UNDERSTANDING AND ANALYZING THE HUMAN MICROBIOME: TAXONOMY IDENTIFICATION AND POTENIAL INTERACTIONS

by

Ammar S. Abbas Naqvi

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Masters of Science
Bioinformatics and Computational Biology

Committee:

Dr. Patrick Gillevet, Thesis Director

Dr. Huzefa Rangwala, Committee Member

Dr. Saleet Jafri, Committee Member

Dr. Donald Seto, Department Chairperson

Dr. Richard Diecchio, Associate Dean for Academic and Student Affairs, College of Science

Dr. Vikas Chandhoke, Dean, College of Science

Date: JANUARY 13, 2010

Spring Semester 2010
George Mason University
Fairfax, VA

Understanding and Analyzing the Human Microbiome: Taxonomic Identification and Potential Interactions

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

Ammar S. Abbas Naqvi
Bachelors of Science
Rochester Institute of Technology

Director: Patrick Gillevet, Professor
Department of Bioinformatics and Computational Biology

Spring Semester 2010
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Patrick Gillevet for guiding me and giving me the opportunity to work in his lab, the Microbiome Analysis Center. It was a great honor and experience to learn from his lab and expertise. I would also like to thank Dr. Huzefa Rangwala for mentoring me through out my graduate studies. I would like to also thank Maliha Abbas, my wife, and family for their unconditional support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

UNDERSTANDING AND ANALYZING THE HUMAN MICROBIOME: TAXONOMIC IDENTIFICATION AND POTENTIAL INTERACTIONS

Ammar S. Abbas Naqvi, MS

George Mason University, 2010

Thesis Director: Patrick Gillevet

The Microbiome Analysis Center, in collaboration with Rush University Medical Center and Case Western University, has been studying and characterizing gut microbiota in normal and diseased states such as ALD and HIV. Currently, the Multitag Pyrosequencing (MTPS) methodology developed by Dr. Gillevet is being used to interrogate the microbiome from dozens of samples at a time. As a result we have been receiving hundreds of samples of both bacterial and fungal microbiome of stool, mucosal biopsy, and oral samples from a large number of subjects and diseases.

The vast volumes of data flowing from diverse sources has necessitated the development of analysis pipelines in order to intelligently and rapidly process the molecular information and to analyze, cluster, and correlate the sample datasets. However, a fundamental and pre-requisite for most research in this particular is being able to efficiently and accurately identify the genus and species information given a set of SSU rRNA sequences. The current implementation of this type of investigation is widespread, but as datasets get very large it proves to be very impractical due to factors

concerning run-time and memory. For this particular study, we have developed and designed a portable and robust tool to identify the bacterial taxonomy and distribution at the species level, specifically in patients with HIV looking at the vaginal microflora.

Another very important aspect of the Microbiome is to understand the relationships of the bacteria between and amongst different classes (ie. healthy, diseased). In order to accomplish this we plan on applying a systems biology approach to the microflora. This study will produce an approach that will specifically look at the gut microbiota in relation to Alcohol Liver Disease at the graphical network level.

A series of challenges is anticipated related to time and memory constraints, informative identification, and proper linkage of taxonomic identification to the clinical information in the microbiome. We discovered distinct and common features amongst our samples that will provide novel insights and ultimately broaden our understanding of how the microbiome influences human health, furthering future research in this rapidly progressing field.

CHAPTER 1

Background

1.1 Background - The Human Microbiome

A microbiome is the entire set of microbial cells, including their genomes and interactions in a particular environment. The human body contains one of the most densely populated microbial environments or microbiomes known on earth. Over 1×10^{14} microbial cells interact with ours, indicating the ubiquity and potentially critical importance of such interactions in our bodies, specifically the digestive tract [1-3]. There are also other sets of microflora sites in our bodies including skin, nose, mouth and vagina, which are all outlined and displayed in Figure 1.1.

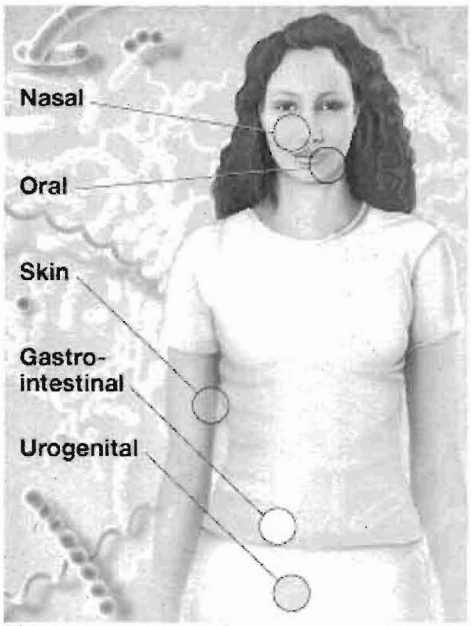


Figure 1.1 Microflora sites in humans (from [12])

Controlled interactions between digestive tract epithelial and immune cells with microbial cells are critical to human health. They are involved with the immune system and its responses, metabolic regulation, and digestion. The human digestive tract is an interface between the human body and the environment and represents an important portal that regulates the level of exposure to environmental factors that is hypothesized to play a key role in the state of health or disease. The microbiome and the gut epithelial barrier are hypothesized to be essential for regulating and maintaining normal mucosal and systemic immune functionality. The underlying paradigm is that the gut microbiome actively interacts with the human host through quorum sensing and immune mechanisms and are in homeostatic equilibrium in the healthy state.

In healthy individuals, controlled interactions of the digestive tract epithelial and immune cells and microbial cells provide a training ground for the body's immune system by regulating metabolic functions, enabling proper digestion and absorption,

providing access to essential vitamins, and conferring protection from intestinal pathogens. However in diseased individuals, the interface connecting interactions may become modified leading to detrimental effects, leading to immune deregulation, and many inflammatory and autoimmune diseases. [1-3]

Some well-studied diseases that have been explained through gut microbiome expression analysis include Inflammatory Bowel Disease, Bacterial Vaginosis, Alcohol Liver Disease, and Obesity. Further development in methods and tools for characterizations of these interactions of the microbiome, mucosal and immune system, and the human genome will allow us to gain valuable insights as to underlying mechanisms of how the microbiome affect health and disease.

1.2 NextGen Sequencing

Nexgen sequencing usually refers to the recently developed technology, which enables for high-throughput systematic sequencing since the Sanger sequencing method. Currently, there are four common and used commercial sequencing platforms for this, including Illumina, 454 Roche, ABI, and Helicos. All are based on DNA replication methods. Due to the increased demand for cost-efficient sequencing there is a drive for the development of high-throughput sequencing technologies that parallelize the sequencing process, producing not only thousands, but also millions of sequences at a time. Nextgen or high-throughput sequencing technologies are both fast and inexpensive. As a result, it has become a driving force for new research. All of the methods listed above, with the exception of 454, are based on generating massive amounts of short reads

(30-45 base pairs). As a result, one of the major strengths of the 454 platforms and method is that it generates longer reads (now 500 bp). The longer reads are much more advantageous when dealing with de-novo analysis. This technology or methodology is specifically known as pyrosequencing and is primarily used in this study and similar studies. [4]

Pyrosequencing is a method of DNA sequencing based on the "sequencing by synthesis" principle. It differs from the other popular methods, because it relies on the detection of a pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides. The technique was developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm in 1996. The method involves taking a single strand of the DNA and then enzymatically synthesizing its complementary strand. The Pyrosequencing method is based on detecting the activity of DNA polymerase by coupling the detection of the pyrophosphate with another chemiluminescent enzyme called luciferase. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step. The template DNA is immobilized, and solutions of A, C, G, and T nucleotides are added and removed after the reaction, sequentially. Chemiluminescence is only produced when the nucleotide solution complements the first unpaired base of the template. The sequence of reagents that produces the luminescent signals allows the determination of the sequence of the template. [5-7] Figure 1.2 shows a general schematic of the procedure.

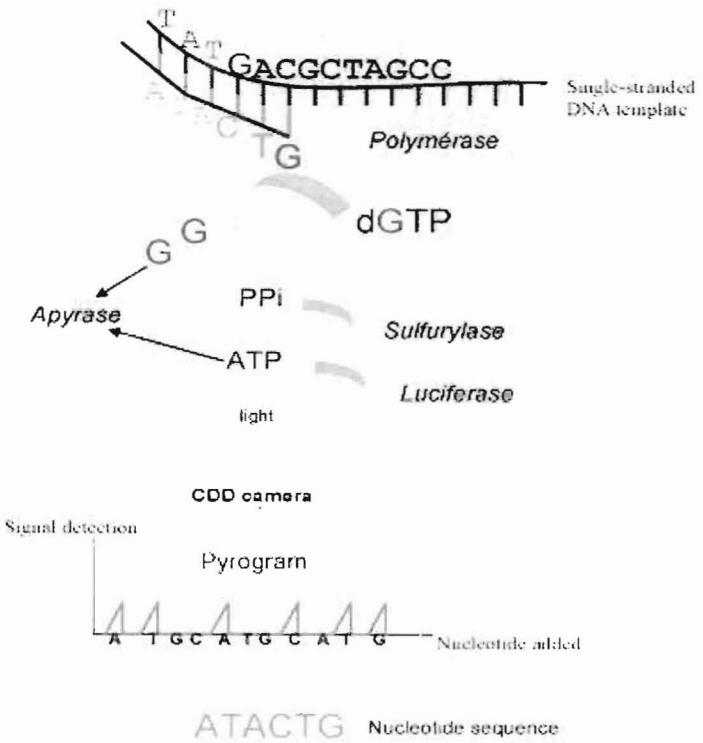


Figure 1.2 Pyrosequencing Schematic (from [13])

Nevertheless, a novel Multitag Pyrosequencing (MTPS) methodology, which has been developed by our group, was used to sequence the samples in our studies. This is an extension of the previously described pyrosequencing procedure. It takes individual DNA molecules and amplifies them on beads using an oil emulsion PCR procedure. Individual beads are then sequenced using the normal pyrosequencing methodology [7]. The improvement resides in incorporating tagged fusion primers to barcode multiple samples at a time. Thus, dozens of samples are sequenced simultaneously allowing us higher throughput.

