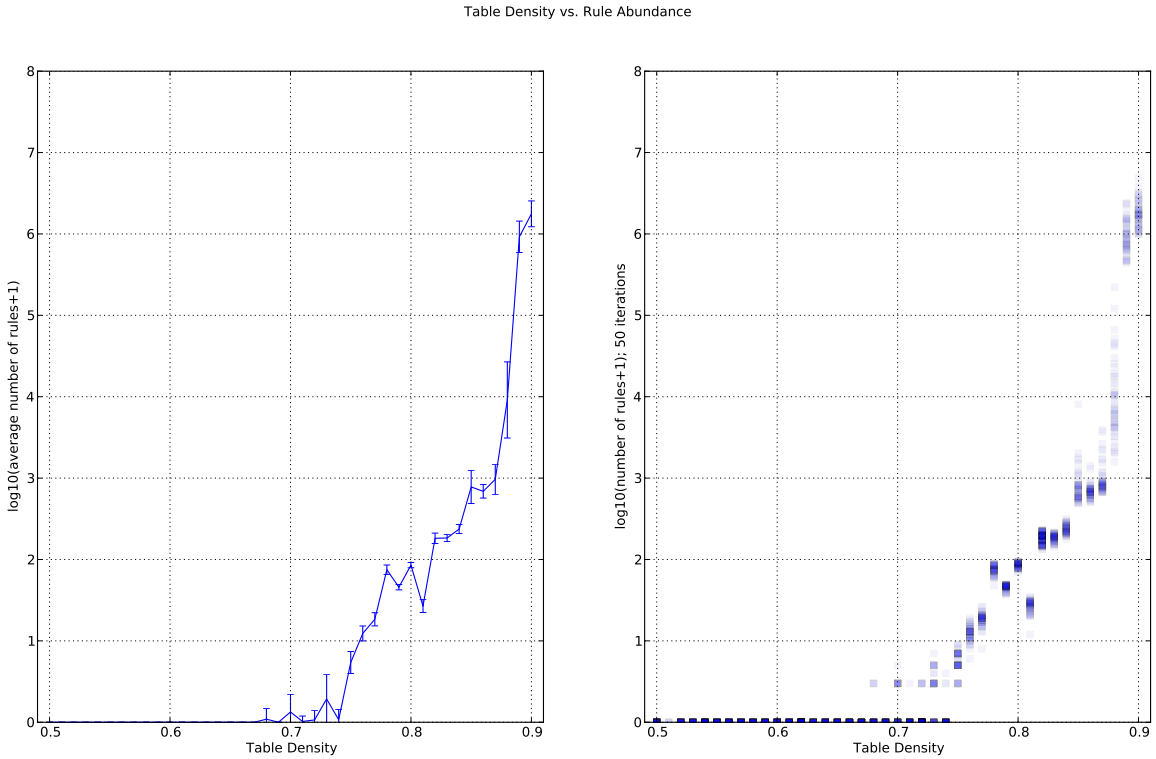

Figure 2.1: Each density was kept at a constant support, and confidence at 0.9, that enabled a thorough examination of the number of rules being produced from *arules*. Each error bar, the standard error, is included to demonstrate the varying number of rules at a given density. The left panel describes the average number of rules produced from the 50 different iterations. The right panel differs in illustrating the entirety of the number of rules per density.

As the general trend suggests, the higher the density of the table the larger the

number of rules presented at each iteration. Further, the higher the table density, the larger the number of associations to be mined by Rule Induction.

The support and confidence were set at a constant level due to memory limitations faced even by a computer cluster. Throughout the entire benchmarking process as well as the duration of the thesis, memory usage was of concern. Even while using the lab cluster machine, the package *arules* still had instances where it would run out of memory even after allocating the highest amount of allotted memory, which was 64 GB of memory. As previously discussed, the number of frequent OTU sets Rule Induction can find is exponential and thus rapidly increases with larger datasets.

I then decided to look at a varying support and confidence to see it's effect on the number of rules presented. Again, due to memory limitations, the density range was kept between 0.5 and 0.7 with a step size of 0.01. We picked one table from each of those different densities, and calculated the number of rules output based on a varying support and confidence between 0.3-0.9 (step size 0.1). Figures 2.2 and 2.2 summarizes the general trends displayed in this analysis.

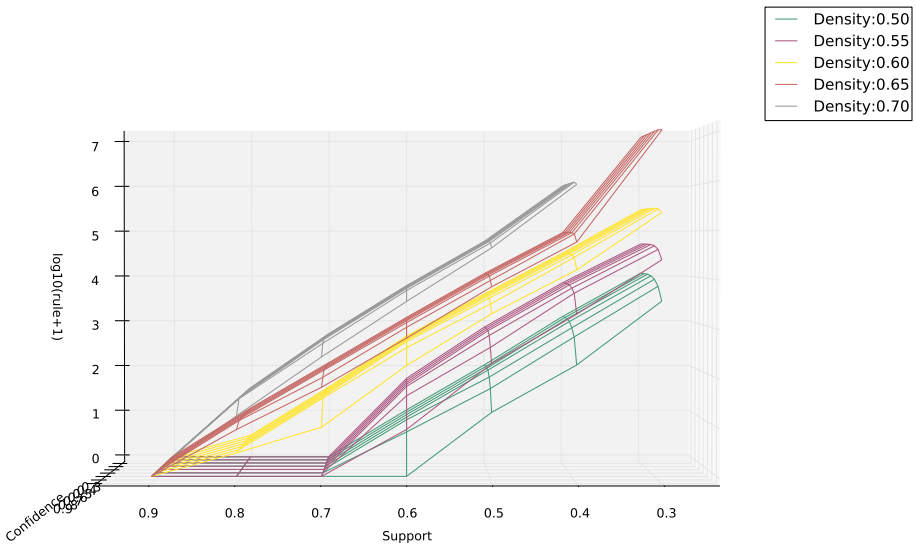

Figure 2.2: Varying the support, confidence and density provides some meaningful insight. This validates the notion of the downward closure lemma, that lower support allows more associations, because many OTU sets don't get immediately pruned out. We can notice here that support drastically changes the number of rules being produced. Even within a log scaling the number of rules goes up a magnitude of 4, which clearly illustrates the effect of rules being produced based on a support. This further validates the use of support when building associations.