1.3 16S rRNA

The 16S ribosomal RNA (rRNA) gene is a region of prokaryotic DNA found in all bacteria and archaea. This specific gene encodes for an essential part of the ribosome. In bacteria, the small subunit is coded for by the 16S rRNA gene, and the large subunit is coded for by the 23S rRNA and 5S rRNA genes. The 16S rRNA gene is a commonly used feature for identifying the taxonomy of bacteria for several reasons. For example, researchers may want to identify or classify only the bacteria within a given environmental sample. Since, the gene is distinct, it is considered a very useful feature for identifying bacterial species, making it very useful for metagenomic and related studies. In addition, the 16S rRNA gene is relatively short at 1.5kb in length making it easier and inexpensive to sequence. Lastly and perhaps the most important reason is that it contains the species-specific signatures that allows us to identify distinct species within a bacterial sample or data set. [8]

1.4 Available Tools

Since the advent of NextGen sequencing technologies there has been a drive or a demand to efficiently analyze human microbiomic or metagenomic data. In addition to sequencing these samples, there is a need to analyze them. Many old statistical methods cannot handle the sheer volume of information being processed, so many tools are currently being developed with modified algorithms and methods in order to facilitate the data processing. Some of the popular tools for data and statistical analysis are UniFrac [9,10] and MEGAN. [11]

1.4.1 UniFrac

Statistical methods and tools have been developed using traditional methods of comparisons, but many of these techniques are limited in nature because they do not account for the different degrees of similarity between sequences and the abundance content or diversity within samples. As a result, a substantial loss of information is attributed to the analysis.

A tool called UniFrac introduced a new method for computing differences between microbial communities based on phylogenetic information and abundance. It measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that separates various taxa from either one environment or the other, but not both. It can be used to determine whether communities are significantly different, to compare many communities simultaneously using clustering and ordination techniques, and to measure the relative contributions of different factors, such as chemistry and geography, to similarities across samples.

It takes as input a single phylogenetic tree, which can be generated using any of the tree making tools (ie. PAUP) that contains sequences derived from at least two different environmental samples and an environmental file describing the origins of the sequences and their abundance. Either the UniFrac distance metric or the P test or both can be used to make the comparisons. It can be used to compare many samples simultaneously, because it satisfies the technical requirements for a distance metric and can thus be used with standard multivariate statistics such as un-weighted pair group method using average linkages (UPGMA) clustering and principal coordinate analysis. [9, 10]

It is more powerful than non-phylogenetic distance measures, because it exploits the different degrees of similarity between sequences. The ability to integrate sequence data from many diverse studies makes it ideal for large-scale comparisons, between environments despite the variability in data collection techniques.

1.4.2. MEGAN

MEGAN is a taxonomic identification tool that allows one to interactively exploring the content of a data set, using the NCBI taxonomy to summarize and order the results. The program uses a basic algorithm to assign each read to the lowest common ancestor of the set of taxa that it hit in the comparison procedure.

As a result, species-specific sequences are assigned to taxa near the leaves of the NCBI tree, whereas widely conserved sequences are assigned to high-order taxa closer to the actual root. It deviates from the analytical pattern of previous metagenomic analysis pipelines and builds on the power of comparing random sequence intervals against known databases regardless of phylogenetic properties. Despite the incompleteness of databases, which is common in these types of studies, this approach is gives one an overall generalization of the dataset. [11]

Nevertheless, the main benefit of this tool is that it allows large data sets to be dissected without the need for assembly or the targeting of specific phylogenetic markers, which can be very time and memory intensive at times. Hence, it speeds up the analysis many fold. Furthermore, it also provides visual and statistical output for comparing different datasets. However, the lack of data in these databases may result in false

negatives reads, but will not result in a significant amount of false positives, since it uses a very conservative approach. With its pros and cons it has demonstrated its usefulness as it has been used on many microbial and metagenomic datasets.

1.5 Other Resources

A plethora of additional resources exist for similar data and computational analysis of these types of samples. A central repository for all microbiomic data is available at the Human Microbiome Project: Data Analysis and Coordination Center. The resource provides a great deal of valuable information, including reference genomes, whole-shotgun sequencing data, and related clinical data [2]. A similar resource is CAMERA, the Community Cyber-infrastructure for Advanced Marine Microbial Ecology Research and Analysis database, which is a web resource providing databases of raw environmental sequence data, associated metadata, pre-computed search results, and high-performance computational tools. Other such tools exist and a great number are coming into existence due to development of new methodologies and analysis pipelines.

1.6 References

1. Hsiao, W.W. and C.M. Fraser-Liggett, Human Microbiome Project--paving the way to a better understanding of ourselves and our microbes. Drug Discov Today, 2009. 14(7-8): p. 331-3.
2. Turnbaugh, P.J., et al., The human microbiome project. Nature, 2007. 449(7164): p. 804-10.

3. Turnbaugh, P.J., et al., A core gut microbiome in obese and lean twins. Nature, 2009. 457(7228): p. 480-4.

4. Mardis, E.R., The impact of next-generation sequencing technology on genetics. Trends Genet, 2008. 24(3): p. 133-41.

5. Ronaghi, M., Pyrosequencing sheds light on DNA sequencing. Genome Res, 2001. 11(1): p. 3-11.

6. Ronaghi, M., Improved performance of pyrosequencing using single-stranded DNA-binding protein. Anal Biochem, 2000. 286(2): p. 282-8.

7. Poirel, L., T. Naas, and P. Nordmann, Pyrosequencing as a rapid tool for identification of GES-type extended-spectrum beta-lactamases. J Clin Microbiol, 2006. 44(8): p. 3008-11.

8. Cole, J.R., et al., The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res, 2005. 33(Database issue): p. D294-6.

9. Lozupone, C., M. Hamady, and R. Knight, UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics, 2006. 7: p. 371.

10. Lozupone, C. and R. Knight, UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol, 2005. 71(12): p. 8228-35.

11. Huson, D.H., et al., MEGAN analysis of metagenomic data. Genome Res, 2007. 17(3): p. 377-86.

12. The Human Microbiome. The Human Microbiome 2009; Available from:

http://nihroadmap.nih.gov/hmp/.

13. Pyrosequencing Technology. 2009 [cited 2009; Available from: http://www.pyrosequencing.com/DynPage.aspx?id=8726.

CHAPTER 2

Taxonomic Identification and Analysis of Multitag Pyrosequence data from Human Microbiome Samples

2.1 Abstract

The lab has been using the Roche GS FLX sequencing platform to produce tens of thousands of sequencing reads from samples of both bacterial communities (microbiome) and fungal communities (mycobiome) of stool, gut mucosa, vaginal washes, and oral washes from a large number of subjects. This vast volume of data from diverse sources has necessitated the development of an analysis pipelines in order to systematically and rapidly identify the taxa within the samples and to correlate the sample data with clinical and environmental features. Specifically, we have developed automated analytical tools for data tracking, taxonomical analysis, and feature clustering.

We have developed a portable and robust tool to identify the taxonomy and abundance of bacteria in the human microbiome and demonstrate the pipeline using cervical vaginal lavage (CVL) samples.

This analysis pipeline will not only provide insight to our specific CVL dataset, but is applicable to other microbiome and to Metabiome samples produced from metagenomic methods and will ultimately broaden our understanding of how the

microbiome influences human health furthering future research in this rapidly expanding field.

2.2 Background

The human microbiome is the entire set of microbial cells, including their genomes in a particular ecological niche on or in the human body. The human body contains one of the most densely populated microbial environments known on earth. Over 1 x 10^{14} microbial cells interact with ours [2, 3], indicating the ubiquity and potentially critical importance of such interactions in our bodies, especially in the digestive tract. There are also other sets of microflora sites in our bodies including skin, nose, mouth, genital tract and lungs. Controlled interactions between the microbes in the protective mucosal gut biofilm and gut epithelial and immune cells are critical to human health. We define the interactions of a human microbiome with the host metabolism and immune systems as the "Metabiome". These interactions are involved in the immune system and its responses, metabolic regulation, quorum sensing, and digestion. In diseased states, the normal microbiome composition can shift (dysbiosis) altering the interactions and the functionality of the Metabiome.

We report the development and validation of an analytical pipeline that calculates the taxonomic distribution across the samples which is practical and addresses major obstacles encountered in the analysis of microbiome samples such as dataset size and taxanomic identification. It results in a quick and efficient method to characterize samples based on taxonomy and abundance.

We evaluated our tool on a dataset related to the lower genital tract microbiome sampled by Cervical Vaginal Lavage (CVL). Our subjects were healthy females and females with HIV with and without Bacterial Vaginosis (BV). BV is a serious condition found in females aged 15 to 44 caused by an imbalance of naturally occurring bacterial flora. A healthy vaginal microbiome normally consists predominantly of Lactobacillus. The bacteria that make up BV are very diverse which may cause an imbalance and lead to detrimental health effects. Specifically, it has been observed that patients with the BV condition have a higher incidence of heterosexually transmitted HIV, indicating there must be some sort of interaction between the microbes and the human body.[10]

Our research lead to the development and validation of a tool that calculates the taxonomic distribution across the samples, which is practical and an applicable pipeline that surpasses major obstacles of dataset size and similarity. It results in a quick and efficient method to characterize samples based on taxonomy and can be applied to a variety of microbiome 16S rRNA datasets.

2.3 Methods

Datasets Used

The data we used for this study was derived from a set of Cervica Vaginal Lavage samples provided by G.T. Spear. [13]. CVL samples were obtained from 21 women divided into 4 groups on the basis of HIV seropositivity (HIV^+ or HIV^- status) and Nugent Gramstain analysis for BV. Samples that were obtained from BV^+ women had scores of 7–10, whereas samples obtained from BV^- women had scores of 0–3. The

HIV^+BV^+, HIV^-BV^+, and HIV^-BV^-) groups each consisted of 5 subjects, whereas the HIV^+BV^- group consisted of 6 subjects. The samples were subjected to a novel multiplexed pyrosequencing method by generating a set of 12 primers that each contained either the 27F or 355R primer that was tagged on the 5' end with a 4-base "bar code." PCR was performed on individual patient samples by use of the unique barcoded primers, and 10–12 samples then were pooled and ligated to the PCR linkers used in the emulsion step of pyrosequencing. All samples were amplified for 30 cycles, as described in Spear et al 2008[13]. Pyrosequencing of the amplified, tagged DNA was performed 454 Life Sciences, with the use of 10–12 separately tagged samples included in a single slot. The data from each well were "deconvoluted" by sorting the sequences into bins on the basis of the bar codes, and the taxa in the samples were normalized by the total number of reads from each bar code.

There are four distinct classes in the set, including those that have Bacterial Vaginosis (BV) and are negative or positive for HIV and those that do not have BV and are either negative or positive for HIV. In this report, we primarily used the two BV classes to validate the clustering, since it is the more diverse set in these samples. WE obtained 15,874 reads for these samples, 905 for the BV- samples and 6,968 for the BV+ samples each consisting of five patients, and these samples contained around 50 different bacterial species.

Performance Metrics

In order to verify the cluster reliability we calculated the p-values between our clusters and contigs output. In addition, we also tested the final taxa distribution against the original pipeline using the correlation coefficient and t-test measures.

Computational Pipeline

As reflected in Figure 2.1, we had already developed a prototype pipeline to identify the taxa and compute the distribution in the samples using a brute force approach that searched the RDP8, RDP10, or Genbank databases using the BLAST algorithm. However, this became impractical with the large number of sequences produced by the Multitag Pyrosequencing data.

Specifically, there were almost 10,000 sequences from the BV-HV- sample that were needed to align and search for against NCBI's nucleotide database. Despite, local high end computing power (Mac OS, 2x3Ghz Dual Core Intel Xeon, 4gb 667 Mhz DDR FB-DIMM), the local BLAST process crashed the local server. As a result, PERL scripts had to be developed so we could divide the original sequence file into smaller parts and then merge and parse the BLAST results. We tested a variety of combinations of the options or parameters available through BLAST and we found that that a minimum cut off for percent identity of 96%, an e-value of .000000001, a bit size of 60 and a word size of 50 were optimal with over 95% of the query sequences producing significant hits.

Nevertheless, even with the more stringent filtering, we were still getting over a thousand hits per query, which limited the performance of the analytical pipeline. As a result, we applied an additional layer of filtering, our own ranking mechanism, where we

sorted the results by the percent identity, bit-score, and then by the coverage of the alignment, respectively and parsed out only the highest match.

Our goal is to produce an abundance table for each taxa in a microbiome sample so we extracted all of the unique accession identification numbers, created a database file with all of these accession numbers, and then ran the file through our local NCBI fastacmd program which outputted the taxonomic information for each hit.

We then linked up each hit from the parsed results with its proper taxonomic identification for each analyzed data subset and then concatenated all of our results into one flat file database. After creating these linkages between the sequences and taxonomy, we normalized the abundance of each the genus and species for each sample and used this normalized abundance table for clustering analysis. Despite our efforts of optimization this approach was still very slow, especially if the number of sequences to analyze are high. Thus we developed the revised pipeline illustrated in Figure 2 and compared the clustering based on CD-HIT with a full assembly program (Seqman) and evaluated the pipeline's performance using a variety of measures. The general outline for the validation of the revised pipeline is as following:

- Cluster the given microbiome data set using the CD-HIT algorithm
- Perform preliminary analysis in order to verify that the initial clusters are reliable
- Assemble the original dataset in Seqman in order to generate consensus sequences of assembled contigs

- Format the two sets (clustering and contig) and run Unifraq in order to measure statistical significance of the two sets generated from the two tools (CD-HIT and Seqman)
- Assess accuracy by comparing the distribution using the different pipelines with each other
- Perform run-time analysis on the different pipelines

Using CD-HIT

In order to cluster the dataset, we used a tool called CD-HIT, which was designed and developed by Weizhong Li at the University of California in San Diego. CD-HIT was originally developed for clustering large protein database at a particular sequence identity threshold and uses an incremental clustering algorithm and was recently modified to analyze nucleotide sequences. It is widely used in educational groups and institutions such as UniProt, Protein Data Bank, European Bioinformatics Institute, and the Venter Institute [14-16].

We fed our fasta-formatted sequences into the program using a minimum threshold value of 98% identity and a word size of 9 that was found to be optimal for to identify taxa at the species level. The results included two files – a cluster file with all of the clusters and its members and a fasta file with the longest or representative sequence for each cluster. We then subjected the representative sequences of each cluster to local BLAST using the Genbank database.

Cluster Validation

Since, the clustering step was the most crucial step in our new methodology we proceeded to verify that our clusters were indeed reliable and stable. We first checked the total members of a particular cluster to see if they all produced the same significant hits using BLAST. Secondly, we investigated the multiple sequence alignments of the assembly of each cluster manually using Seqman, a NextGen assembler developed by DNASTAR Inc, Madison, WI [17]. For these particular assessments we chose 5 random clusters of three or more members and performed the procedure on each of those clusters. For the BLAST method we used megablast and used the same options as we would in our original pipeline with a minimum percent identity of 96%, an e-value cutoff of 1 x -10, a word-size of 50 and a bit-score cut off of 60. Similarly for the Seqman assembly, we used the same percent identity.

To further and verify the validity of the clusters, we ran Unifrac.[9] on the contigs and clusters. Unifrac uses both phylogenetic information and abundance information to statistically compare microbial communities. We first divided the datasets into forward and reverse reads and then assembled each of the two sets (forward and reverse) into distinct contigs produced by Seqman and clusters generated by CD-HIT.

We then took the respective contigs and clusters and aligned them with CLUSTALW [18, 19] and then used PAUP [20, 21] to generate a tree file. We used a PERL script to generate an environmental file that indicated where each sequence had come from, including the sample or patient origins and either a Seqman or CD-HIT environmental type. We fed these two files into the Unifrac [11] tool which produced a matrix showing the P-value of the comparisons for each environment against each other.

The matrix was computed with the Bonferroni correction to correct the P-values that is reported for multiple comparisons which is performed by multiplying the raw P-value by the number of permutations. [10, 11]

Taxonomic Identification and Distribution

After verifying the reliability of the clusters, we continued with extracting sequence information of the representative of each cluster, which the program chose based on the length of the sequence, and ran BLAST on those representative sequences. Hence, instead using BLAST on each sequence in our full dataset, we just used BLAST on a single member from each cluster speeding up the NCBI database search significantly, usually by more than ten times.

After retrieving the best hit, which we ranked by percent identity, e-value, and bit-score respectively we retrieved the genbank accession numbers for each hit following the same procedure as we did in our original pipeline. We made a separate file for the accession numbers, which we then fed into a script, Fastacmd that looked up the genus and species information. Fastacmd is a widely used program that NCBI offers as one of its tools to interact with the genbank database.

Once we annotated the genus and species information back to the representative sequences that were BLASTed, we then went back to the cluster file that was produced previously and annotated each cluster member with the appropriate information.

Next we applied the same shell and Perl scripts that were originally developed used to count species abundance information to this annotated and labeled dataset, so we can visualize and inspect the species distribution in the entire set.

Comparative Analysis

Finally, we compared our results from this newly developed pipeline with the original pipeline in order to ensure and assess the overall precision and accuracy of the revisions and modifications introduced was comparable and reliable.

Run-time Analysis

As a final step to confirm that the CD-HIT version was indeed very quick and portable we compared the run-time against our original BLAST pipeline and another commonly used pipeline, where the reads are assembled and then the contigs are used with BLAST.

2.4 Results and Discussion

Using CD-HIT

Revising our old pipeline (Figure 1) by implementing our newly developed procedure shown Figure 2. 2, we first clustered the raw reads. Our results from the crucial clustering step are shown in figure 3. In the BV-negative (BV-) set of 8,905 we found that there were a total of 1,063 clusters that were produced with our threshold. Similarly the BV+ set of 6,968 reads generated 1,000 clusters. Figure 2.3 represents the cluster size

and sequence membership distribution of the BV- set, which is almost identical to the BV+ class. Almost 50% of the sequences are in clusters that are of cluster sizes 50 to 300, nevertheless, there are significant numbers of sequences in other cluster sizes as well. Fortunately, the clustering step has reduced the number of queries we need to feed to the BLAST by a factor of more than 8, which will inevitably speed up the process.

Cluster Validation

We checked 5 different clusters and all the members of each clusters hit the same matches when compared to other cluster members. However, in order to confidently verify their reliability, we continued on to the more robust test on our cluster stability with Unifrac. Looking at matrix displayed in table 2.1, the last two letters (ie. FC) display an environment in our dataset, which in our case is a patient. For example SBV7397_II_FC comes from environment "FC." And the "SBV" portion just indicates that this read had originated from Seqman or an assembly, while "BV" indicates a CD-HIT or clustering origin. We can conclude that each read reflecting the same environment has a p-value score of 1.0, verifying a very strong similarity. For example read BV7397_II_FC, a CD-HIT cluster sequence, is equivalent to SQ7397_II_FC, a Seqman contig sequence. These comparisons tell us that the clusters and their respective members all have very strong alignments amongst each other, which indicate a strong confirmation of the reliability and stability of our clusters.