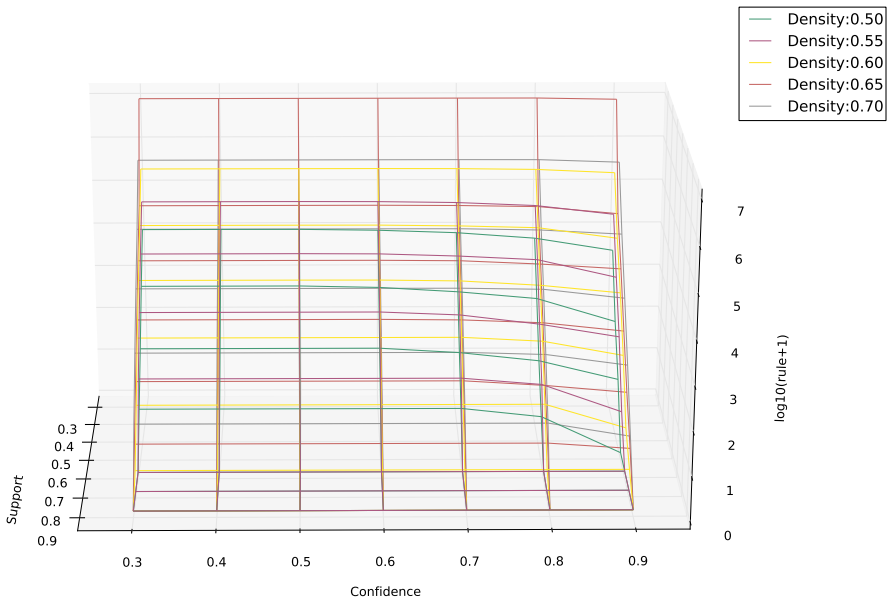

Figure 2.3: Varying the support, confidence and density provides some meaningful insight. In this example we see that the number of rules is only slightly changed with the varying confidence. The only drastic change is between 0.8 and 0.9. This exemplifies the effect of confidence on number of associations built is not very dependant on confidence as a metric.

2.3 The Effect of Subsampling on Rules Produced

Another avenue of interest for understanding Rule Induction, in the context of arules, is subsampling. My methods were to take a single synthetically generated table and do 100 random subsamplings on that table, at the same depth, to create 100 distinct variations of the original table. We chose to do 100 random subsamplings and not anything higher due to computational constraints. We felt that 100 subsamplings would effectively yield good subsamplings in our 20 by 100 subsampling space. We then took the top 100 rules, sorted based on lift measure, from each of the 100 different tables. Within there we checked to see how many of the rules in each subsampling were reproduced in any of the other subsampled tables. In a completely robust system the top rules generated from each subsampling would always be the same. We define robust in terms of Rule Induction as being able to produce rules that are identical regardless of any subsampling, because Rule Induction would be able to find the strongest structure within the data regardless of other less abundant OTU's being present or absent. See Figure 2.3 to look at the distribution of rules throughout each subsampling.

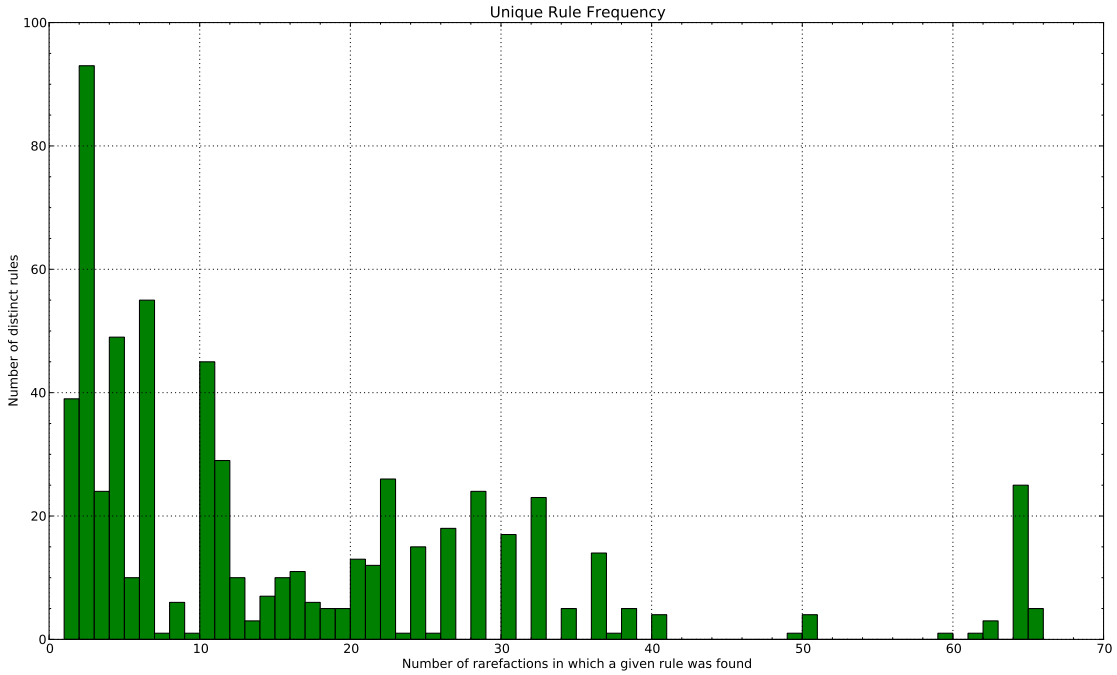

Figure 2.4: In this graph we note the number of distinct rules found in each of our subsamplings. at the far left we have rules which were found in only a single subsampled table. For example on the far right of the graph, we have 5 distinct rules which were found in 65 of the subsampled tables.

Chapter 3

Variability Amongst Rules

We now transition into rules and distinguishing variables within the context of the microbiome. There are many approaches in understanding the microbiome and these differences can play a very important role into the analysis of the microbiome. The effects of these different variables are of interest to be able to understand the differences within OTU structure, but more importantly to see what is constant regardless of the variable. The issue of technical bias, as presented in Technology and Region is of interest and seeing how robust the rules are to this bias can further validate those rules.

To illustrate further, we used a Jaccard Index of comparison between the rules based on the varying factors. We define Jaccard Index [38] to be a variation of the well established Jaccard coefficient with sets A and B, namely [28]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

For our purposes in order to compare rules where the composition of two rules may look like $R_1 = \{R_1^l \rightarrow R_1^r\}$ and $R_2 = \{R_2^l \rightarrow R_2^r\}$. We define the Jaccard Index of similarity to be

$$J(R_1,R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|} \ \textit{with} \ R_1 = R_1^l \cap R_1^r \neq \emptyset \ \textit{and} \ R_2 = R_2^l \cap R_2^r \neq \emptyset$$

This was applied to different rules and was used to tell the similarity/dissimilarity to the compared rules.