Taxonomic Identification and Distribution

Finally we continued to the next step in our pipeline, which was the final step to verify its accuracy. We compared the distribution from the different pipelines (BLAST, contig, and revised) on the samples. As we can observe in Figure 2.4 there was some loss of information, but nonetheless, the abundance information was quite similar and comparable validating our newly proposed pipeline. Nevertheless, we calculated the correlation coefficients' and t-tests for all three RDB levels and all were positive with the strongest correlation being on level 3 and the weakest on level 5, which is reflected in table 2.2.

Run-Time Analysis

Our dataset consisted of 8,906 sequence reads. We compared our CD-HIT method with our previous BLAST pipeline and another common method of running BLAST on the contigs. Figure 2.5 shows the run time in seconds for all three types of methods. The run-time for the contig method using Seqman and the clustering method using CD-HIT all included the additional steps (ie. assembly and clustering). As we can see, CD-HIT is significantly quicker being almost 20 times faster than our original BLAST pipeline and almost 40 times faster than the popular contig pipeline.

2.5 Conclusions

We addressed a series of challenges related to computer time and memory constraints, informative identification, and proper linkages of taxonomy to the clinical

data to develop an accurate analytical pipeline to determine the taxa abundance in large datasets from human microbiome samples.

Our implementation and proposed pipeline has made this type of analysis practical as current implementation of this type of investigation analyze the individual sequencing reads, but as datasets get very large this proves to be impractical due to run-time and memory considerations.

The clusters that were produced were reliable enough to conclude that CD-HIT is a tool that can successfully cluster microbiome sequences that are 97-100% similar. The taxonomic distribution was also significantly similar showing us that we can simply BLAST a representative sequence from each cluster and still obtain comparable results. Finally, the runtime of the pipeline reflected a 20-40-fold increase in speed when compared to other well-defined methods. Usually researchers have traditionally used the contig method to make the procedure, specifically the BLAST step, less exhaustive and more practical, but our clustering pipeline proves to be of greater efficiency. Hence, the revised pipeline makes taxonomic identification and distribution analysis much more efficient and practical, especially when the total number of samples or reads reach a high number. In the future, we would test the pipeline on metagenomic and other microbial samples to further our confidence.

2.6 References

1. Turnbaugh, P.J., et al., A core gut microbiome in obese and lean twins. Nature, 2009. 457(7228): p. 480-4.

2. Turnbaugh, P.J., et al., The human microbiome project. Nature, 2007. 449(7164): p. 804-10.

3. Spear, G.T., et al., Comparison of the diversity of the vaginal microbiota in HIV-infected and HIV-uninfected women with or without bacterial vaginosis. J Infect Dis, 2008. 198(8): p. 1131-40.

4. Li, W., L. Jaroszewski, and A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 2001. 17(3): p. 282-3.

5. Li, W. and A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 2006. 22(13): p. 1658-9.

6. Li, W., L. Jaroszewski, and A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics, 2002. 18(1): p. 77-82.

7. Swindell, S.R. and T.N. Plasterer, SEQMAN. Contig assembly. Methods Mol Biol, 1997. 70: p. 75-89.

8. Thompson, J.D., T.J. Gibson, and D.G. Higgins, Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics, 2002. Chapter 2: p. Unit 2 3.

9. Fukami-Kobayashi, K. and N. Saito, [How to make good use of CLUSTALW]. Tanpakushitsu Kakusan Koso, 2002. 47(9): p. 1237-9.

10. Matthews, L.J. and A.L. Rosenberger, Taxon combinations, parsimony analysis (PAUP*), and the taxonomy of the yellow-tailed woolly monkey, Lagothrix flavicauda. Am J Phys Anthropol, 2008. 137(3): p. 245-55.

11. Wilgenbusch, J.C. and D. Swofford, Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics, 2003. Chapter 6: p. Unit 6 4.

12. Lozupone, C. and R. Knight, UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol, 2005. 71(12): p. 8228-35.

13. Lozupone, C., M. Hamady, and R. Knight, UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics, 2006. 7: p. 371.

2.7 Figures

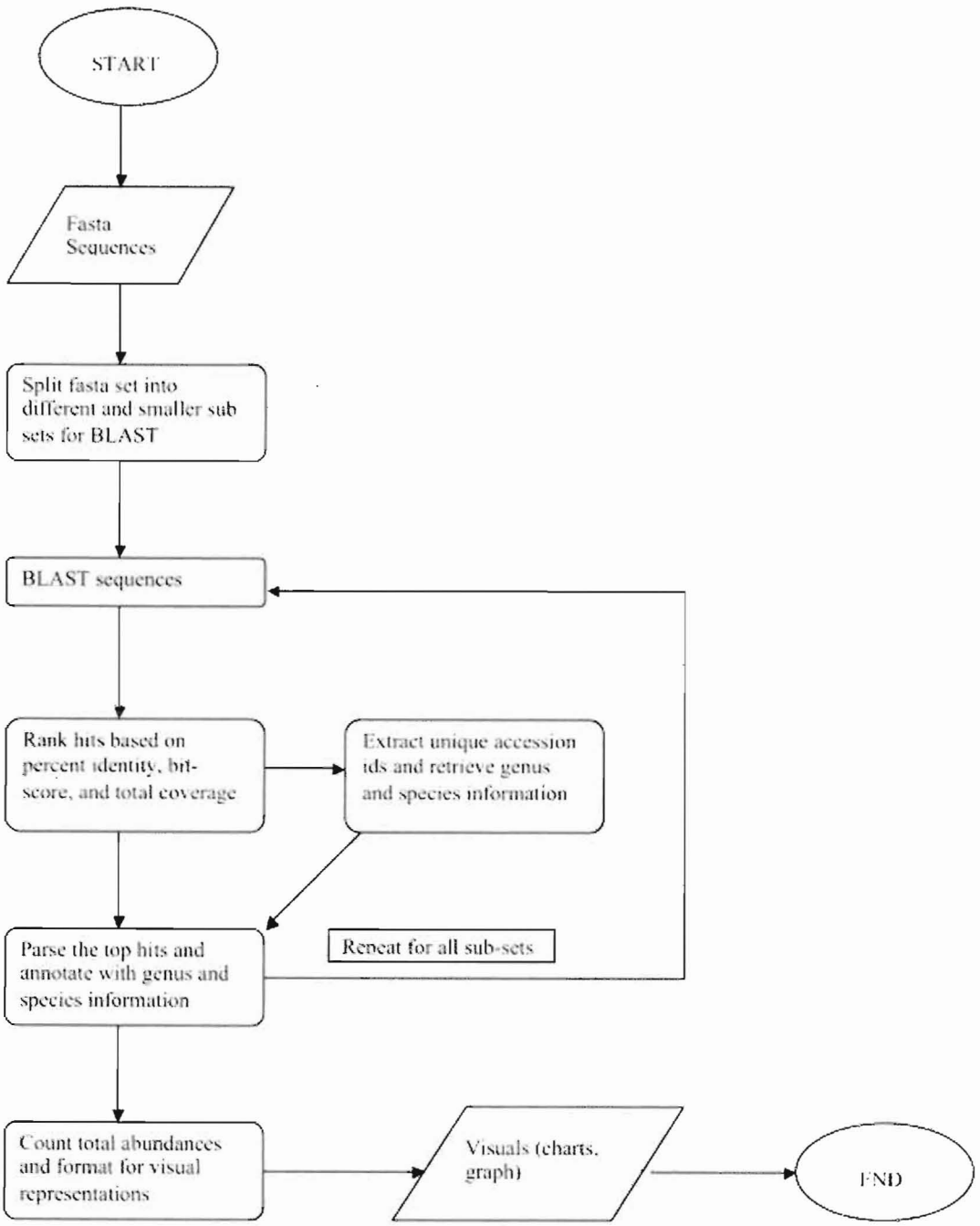


Figure 2.1- Original Taxonomic Identification Pipeline
Popular flow of methods and procedures to obtain taxonomic identification and distribution with each box being a discreet step in the process

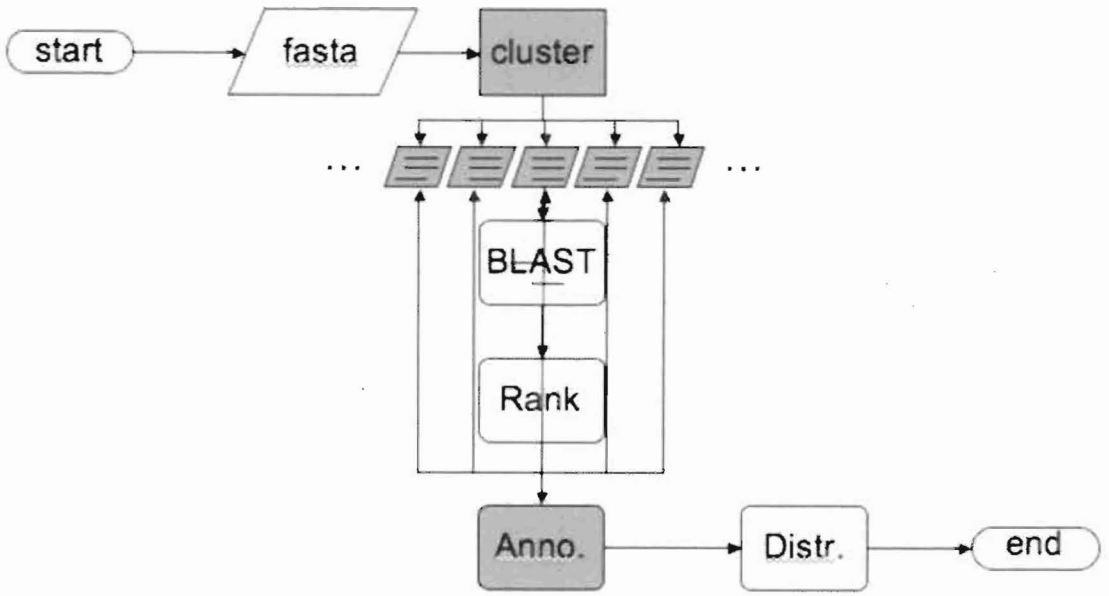


Figure 2.2 - Revised Analysis Pipeline
Revised and proposed flow of methods and procedures to obtain taxonomic identification and distribution with each box being a discreet step in the process

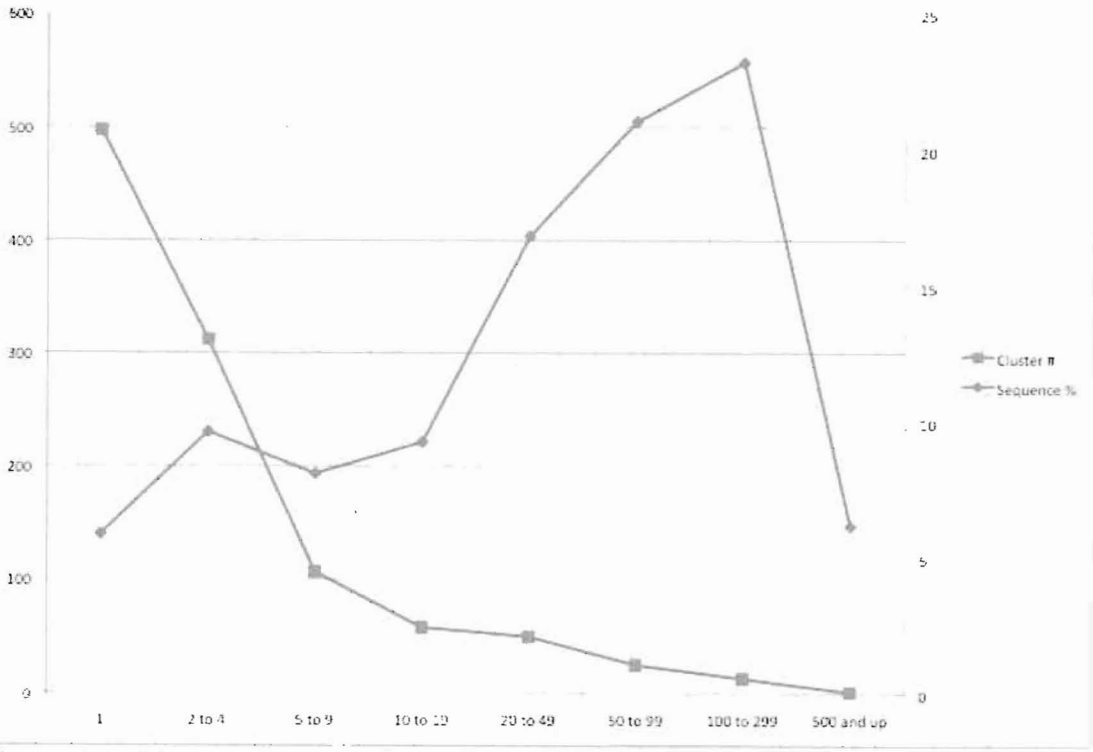


Figure 2.3 - Line graph of cluster size versus percentage of total sequences
Line graph with the x-axis as the cluster size and the y-axis as the percentage of sequences contained in that particular cluster with CD-HIT for BV-.

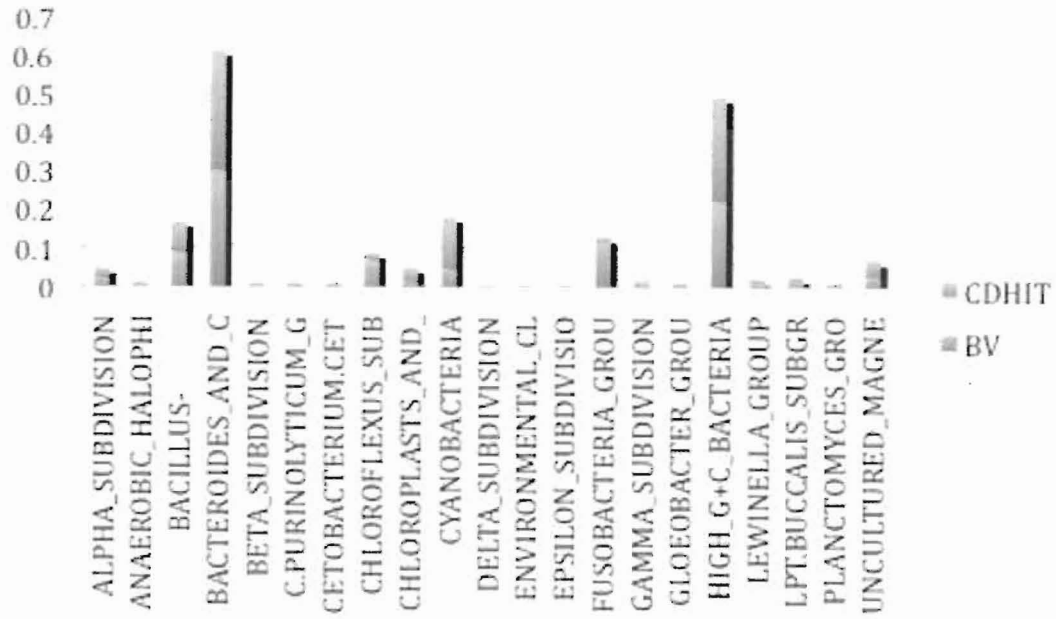


Figure 2.4 - Taxonomic distribution in the CD-HIT and original (BV) pipelines on RDB level 3
Stacked histogram of the taxa content of the old and revised pipelines

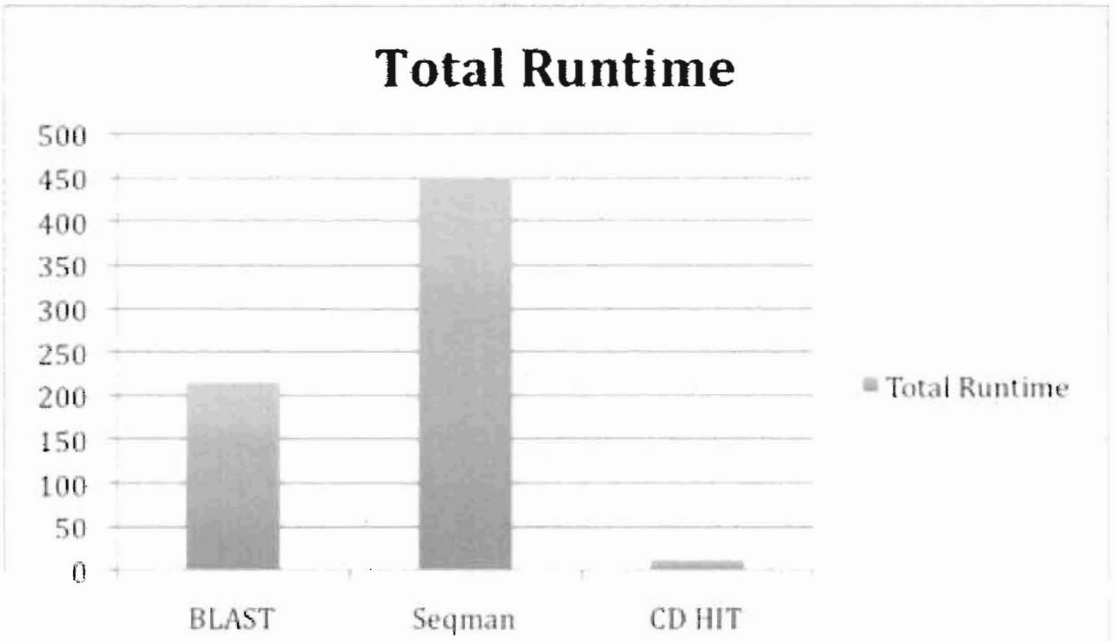


Figure 2.5 - Runtime (in seconds) for each type of method
Histogram measuring the time (seconds) of the two common methods (straight BLAST, contig) and the new CD-HIT version

2.7 Tables

Table 2.1 Unifrac Table of P-Values - Unifrac result data matrix of the P-values of each environment in the samples (BV-,BV+).