3.1 Rules Conserved Over Technology

With the recent breakthrough in Next-Generation Sequencing we are now able to gather large amounts of data from microbial communities. Two sequencing technologies, Roche 454 and Illumina, are widely used in many research contexts and have provided the ability to understand many organisms on a genomic level [34]. Our hypothesis is that we will observe substantial overlap between rules when viewing the same samples with different sequence technologies, because each technology will provide a comparable view of the community sampled [33]. We do recognize that each sequencing protocol is substantially different and the depth of reads are very different as well. The studies that were used in this analysis:

(1) Costello, Elizabeth K., et al. [16] provides insight on the biogeography on the human body using 454 sequencing.

(2) Song, Sejin., et al. provides insight on the biogeography on the human body using Illumina sequencing.

Each of the respective studies have taken samples from different body sites and aimed to establish a healthy criteria for the human body. Study 1, however, uses 454 sequencing technology and study 2 uses Illumina sequencing technology. The OTU table was split on a Body Site category. There were three OTU tables describing the various body sites: feces, oral cavity and skin (namely the hands), which were the same between the two studies. Rule Induction was used on each study and body site with the same support and confidence (0.1 and 0.9 respectively) to generate rules. The top 500 rules sorted by lift were taken and then compared using the Jaccard Index

Jaccard Indices under Technology and Body Site

	Jaccard Index					
	0.0	0.2	0.25	0.33	0.5	1.0
Feces	124936	314	0	0	0	0
Oral Cavity	125020	230	0	0	0	0
Skin	124848	381	21	0	0	0
Total Number of Observances = 125250						

as defined above. The tables in 3.1 describe each Jaccard index that was computed throughout the body sites at each varying technology. The rules output from the two different studies were placed in a 500 by 500 matrix where each index (i, j) refers to a that rule number from each respective dataset. A matrix was computed and only the Jaccard Indices upper triangle along with the diagonal so as not to double count due to the symmetric nature of the Jaccard Matrix.

The results clearly show no overlap and any overlap provided by the Jaccard Matrix doesn't give any true insight. We believe that this is due to the fact that each community that was produced came from two different studies and thus didn't accurately represent the initial community. We need to check that the underlying OTU's are found across the samples when each OTU table is generated.

3.2 Rules Conserved over Region

Much of the diversity stemming from within a microbial community can be characterized by highly conserved nature of the 16S rRNA [14]. 16S rRNA is very useful for identifying microbes due to the fact that it is highly conserved amongst different types of microbes. There are 9 hypervariable regions associated with a 16S rRNA, and each of those regions is said to have a high sequence diversity, which can reveal a different species within various microbes. Our hypothesis is that we will observe a low number of overlapped rules between different regions under the assumption that the primers used to associate to each region pick up different taxa. The studies that were used were used in this analysis:

(1) HMP V1-V3 region, from the QIIME database [12], where the V1-V3 region spanned a majority of the samples. This HMP dataset represents 16S rRNA genes from body sites over human subjects. These are the same samples but sequenced twice.

(2) HMP V3-V5 region, from the QIIME database [12], where the V3-V5 region spanned over all of the samples.This HMP dataset represents 16S rRNA genes from body sites over human subjects. These are the same samples but sequenced twice.

The difference between these two studies is simply the region of the 16S rRNA genes. Specifically the V1-V3 region, which didn't include all of the samples, and the V3-V5 region. Rule Induction was used on each 16S rRNA region with the same support and confidence (0.2 and 0.9 respectively) to generate rules. The top 231 rules sorted by lift were taken and then compared using the Jaccard Index as defined above. Note that there were only 231 rules that were output from each of the datasets. The number of associations wasn't nearly as high, due to the sparsity of the matrix. The

Jaccard Indices under Variable 16S Region

	Jaccard Index						
	0.0	0.2	0.25	0.33	0.5	0.66	1.0
V1-3 vs V3-5 Region	23660	2582	446	34	64	8	2
Total Number of Observances = 26796							

tables in 3.2 describe each Jaccard index that was computed throughout the varying regions. The rules output from the two different studies were placed in a 231 by 231 matrix where each index (i, j) refers to a rule number from each respective dataset. A matrix was computed as shown in Section 3.1.

The results show some overlap and the distribution is more spread than any of the other studies. The results show that regardless of the region sequenced there are still rules that can overlap. We believe that this is due to the fact that even though each region pick up different taxa, there are shared taxa that are discovered in the 16S rRNA gene entirely. This data was not stratified by body site and was taken at whole. An interesting next step would be to understand the rule generation by body site.

3.3 Rules Conserved Over Time

Time series have been of recent interest to many researchers studying the microbiome. Being able to characterize the variations within our microbiome and ourselves on a time scale is very important to see how our the dynamic behavior of our microbiome and the environment. Our hypothesis is that we will see a large overlap of rules where the time points are closely linked, but no overlap over longer time frames. The study that was used were used in this analysis:

(1) Caporaso, J Gregory., et. al. provides the largest human microbiota time series analysis [13], over 396 time points. This study follows two human samples, and looks at 4 different body sites.

This study was analyzed at several different time points and split according to each of those time points. Rule Induction was used on certain time points with the same support and confidence (0.1 and 0.9 respectively) to generate rules. The top 500 rules sorted by lift were taken and then compared using the Jaccard Index as defined before. The tables in 3.3 describe each Jaccard index that was computed throughout the varying regions. The rules output from the two different studies were placed in a 500 by 500 matrix where each index (i, j) refers to a that rule number from each respective dataset. A matrix was computed and only the Jaccard Indices upper triangle along with the diagonal so as not to double count due to the symmetric nature of the Jaccard Matrix.

The results show there was overlap between the initial time point and one day after. There was an overlap between day 10 and all the other days except for day 400. This is the same with day 100 and day 200, 299 but there isn't any similarities to 400. At day 400 there was such a drastic change that no other day resembled the rules within each of those days. The only day that was reasonably similar was day

Jaccard Indices under Time Point Comparisons

	Jaccard Index					
	0.0	0.2	0.25	0.33	0.5	1.0
0 vs 1	113103	11014	0	1	1100	32
0 vs 10	96466	27476	3	1	1304	0
0 vs 100	105065	19736	0	1	448	0
0 vs 200	99181	25213	6	1	849	0
0 vs 299	105920	18673	0	0	657	0
0 vs 400	125123	126	0	1	0	0
1 vs 10	124195	862	138	1	54	0
1 vs 100	124595	585	69	1	0	0
1 vs 200	124653	458	138	1	0	0
1 vs 299	120402	4764	0	0	84	0
1 vs 400	108081	16553	69	1	546	0

Table Continued on the next page

Jaccard Indices under Time Point Comparisons (cont.)