	SBV0329_II_FB	SBV7397_II_FC	SBV7400_II_FD	SBV8123_II_FE	SBV9055_II_FA	BV0329_II_FB	BV7397_II_FC	BV7400_II_FD	BV8123_II_FE	BV9055_II_FA
BV0329_II_FB	1	0.3	0.77	0	0.11	0	0.13	0.01	0	0.22
BV7397_II_FC	0.01	0.98	0.04	0	0	0.13	0	0.13	0	0.23
BV7400_II_FD	0.01	0.27	1	0.13	0.47	0.01	0.13	0	0	0.02
BV8123_II_FE	0	0	0.08	1	0	0	0	0	0	0.04
BV90	0	0.47	0.88	0.43	1	0.22	0.23	0.02	0.04	0

55_II_FA										
SBV0329_II_FB	0	0.9	0.91	0.11	0.79	1	0.01	0.01	0	0
SBV7397_II_FC	0.9	0	0.56	0.01	1	0.3	0.98	0.27	0	0.47
SBV7400_II_FD	0.91	0.56	0	0.8	0.99	0.77	0.04	1	0.08	0.88
SBV8123_II_FE	0.11	0.01	0.8	0	0.95	0	0	0.13	1	0.43
SBV9055_II_FA	0.79	1	0.99	0.95	0	0.11	0	0.47	0	1
	SBV264_I_FC	SBV612_I_FE	SBV787_I_FG	SBV817_I_FH	SBV914_I_FO	BV264_I_FC	BV612_I_FE	BV787_I_FG	BV817_I_FH	BV914_I_FO
BV264_I_FC	0.94	0	0	0	0.03	0	0	0	0	0
BV612_I_FE	0	0.94	0.27	0	0	0	0	0.26	0	0
BV787_I_FG	0.05	0.02	0.84	0	0	0	0.26	0	0	0
BV817_I_FH	0	0	0	1	0.04	0	0	0	0	0.07
BV914_I_FO	0.02	0	0.12	0.47	0.97	0	0	0	0.07	0
SBV264_I_FC	0	0.03	0.1	0	0.15	0.94	0	0.05	0	0.02
SBV612_I_FE	0.03	0	1	0	0	0	0.94	0.02	0	0
SBV787_I	0.1	1	0	0	0.16	0	0.27	0.84	0	0.12

FG										
SBV817_I_FH	0	0	0	0	0.82	0	0	0	1	0.47
SBV914_I_FO	0.15	0	0.16	0.82	0	0.03	0	0	0.04	0.97

Table 2.2. Statistical Measures of Taxa Distribution - Correlation Coeffecients and Student T-Tests calculations between the original and revised pipelines.

RDB Level	Correlation Coeffecient	Students T-Test
Level 3	0.912903583	1
Level 4	0.912903583	1
Level 5	0.370181378	1

CHAPTER 3

Network-based Modeling for Analyzing the Human Gut Microbiome

3.1 Abstract

Background

The human gut contains one of the most populated microbial communities in the world. The influence of these microbial communities on the human development, immunity, and physiology is largely unstudied. In this paper we used a network-based approach to characterize the microflora in colonic mucosal samples and correlate potential interactions between the identified species with respect to the healthy and diseased states.

Results

We performed our analysis on the abundance data produced from the 16S rRNA sequencing of the bacteria within mucosal microbiome samples. We analyzed the modeled network by computing several local and global network statistics, identified recurring patterns or motifs, fit the network models to a family of well-studied graph models.

Conclusions

This study has demonstrated, for the first time, an network analysis approach that differentiated the gut microbiota in a disease state [Alcoholic Liver Disease] and Healthy

state [Healthy subjects]. The results indicate that by investigating the topological network of taxa identified in different gut microbiota samples, we can essentially predict a person's state of health or identify the bacteria or bacterial relationships that differentiate the disease state from healthy state. Our study suggests that we may be able to use network-based analysis of bacteria in studying the role of microbiota in the pathogenesis, diagnosis and clinical course of the human diseases where dysbiosis has been implicated.

3.2 Background

A microbiome is the entire set of microbial cells, including their genomes and interactions within the various ecological environment of the human body. The human gut contains one of the most densely populated microbial environments known on earth. Over 1×10^{14} microbial cells interact with human cells, indicating the ubiquity and potentially critical importance of such interactions in our bodies [1]. The interactions between digestive tract epithelial and mucosal immune cells with microbial cells are critical to human health. These interactions are involved with the immune system and its responses, metabolic regulation, and digestion and we define these microbial interactions with the host the Metabiome.

In diseased states, these interactions may be altered resulting in disrupted functionality and organ failure [1][2]. In this project we investigate the microbiome with respect to Alcoholic Liver Disease, a serious condition due to heavy alcohol consumption. It is a major health problem in the United States consuming 15% of total

health care dollars [3] and is associated with 20% mortality rate [4]. To date, the impact of chronic alcohol consumption on gut microbiome composition has not been fully studied. New advances in molecular biology have now made it possible to fully interrogate the microbiota in complex biological environment like the human gut. For our studies, we analyzed the microbiome composition in mucosal samples identified using the first two variable regions of the 16S rRNA. This region contains taxa specific signatures that allows identification of the taxa down to the species level. Using the identified species, we follow a network-based approach to model the correlations between the different microbes. We are able to show significant differences amongst the potential microbial correlations within the healthy and diseased patients, using network analysis [5][6], motif finding algorithms [7], and network fitting algorithms [8]. Previous network-based analysis of microbial communities has involved the evolutionary relatedness across species [9][10]. To the best of our knowledge this is the first attempt to investigate microbial taxa networks and diversity within the human gut microbiome of Alcoholic Liver Disease (ALD) patients.

3.3 Methods

Our analysis of the potential correlations amongst microbes within the gut follows the following five steps: (i) identification of abundant species within the patient sample using 16S RNA sequencing, (ii) defining the network collectively for a patient type, (iii) computing network statistics and set operations, (iv) motif finding, and (v) fitting the network to a family of graph models.

Datasets

The data we used for this study was the mucosal microbiome composition from Alcoholic Liver Disease (ALD) and Healthy Control patients produced by Multitag Pyrosequencing (MTPS) of the 16S rRNA of mucosal biopsies from the gut. There were five distinct clinical patient classes we studied, which included the healthy controls, alcoholics with liver disease, alcoholics without liver disease, sober alcoholics with liver disease, and sober alcoholics without liver disease denoted as Healthy, Alcoholic (+), Alcoholic (-), Sober (+), and Sober (-), respectively. These clinical samples were obtained by the Rush University Medical Center in Chicago, Illinois from a total of 51 middle-aged male and female patients (refer to the clinical paper [13]). In clinical terminology, sober patients are those patients that were alcoholics and have stopped drinking due to adverse health effects. In Table 3.1 we report the general statistics of the entire set, including the defined classes, patients, 16S reads, number of family-level taxa identified within the samples.

Taxonomic Identification

Molecular methods that examine the 16S ribosomal RNA (rRNA) gene are routinely used to identify the phylotypes of the bacteria comprising the Human Microbiome [14]. We used the bacterial primers 27F and 355R to amplify the first two variable region in the 16S rRNA and then using the new generation Roche GS-FLX high-throughput

instrument [12] in combination with a multi-tag approach to produce several thousand 16S sequence reads for each sample [11].

We identify the taxa or phylogenetic class for each read by performing a BLAST [15] search against the Ribosomal Database Project (RDP 8.1) [16][17]. In our study we determined the taxa using blast parameters "-e 0.01 –p 0.97 –w 60 –m 8". The RDP is a database of 16S rRNA small sub-unit sequences for the Bacteria and Archaea organisms along with annotation information. The RDP provides a hierarchical phylogenetic categorization for the 16S rRNA sequences. It also provides several web-services related to the identification of taxonomical distribution of microbiome samples.

We identified the taxonomical information for each sequence reads by annotating each read with the best BLAST search hit to the RDP database. We used the fifth level in the RDP 8 taxonomic hierarchy, the family level, for the taxa assignment as it provided sufficient resolution of the species in the sample microbiomes.

We computed the taxa distribution for each set of 16S rRNA reads obtained from the biopsy of a patient and filtered out any taxa that have less than 1% normalized abundance. Every patients microbiome is slightly different and much of this difference lies in the low abundant taxa below 1% [1]. Thus, filtering the data also helps identify trends in the microbiome above this background variation. We also experimented with using the Bayesian-based RDP10 [17] classifier to determine the taxonomical information and found the differences in the final analysis not as significant.

Network Modeling

We modeled a networks for each the five patient types. An undirected graph G = (V,E) is used to represent the potential correlations between the different taxa within patient groups. The set of vertices V represents the set of identified taxa, and an edge $E_{i,j}$ exists between vertices V_i and V_j if both these taxa were found together above a defined threshold in the reads obtained from a patient's sample. The edge $E_{i,j}$ indicates a potential correlation between the bacterial species. We can also compute the weight of the edge $E_{i,j}$ by counting either the number of patient samples where both these taxa were present and abundant, or by using the abundance information of the interacting partners. For the purpose of this study, we neglect the edge weights and use an undirected, un-weighted graph representation. We will investigate weighted graphs in the near future.

The network models are visualized in the popular open-source Cytoscape [5] software, developed for visualization of protein-protein interaction networks obtained from high throughput proteomic studies. Figures 3.1 (a)-(e) shows the network models for the five patient classes – Healthy, Alcoholics (+), Alcoholics (-), Sober (+), and Sober (-), respectively. We observed that the healthy group network is denser, which we measure in our network connectivity analysis, in comparison to the other networks reflected by the larger number of taxa that co-occur. There are distinctions in the different networks, which can be correlated to different patient classes.

Network Statistics and Operations

We computed several global and local network properties to distinguish between the different patients. In Table 2 we report some of the computed statistics for the different networks. The global network properties that were calculated were the degree distribution, the average network diameter, and the average clustering coefficient. The degree of a node represents the number of neighbours for a given node. The average network diameter is the average shortest path length over all pairs of nodes in the network. The clustering coefficient of a node z in a network, is the probability that two nodes x and y which are connected to the node z are themselves connected. The average of this over all nodes z of a network is the clustering coefficient of the network [8]. We used GraphCrunch to quickly and efficiently compute these network statistics.

We also performed network operations that involved overlapping the different class networks in order to find the intersection, union and difference of particular nodes and edges (correlations). We also computed a core network i.e., the intersection or the common set of the entire five patient networks that may represent the "core" microbiome. This network is shown in Figure 1(f) (See Results for more details.)

Motif Detection

Motifs within networks are frequently occurring sub-graphs generally made up of small number of vertices or nodes. Motifs may reveal important structural principles or a unique signature in complex network models [18][19]. We attempted to use a motif-finding algorithm to differentiate between the various patient-derived networks.