	Jaccard Index					
	0.0	0.2	0.25	0.33	0.5	1.0
10 vs 100	90755	31852	0	2	2605	36
10 vs 200	69789	48701	0	1	6610	149
10 vs 299	84526	34682	0	0	5883	159
10 vs 400	125248	0	0	2	0	0
100 vs 200	78261	42402	0	1	4520	66
100 vs 299	91575	30332	0	0	3289	54
100 vs 400	125249	0	0	1	0	0
200 vs 299	80169	40359	0	0	4722	0
200 vs 400	121817	3431	0	2	0	0
299 vs 400	124982	250	18	1	0	0
Total Number of Observances = 125250						

1. This is likely due to a constraint on the number of samples included in the study, which ultimately lacks enough data for true understanding of the Time series.

Chapter 4

Rule Induction Applied to Microbial Datasets

After understanding and interpreting the rules produced from Rule Induction we now seek to understand how this will apply to biological datasets in the context of accurately predicting those co-occurrence networks. Much of this chapter details the process to go about understanding the nature of Rule Induction in the context of a more broad and general analysis.

We look at two biological datasets with a known structure to see if Rule Induction can pick the OTU structure associated with various metadata. This would validate the Rule Induction approach and increases our confidence in the results obtained. We use the following studies for this purpose:

(1) Lozupone, Catherine., provides insight on the various effects of HIV on the Gut Microbiome. Note this is an unpublished dataset.

(2) Yatsunenko, Tanya., et. al. [43] examines the Gut Microbiome over different regions and at different ages.

Within each of these datasets we explored the idea of integrating metadata categories with rules so we could mine rules linking OTU's with various interesting metadata. These are detailed in each of the following subsections.

4.1 Rarefying, Filtering and Discretizing OTU Tables

One of the challenges faced within the context of general analysis is circumventing the limitation of presence/absence read data. The *arules* package will interpret any OTU read that is non-zero as present, regardless of the abundance. Many issues can arise from this, most importantly the issue of spurious rules. With a table that can have many more associations, one runs a risk of mining rules that provide no insight based on the presence/absence limitation. Another issue to pay attention to is the memory usage and the specific limitation to computing frequent OTU sets if the OTU sets are vastly large with contrived associations.

The first step is to be able to create an even sampling space in which to analyze any data given. Many microbiologists have taken to measures such as rarefaction to overcome this hurdle. Rarefaction compares observed richness among various sites that have been unequally sampled. Rarefaction results in an averaging randomization of the observed accumulation per sample [24, 26].

There are many workarounds to this specific concern including attaching weights within *arules* that correspond to relative abundances, pre-mining the associations using an algorithm that takes into effect the relative abundances, or mining on multiple levels with previous knowledge of large associations [25]. Many of these tactics could have produced meaningful rules, but we used various filtering and discretization measures in order to bypass the matters of memory, false counts all the while still accounting for the abundances.

These methods employ the use of a global filtering method [10] to initially remove any non-contributing OTU's that were found. Using the recommended OTU threshold for a mock community at 0.00005, meaning the fraction of the entire observation count in a given dataset, we perform a global filtering. From there a discretization method is utilized to then convert all the reads to either a 1 or 0 (signifying

presence or absence). In the discretization step, or local filtering, each sample in the study is looked at an individual basis and the microbial composition is analyzed. Based on the total number of observances *per* sample we use different thresholds to convert that OTU in a sample present or absent. Specifically, we use 1%, 0.1% and 0.02% abundance thresholds. For example if we were to use 1% filtering, we would add up the number of observations in a sample, find 1% of that number, and call that our threshold. Then we look at the OTU composition in that given sample and if that OTU abundance is greater than the threshold we label that OTU as present (1) and if it is less than or equal to the threshold we label that OTU as absent (0). This workaround provides a way to account for abundance and filter out low-abundance reads.

We perform this analysis on the HMP V1-3 dataset and analyze the effect of local filtering thresholds on the number of rules produced. Figure 4.1 displays the results of our inquiry. We notice that the number of rules drastically decreases with a stronger threshold, which we think will lead to stronger associations.

Throughout any of the analysis requiring Rule Induction we maintain the same techniques to create a dataset that is preprocessed to efficiently mine rules.