However, searching through these networks is a NP-complete problem. FANMOD is one of the heuristic-based motif finding algorithm that has proven to be efficient in determining small motifs within biological networks [7] in comparison to other motif finding algorithms like MAVISTO [20]. FANMOD also determines the significance of a discovered motif by counting the occurrence of the identified motif in a set of randomized networks (generated with the same degree distribution).

Network Fitting

We also compared our generated network models to five sets of well-studied family of random graph models, which are the Erdös–Rényi [21], Erdös–Rényi with same degree distributions [22], Scale-free Barabasi-Albert [23], N-dimensional geometric [24], and Stickiness [25] denoted by ER, ER-DD, SF-BA, GEO-3D, and STICKY, respectively. All of these random null models are common in real networks, such as social, protein interaction, and World Wide Web networks [19]. We used GraphCrunch [8], to assess how well our networks fit some of these random graph models. GraphCrunch ensures that the different random models have the number of nodes and edges that are within 1% of the input network.

Since the comparison of a pair of networks leads to subgraph isomorphism (i.e., NP-complete), GraphCrunch uses a set of heuristics for computing the local sub-graph dependent statistics. In particular, the RGF-distance and GDD-agreement are local measures of structural similarities between two networks. These measures are based on 3-5 node graphlets, which are small-connected non-isomorphic induced sub-graphs of large

networks. RGF-distance compares the frequencies of the appearance of all 3–5-node graphlets in two networks (real and random), while the GDD-agreement generalizes the notion of the degree distribution to the spectrum of graphlet degree distributions.

3.4 Results and Discussion

Network Topology

In Table 3.2 we present the average correlation coefficient and the average diameter for the five patient-derived networks. In Figure 3.2, we also present the cumulative distribution function (CDF) of the degree distributions per node (taxa) for the five defined categories. Nodes with a higher degree indicate the specific taxa potentially having interaction with more taxa. From the Figure 3.3 we observe the difference between the "Healthy", "Sober" and "Alcoholic" classes. The x-axis is the degrees per nodes in each of the networks, while the y-axis is the percentage of interactions that occur relative to each class network. As an example, the plot shows that 60% of the species interact with less than 50, 40, and 30 (approximately) species in the Healthy, Sober, and Alcoholic classes, respectively. This indicates a potential larger number of interactions within the Healthy class. The graphs did not distinguish between the classes with or without liver disease.

Core Microbiota

Using Cytoscape [5] we compute the intersections and differences amongst the edges or taxa relationships across the different networks. We compute a core network,

which presents the common correlations between nodes (shown in Figure 1(f)). This network has 68 vertices and 326 edges and may potentially represent the "core" microbiome relationships in our gut. In Figure 3.3(b) we show the Healthy network with the "core" network highlighted in yellow. We can see that there are significantly many edges (1057) and vertices (112) involved in the non-core part of the network.

We also compared the Healthy network to four combinations of non-healthy networks – (i) Alcoholics (+) and Alcoholics (-), (ii) Sober (+) and Sober (-), (iii) Alcoholics (+) and Sober (+), and (iv) Alcoholics (-) and Sober (-). In Figure 3(a) we show the number of common and distinct interactions between the healthy and the four union networks. We observe that the Healthy network shares the most potential interactions with the Sober (Sober (+) and Sober (-)) and no Liver Disease (Alcoholics (-) and Sober (-)) networks. We also note that there are significant differences (distinction) between Healthy and the other classes.

Motif Finding

We used the Fanmod motif detection algorithm to search for three, four, and five vertex sub-graphs, while generating 1,000 random models with a locally constant number of bidirectional edges and 3 exchanges and tries per edge for computing the statistically significant motifs. As in other biological networks [26] we found the feed-forward 3-node motif to be present in all the patient-derived networks with at least a 20% frequency. We present the most significant (highest z-scores) 4-node motifs discovered across the five networks in Figure 4, along with the frequency of occurrence in the

network and the random networks. We found a total of four motifs in the Healthy and Sober classes, and 3 motifs in the Alcoholics classes that were significant. The first two motifs are much more abundant in the Healthy and the Sober (-) classes when compared to the others. On the other hand the third motif is more abundant in the Alcoholics and the Sober (+) classes. These results reflect that there may be specific patterns of interactions existing within different patient classes.

Network Model Fitness

Using GraphCrunch [8], we generated 30 instances of all the five random models (ER, ER-DD, GEO-3D, SF-BA, and Sticky) for each of our patient-derived networks. We compared the real networks to these set of random instances using a set of both global and local properties of networks that were described earlier to find the best-fit model. In Figures 5(a) and 5(b) we plot the GDD-agreement (arithmetic mean) and RGF-distance between our patient networks and derived random models, respectively. Analyzing Figure 5(a), we notice that the trend is similar for the families of random models, though the agreement for the Healthy and Sober (-) networks is the lowest, and highest for the Sober (+) class. In Figure 5(b), we see that the STICKY and ER-DD model fits best with our patient-derived networks, since we show a distance measure.

3.5 Conclusions

In this work we presented an approach to model the potential interactions within the microbiome using a network-based approach. We used a range of network analysis

tools to characterize the modeled networks for different classes of patients. In particular, we analyzed the 16S rRNA sequences in the gut microbiome from healthy patients and alcoholic patients (with or without liver disease).

We found a core set of correlations (interactions or relationships) that exist in all of these networks that may suggest that there is a core set of metabolic or immune functions that are provided by the human gut microbiome. We also found potential interactions that only occur in the Healthy patients reflecting that these relationships may be crucial for good health and that disruption of these interactions may lead to instability in the ecosystem or disease. We also found that the Healthy network was denser, with a higher degree distribution per node and a greater number of motifs present. One potential hypothesis that needs evaluation is that the healthy gut microbiome is more robust and adaptable to changing environmental conditions. The comparison against specific random null model networks gives us further insight in the topology of these networks.

To test the robustness and significance of our analysis we need to apply the same technique to larger alcohol liver disease datasets. We also would like to test it for other potential diseases, and for microbial communities in other species or even environments. In the future we aim to use a weighted graph representation and validate the abundance of interactions using metabolomic, metaproteomic, and metagenomic studies.

3.6 References

1. P.J. Turnbaugh, R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, and J.I. Gordon, "The human microbiome project.," Nature, vol. 449, Oct. 2007, pp. 804–

810.

2. F. Backhed, R.E. Ley, J.L. Sonnenburg, D.A. Peterson, and J.I. Gordon, "Host-Bacterial Mutualism in the Human Intestine," Science, vol. 307, 2005, pp. 1915-1920.
3. P.G. O'Connor and R.S. Schottenfeld, "Patients with Alcohol Problems," N Engl J Med, vol. 338, 1998, pp. 592-602.
4. B.S. Maher, M.L. Marazita, W.N. Zubenko, D.G. Spiker, D.E. Giles, B.B. Kaplan, and G.S. Zubenko, Genetic segregation analysis of recurrent, early-onset major depression: Evidence for single major locus transmissio, 200.
5. P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," Genome Research, vol. 13, 2003, pp. 2498-2504.
6. S. Brohee, K. Faust, G. Lima-Mendez, O. Sand, R. Janky, G. Vanderstocken, Y. Deville, and J. van Helden, "NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways," Nucl. Acids Res., vol. 36, 2008, pp. W444-451.
7. S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," Bioinformatics, vol. 22, 2006, pp. 1152-1153.
8. T. Milenkovic, J. Lai, and N. Przulj, "GraphCrunch: A tool for large network analyses," BMC Bioinformatics, vol. 9, 2008, p. 70.
9. J. Xu, M.A. Mahowald, R.E. Ley, C.A. Lozupone, M. Hamady, E.C. Martens, B.

Henrissat, P.M. Coutinho, P. Minx, P. Latreille, H. Cordum, A. Van Brunt, K. Kim, R.S. Fulton, L.A. Fulton, S.W. Clifton, R.K. Wilson, R.D. Knight, and J.I. Gordon, "Evolution of Symbiotic Bacteria in the Distal Human Intestine," PLoS Biol, vol. 5, 2007, p. e156.

10. R.E. Ley, C.A. Lozupone, M. Hamady, R. Knight, and J.I. Gordon, "Worlds within worlds: evolution of the vertebrate gut microbiota," Nat Rev Micro, vol. 6, Oct. 2008, pp. 776-788.
11. G. Spear, M. Sikaroodi, M. . Zariffard, A. Landay, A. French, and P. Gillevet, "Comparison of the Diversity of the Vaginal Microbiota in HIV-Infected and HIV-Uninfected Women with or without Bacterial Vaginosis," The Journal of Infectious Diseases, vol. 198, 2008, pp. 1131-40.
12. M. Ronaghi, "Real-time DNA sequencing using detection of pyrophosphate release," Anal. Biochem., vol. 242, 1996, pp. 84-89.
13. A. Kesharvaian and Patrick M Gillevet, "Clinical Study of Human Alcohol Liver Disease," Hepatology (Under Review).
14. K. Wilson and R.B. Blitchington, "Human colonic biota studied by ribosomal DNA sequence analysis," Appl Environ Microbiol, vol. 62, 1996, pp. 2273-2278.
15. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic Local Alignment Search Tool," Journal of Molecular Biology, vol. 215, 1990, pp. 403-410.
16. J.R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, and J.M. Tiedje, "The Ribosomal Database Project: improved alignments and new tools for rRNA

analysis.," Nucleic Acids Res, Nov. 2008.