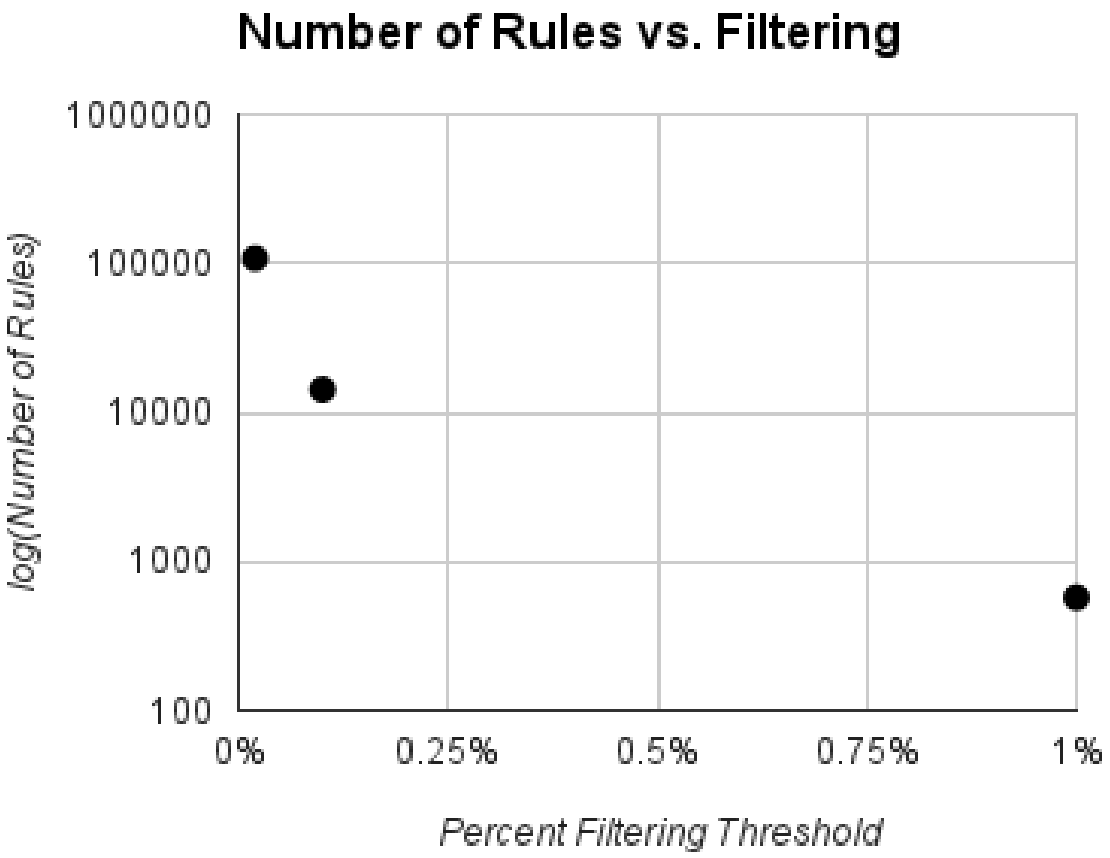

Figure 4.1: The effect of varying local thresholds on the number of rules being produced. Many of the associations can be labeled as spurious without providing insight and is thus shown to decrease with a stronger (1 percent) local filter.

4.2 HIV and the Gut Microbiome

One interesting study presented is the effect of HIV on the Gut Microbiome. The study that was carried out looked at several different factors such as the severity of the disease (Chronic vs Acute). Throughout the analysis done by Lozupone the primary findings showed two distinct categories between people afflicted with HIV and people who were healthy. Using ANOVA correlations with feature importance scores, Lozupone was able to find a strong link between the Prevotella taxa, abundant in the gut with people who had Chronic HIV. The author was then able to also correlate

the Bacteroides taxa abundance in the gut with healthy individuals. Those results are described in 4.2.

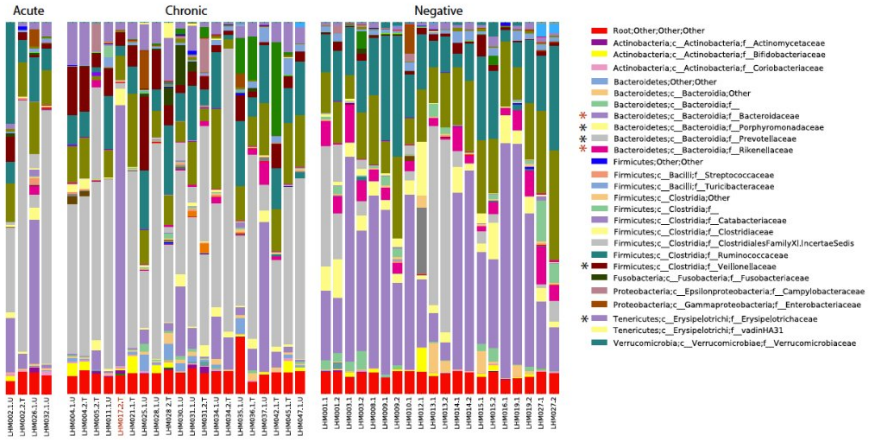

Figure 4.2: Lozupone observed that many healthy samples showed a strong link to Bacteroides, and people infected with HIV showed a link to Prevotella. The taxon strings are provided on the right and the abundances correspond to the color given in the figure. Results provided by Cathy Lozupone.

Using Rule Induction we were able to find a similar correlations with respect to HIV status. In both cases we filtered any mined rules to make sure our metadata was in the antecedent position of the rule structure. In Figure 4.2 we strongly correlated rules that were dominated by Bacteroides with healthy individuals. In Figure 4.2 we found a 2 major rules containing a majority Prevotella to associate with HIV.

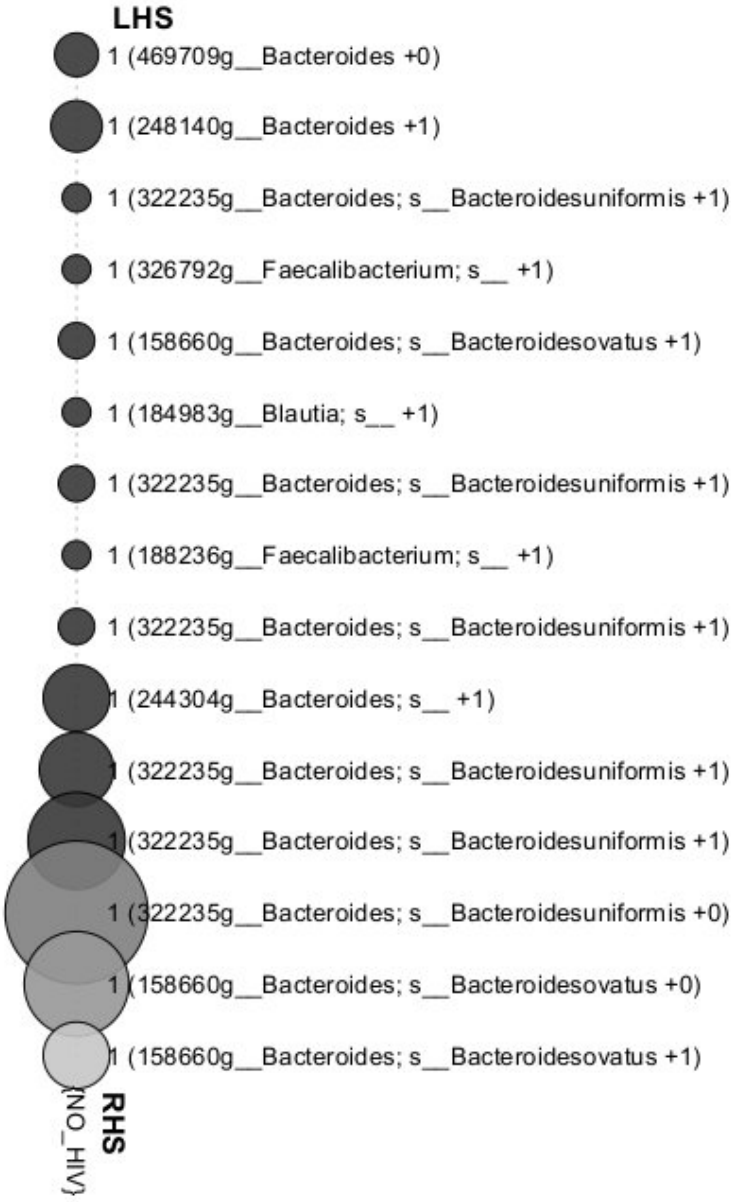

Figure 4.3: Result given by Arules Viz [22] to show Bacteroides clustering yields an healthy individual. The larger the circle the higher the support for that LHS implying the RHS. The darker the circle the higher the lift for that LHS implying the RHS.

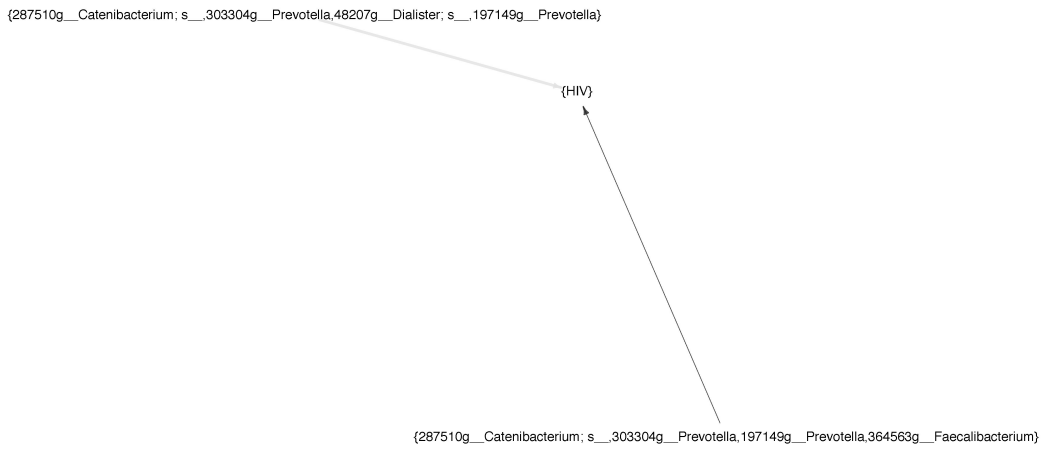

Figure 4.4: Result given by Arules Viz [22] to show Prevotella clustering yields an HIV infected individual. This graph doesn't account for the support and confidence provided in each implication. This simply shows a graph of the rules with the LHS implying the same RHS, HIV.

However, Rule Induction doesn't completely elicit those associations. A problem arises when there are non-discriminatory OTU's present. That OTU will be viewed upon as highly frequent and will be in most associations, as noticed with Faecallbacterium in this study. Rule Induction doesn't prune out for non-discriminatory OTU's and thus loses power in its methods when it comes to evaluation.

4.3 Human Microbiome Viewed Across Geography and Age

This study examines how gut microbiomes are unique across the world specifically looking at the Amazonas of Venezuela, rural Malawi and US metropolitan areas [43]. The study was carried out and looked at how age and location affected humans microbiomes. For our analysis we only decided to look at geography and characterize the differences between gut microbiomes. Using supervised learning methods, namely Random Forest Classifiers, as well as clustering analysis, the author was able to find a strong link between the Prevotella taxon, abundant in the gut with people from

Malawi and Venezuela. The author was then able to also correlate the Bacteroides taxon abundance in the gut with the US population. Those results are described in 4.3. Those results yield great insight into how westernization plays a role in formation of gut microbiomes.

```
Rules showing the presence of Bacteroides in the US population:

   { US,
     186676 " f__Bacteroidaceae", " g__", "s__",
     197072 " f__Bacteroidaceae", " g__Bacteroides", " s__",
     190796 " f__Bacteroidaceae", " g__", " s__"}
                          ====>
     {513445 " f__Bacteroidaceae", " g__Bacteroides", " s__"}


   { 4189999 " f__Bacteroidaceae", " g__Bacteroides", " s__",
     188735 " f__Bacteroidaceae", " g__Bacteroides", " s__",
     2099573 " f__Bacteroidaceae", " g__Bacteroides", " s__"]}
                          ====>
   { US}

Rules showing the presence of Prevotella in the population of
Malawi & American Indians from Venezuela:

   { Malawi,
     515539 " f__Prevotellaceae", " g__Prevotella", " s__copri",
     198502 " f__Prevotellaceae", " g__Prevotella", " s__copri"}
                          ====>
    {197994 " f__Prevotellaceae", " g__Prevotella", " s__copri"}

   { Malawi,
     293717 " f__Prevotellaceae", " g__Prevotella", " s__copri",
     185522 " f__Prevotellaceae", " g__Prevotella", " s__copri"}
                          ====>
    {198502 " f__Prevotellaceae", " g__Prevotella", " s__copri"}

   { AmerIndians,
     328936 " f__Prevotellaceae", " g__Prevotella", " s__copri",
     295554 " f__Prevotellaceae", " g__Prevotella", " s__copri",
     2075910 " f__Prevotellaceae", " g__Prevotella", " s__copri"}
                          ====>
    {289977" f__Prevotellaceae", " g__Prevotella", " s__"}
```

```
{ AmerIndians,
  182123 " f__Prevotellaceae", " g__Prevotella", " s__",
  328936 " f__Prevotellaceae", " g__Prevotella", " s__copri",
  295554 " f__Prevotellaceae", " g__Prevotella", " s__copri",
                        ====>
  {289977 " f__Prevotellaceae", " g__Prevotella", " s__"}
```

Based on a preliminary analysis, OTU tables manually incorporated metadata by taking advantage of the presence/absence feature of *arules*. Further, rules were associated with one another and after searching for rules that incorporated the metadata, we were able to find and confirm the results given in [43]. One shortcoming is the lack of infrastructure to be able to detect and discard non-discriminatory OTU's. For instance in many of the other rules found groups of Firmicutes were associated with each of the populations but they were each distinct groups. Because of those associations some of the discriminatory rules were interspersed among non-discriminatory rules. These rules were hard to sort because they showed the same association score (lift score) because they were of the same order and happened to appear in the same number of samples.

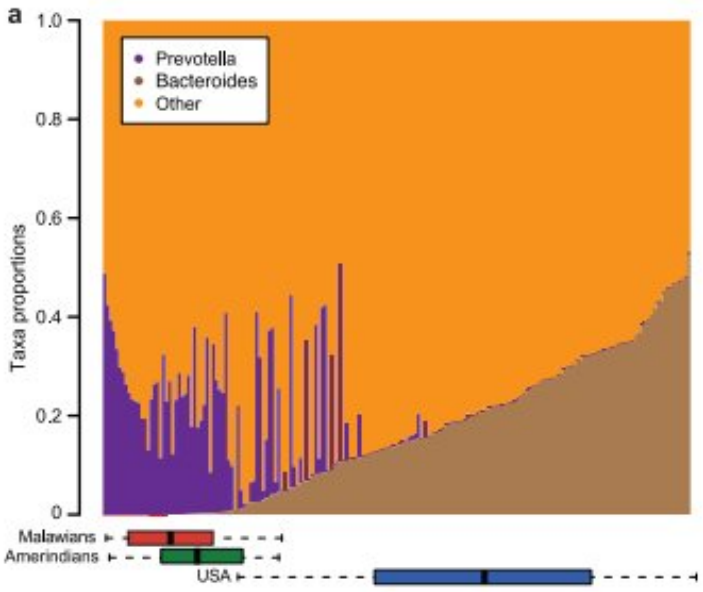

Figure 4.5: Yatsunenko et. al. showed that microbiota differ by within each geographical region. Both the Malawi and Amerindian populations were linked to have an abundance of Prevotella in their gut. The American population was linked to having a higher abundance of Bacteroides in their gut. Reproduced from [43]

Chapter 5

Discussion

In this thesis, we have presented a novel method to analyze and detect co-occurrence patterns within human microbial data. To gain confidence in this technique and results it produced, we applied the method of Rule Induction to well-known data sets where the co-occurrence patterns were independently derived using other methods. Reproducing known results validated our approach and methodology. With regards to this thesis we used the package *arules* from R to aid us in mining rules. The use of R, Python and related libraries was the technology we chose to use for several reasons, the primary one being their popularity within the biology community. The open source nature of these software packages was another important reason that influenced our choice.

Our analysis began with a basic benchmarking of Rule Induction. An approach was taken to test the strength of two required interest measures (support and confidence). We furthered our understanding by adding an extra dimension of variability into our benchmarking, by paying attention to the table density. Our analysis has provided a way to understand how each of these different variables interact with one another, when producing rules. Given a low support and confidence with a very dense table, the number of associations is very large which can lead to artificial rules that don't give any useful insights. This point is very important to highlight since the larger the database of pruned associations the more prone Rule Induction is to min-

ing rules of no value. This benchmarking can pave the path into choosing an optimal support and confidence based on a given dataset by accounting for the density of that dataset.

However, throughout our analysis, we weren't able to produce unique rules over subsampled space to illustrate the notion that subsampling wouldn't affect the associations we mined. The limitation here was the presence and absence factor. If there are OTU's that are sparse they can be present in some subsamples but absent in other and are thus grouped with the highly abundant always occurring OTU's. However, because relative abundance isn't taken into account and each observance is seen in a binary fashion, associations will vary based on what OTU's are present.

Further, this thesis describes an approach taken to understand and interpret different variables in a biological context using the methods of Rule Induction. Within the field of the microbiome sequencing technologies are rapidly advancing themselves and can change within a matter of a few years. Two of the most common technologies are known as 454 and Illumina sequencing. One of the goals of the study was to highlight those structures that are present in those samples and will be found regardless of the technology the sample was sequenced on. The analysis yielded little to no similarity between rules from three different body sites that were sequenced using different technologies. This shortcoming may be due to the fact that each dataset was from a different study, and although there was an overlap in the samples collected, the studies were conducted in different settings and there may be confounding factors resulting in the different OTU structure.

I carried out this same procedure to focus on varying 16S rRNA regions at which researchers can choose to sequence based on a sequence diversity within different regions. 16S rRNA is a highly conserved region among different species of microbes which can assist in identifying various microbes. But within the 16S rRNA there are several different regions that can allow species distinction within microbes. Looking

for a conservation of rules between different regions can yield insight into identification of species that aren't dependant on different regions. Within this comparison there was some overlap amongst the rules. The majority of the rules had no overlap and had a Jaccard Index of 0. There is still reason to believe that regardless of the region some microbes are classified in the same manner and are present regardless.

Finally, our analysis turned to time series and aimed to see what rule order was maintained over time, at different time intervals. Again many of the rules weren't conserved over time but there were certain time points where there was a clear overlap of rules produced. Without any surprise there was overlap between the initial time point and one day after. There was an overlap between day 10 and all the other days except for day 400. This is the same with day 100 and day 200, 299 but there isn't any similarities to 400. At day 400 there was such a drastic change that no other day resembled the rules within each of those days. The only day that was similar was day 1.

In this thesis, we also presented a framework of Rule Induction applied to microbial datasets. Many of the rules mined using the rarefied, filtered and discretized methods were able to have some value and confirmed the insights already provided by each of those studies. It is noteworthy that non-discriminatory OTU's can provide a false rule implication. If an OTU is present in every sample that OTU should not be considered. In terms of the Global Gut dataset we were able to find and verify the classifications between Bacteriodes within the US population and Prevotella within the Malawi and AmerIndian population. Rule Induction was very effective in mining rules of taxa similarity. For instance in many of the other rules found groups of Firmicutes were associated with each of the populations but they were each distinct groups. Because of those associations some of the discriminatory rules were interspersed among non-discriminatory rules. These rules were hard to sort because they showed the same association score (lift score) because they were of the same order

and happened to appear in the same number of samples.

Much of this analyses done and presented are only to provide insight on the co-occurrence problem. This is simply another method geared to interpret various co-occurrence structures within microbial data and can reasonably identify various structures associated with various metadata. More approaches can be taken to understand this very intricate and complex problem to be able to elicit vital networks which could have major implications in understanding the microbiome and general microbiology.

Bibliography

[1] Market basket analysis. http://www.allanalytics.com/author.asp?section$_id$ = 2037doc_id = 253387. *Accessed* : 2012 − 10 − 10.

[2] Nitrogen cycle. http://commons.wikimedia.org/wiki/File:Nitrogen$_C ycle.svg$. *Accessed* : 2013 − 04 − 01.

[3] Rakesh Agrawal et al. Fast discovery of association rules.

[4] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications, volume 27. ACM, 1998.

[5] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In ACM SIGMOD Record, volume 22, pages 207–216. ACM, 1993.

[6] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Data Engineering, 1995. Proceedings of the Eleventh International Conference on, pages 3–14. IEEE, 1995.

[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, volume 1215, pages 487–499, 1994.

[8] John Aitchison. On criteria for measures of compositional difference. Mathematical Geology, 24(4):365–379, 1992.

[9] Michael J Berry and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997.

[10] Nicholas A Bokulich, Sathish Subramanian, Jeremiah J Faith, Dirk Gevers, Jeffrey I Gordon, Rob Knight, David A Mills, and J Gregory Caporaso. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. Nature methods, 10(1):57–59, 2012.

[11] Sergey Brin, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In ACM SIGMOD Record, volume 26, pages 255–264. ACM, 1997.

[12] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. Nature methods, 7(5):335–336, 2010.

[13] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. Genome Biol, 12(5):R50, 2011.

[14] Soumitesh Chakravorty, Danica Helb, Michele Burday, Nancy Connell, and David Alland. A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods, 69(2):330–339, 2007.

[15] Jose C Clemente, Luke K Ursell, Laura Wegener Parfrey, and Rob Knight. The impact of the gut microbiota on human health: an integrative view. Cell, 148(6):1258–1270, 2012.

[16] Elizabeth K Costello, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. Science, 326(5960):1694–1697, 2009.

[17] Sridevi Devaraj, Peera Hemarajata, and James Versalovic. The human gut microbiome and body metabolism: Implications for obesity and diabetes. Clinical chemistry, 2013.

[18] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. PLoS computational biology, 8(7):e1002606, 2012.

[19] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. PLoS Computational Biology, 8(9):e1002687, 2012.

[20] James N Galloway, Alan R Townsend, Jan Willem Erisman, Mateete Bekunda, Zucong Cai, John R Freney, Luiz A Martinelli, Sybil P Seitzinger, and Mark A Sutton. Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. Science, 320(5878):889–892, 2008.

[21] Adriana Giongo, Kelsey A Gano, David B Crabb, Nabanita Mukherjee, Luis L Novelo, George Casella, Jennifer C Drew, Jorma Ilonen, Mikael Knip, Heikki Hyoty, et al. Toward defining the autoimmune microbiome for type 1 diabetes. The ISME journal, 5(1):82–91, 2010.

[22] Michael Hahsler and Sudheer Chelluboina. Visualizing association rules: Introduction to the r-extension package arulesviz. R project module, 2011.

[23] Michael Hahsler, Bettina Grun, and Kurt Hornik. arules: Mining association rules and frequent itemsets, 2006, url http://cran. r-project. org/, r package version. In SIGKDD Explorations. Citeseer, 2007.

[24] Kenneth L Heck Jr, Gerald van Belle, and Daniel Simberloff. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology, pages 1459–1461, 1975.

[25] T Pei Hong, T Jung Huang, and Ch Sheng Chang. Mining multiple-level association rules based on pre-large concepts, 2009.

[26] Jennifer B Hughes, Jessica J Hellmann, Taylor H Ricketts, and Brendan JM Bohannan. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and Environmental Microbiology, 67(10):4399–4406, 2001.

[27] Mike SM Jetten, Markus Schmid, Ingo Schmidt, Mariska Wubben, Udo van Dongen, Wiebe Abma, Olav Sliekers, Niels Peter Revsbech, Hubertus JE Beaumont, Lars Ottosen, et al. Improved nitrogen removal by application of new nitrogen-cycle bacteria. Reviews in Environmental Science and Biotechnology, 1(1):51–63, 2002.

[28] Michael Levandowsky and David Winter. Distance between sets. Nature, 234(5323):34–35, 1971.

[29] Ruth E Ley, Fredrik Backhed, Peter Turnbaugh, Catherine A Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America, 102(31):11070–11075, 2005.

[30] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. BioMed Research International, 2012, 2012.

[31] David Lovell, Warren Muller, Jen Taylor, Alec Zwart, and Chris Helliwell. Caution! compositions! CSIRO, 2010.

[32] Catherine Lozupone, Karoline Faust, Jeroen Raes, Jeremiah J Faith, Daniel N Frank, Jesse Zaneveld, Jeffrey I Gordon, and Rob Knight. Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. Genome research, 22(10):1974–1984, 2012.

[33] Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read, and Konstantinos T Konstantinidis. Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. PloS one, 7(2):e30087, 2012.

[34] Elaine R Mardis. Next-generation dna sequencing methods. Annu. Rev. Genomics Hum. Genet., 9:387–402, 2008.

[35] David Alexander Antonius Mossel, Janet EL Corry, Corry B Struijk, Rosamund M Baird, et al. Essentials of the microbiology of foods: a textbook for advanced studies. John Wiley & Sons, 1995.

[36] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In Database TheoryICDT99, pages 398–416. Springer, 1999.

[37] David M Raskin, Rekha Seshadri, Stefan U Pukatzki, John J Mekalanos, et al. Bacterial genomics and pathogen evolution. Cell, 124(4):703–714, 2006.

[38] Prerna Sethi and Sathya Alagiriswamy. Association rule based similarity measures for the clustering of gene expression data. The open medical informatics journal, 4:63, 2010.

[39] Michael S Silverman, Ian Davis, and Dylan R Pillai. Success of self-administered home fecal transplantation for chronic clostridium difficile infection. Clinical Gastroenterology and Hepatology, 8(5):471–473, 2010.

[40] Roy D Sleator. The human superorganism–of microbes and men. Medical hypotheses, 74(2):214–215, 2010.

[41] Hannu Toivonen et al. Sampling large databases for association rules. In Proceedings of the International Conference on Very Large Data Bases, pages 134–145. INSTITUTE OF ELECTRICAL & ELECTRONICS ENGINEERS (IEEE), 1996.

[42] Outi Vaarala, Mark A Atkinson, and Josef Neu. The perfect storm for type 1 diabetes the complex interplay between intestinal microbiota, gut permeability, and mucosal immunity. Diabetes, 57(10):2555–2562, 2008.

[43] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, et al. Human gut microbiome viewed across age and geography. Nature, 486(7402):222–227, 2012.

Chapter 6

Appendix A: More on Rule Induction

(1) Below I provide the pseudocode for the Apriori algorithm:

Apriori(Transactions, support)

$L_1 \leftarrow \{large1 - itemsets\}$

$k \leftarrow 2$

$While$L$_{k-1} \neq \emptyset$

$C_k \leftarrow \{c|c = a \cup \{b\}$ˆ$a \in$L$_{k-1}$ ˆ b $\in \cup$L$_{k-1}$ ˆ b $\notin a\}$

$fortransactionst \in Transactions$

C$_t \leftarrow \{c|c \in$C$_k$ ˆ c $\subseteq t$

$forcandidatesc \in$C$_t$

count(c) $\leftarrow count(c) + 1$

L$_k \leftarrow \{c|c \in$C$_k$ ˆ count(c) $\geq support$

$k \leftarrow k + 1$

$return \cup$L$_k$

(2) Downward Closure Lemma: If the candidate itemset A is expected to be small in the current pass over the database, then no extension $A + O_j$ of O, where $O_j > O_k$ for any O_k in A is a candidate itemset in this pass. [5]