17. J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, G.M. Garrity, and J.M. Tiedje, "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis.," Nucleic Acids Res, vol. 33, Jan. 2005, pp. D294–D296.
18. R. Milo, S.S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," Science, vol. 298, 2002, pp. 824 - 827.
19. M.E.J. Newman, "The structure and function of complex networks," SIAM Review, vol. 45, 2003, pp. 167 - 256.
20. F. Schreiber and H. Schwobbermeyer, "MAVisto: a tool for the exploration of network motifs," Bioinformatics, vol. 21, 2005, pp. 3572-3574.
21. P. Erdos and A. Renyi, "On random graphs," Publicationes Mathematicae, vol. 6, 1959, pp. 290 - 297.
22. M. Molloy and B. Reed, "A critical point of random graphs with a given degree sequence," Random Structures and Algorithms, vol. 6, 1995, pp. 161 - 180.
23. A.L. Barabasi and R. Albert, "Emergence of scaling in random networks," Science, vol. 286, 1999, pp. 509 - 512.
24. M. Penrose, Oxford University Press, 2003.
25. N. Przulj and D. Higham, "Modelling protein-protein interaction networks via a stickiness index," Journal of the Royal Society Interface, vol. 3, 2006, pp. 711 - 716.

26. S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," Nature Genetics, vol. 31, 2002, pp. 64 - 68.

3.7 Figures

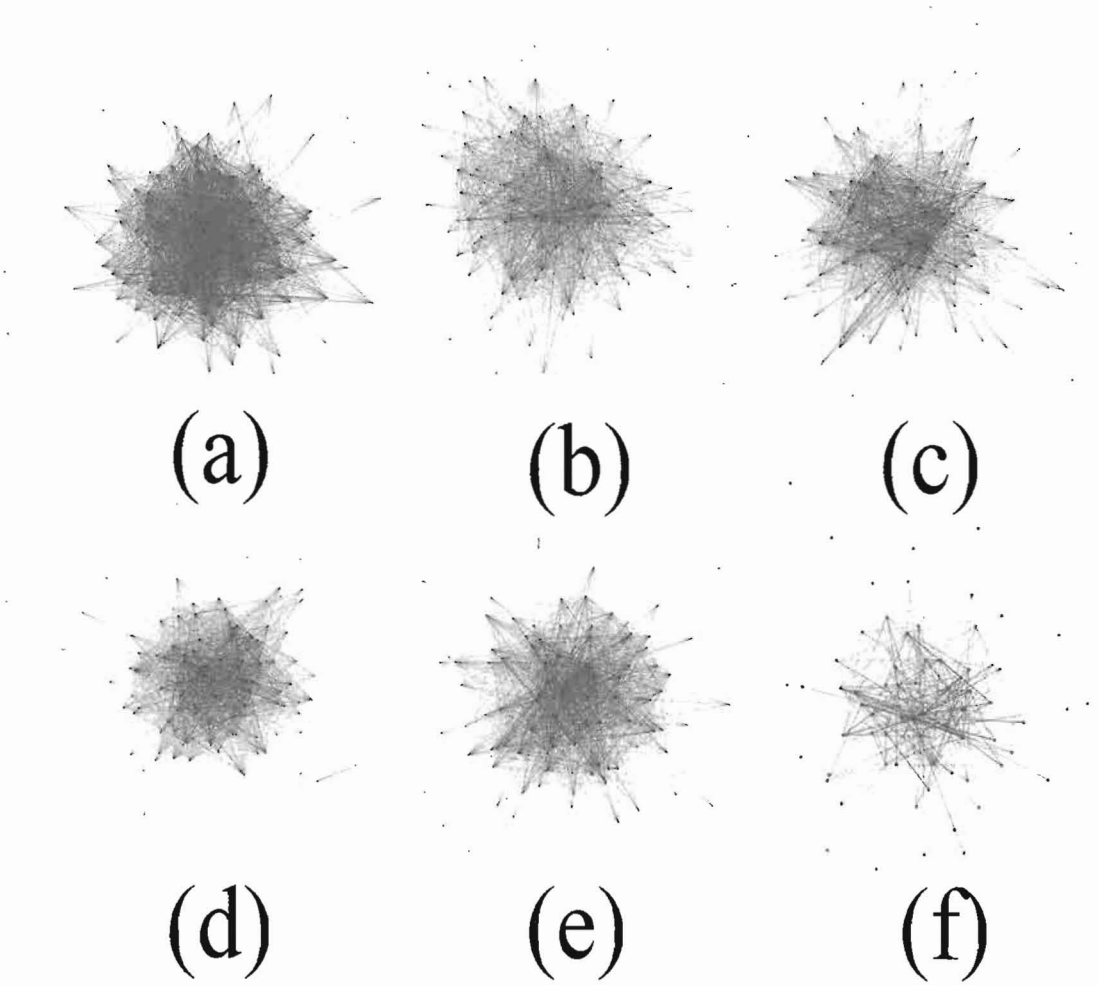


Figure 3.1: Network Representation for the (a) Healthy (b) Alcoholics (+) (c) Alcoholics (-) (d) Sober (+) (e) Sober (-), and (f) Core Classes. These networks are visualized using the Cytoscape network modeling tool.

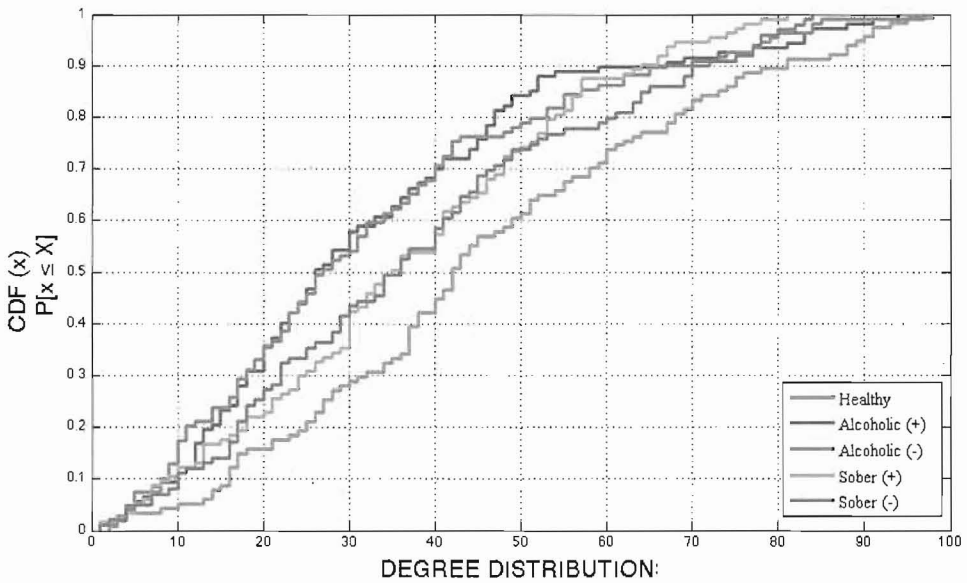
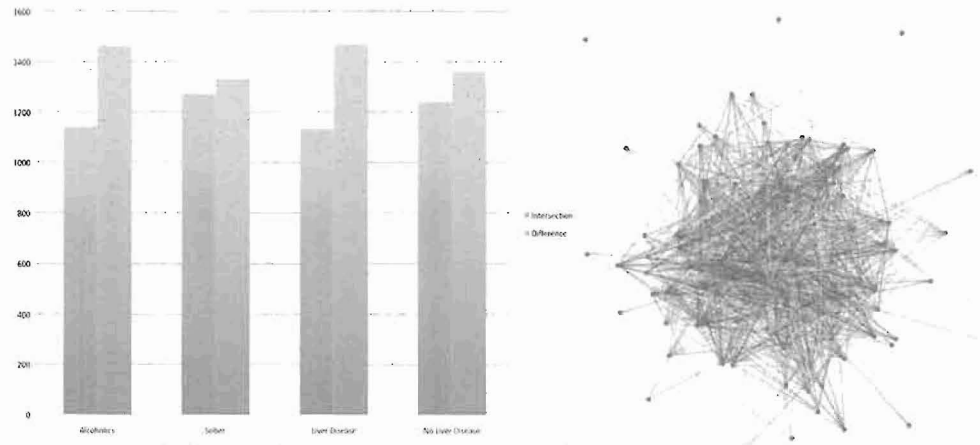


Figure 3.2: Degree Distribution (Cummulative Distribution Function) for the five patient classes.



(a) (b)

Figure 3.3: Network Operations. (a) Intersection and Difference statistics of union networks (Alcoholics, Sober, Liver Disease, and No Liver Disease) with respect to the Healthy class. (b) Superimposition of the Core Network (yellow) over the Healthy Network in blue.

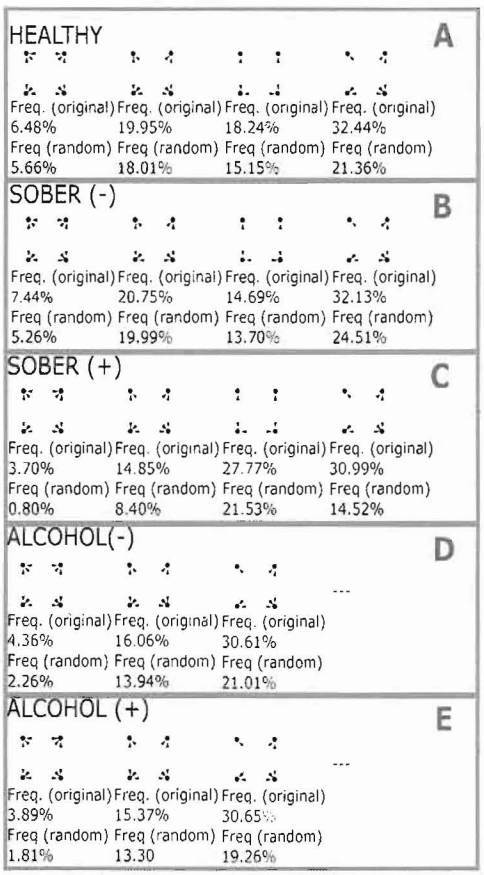
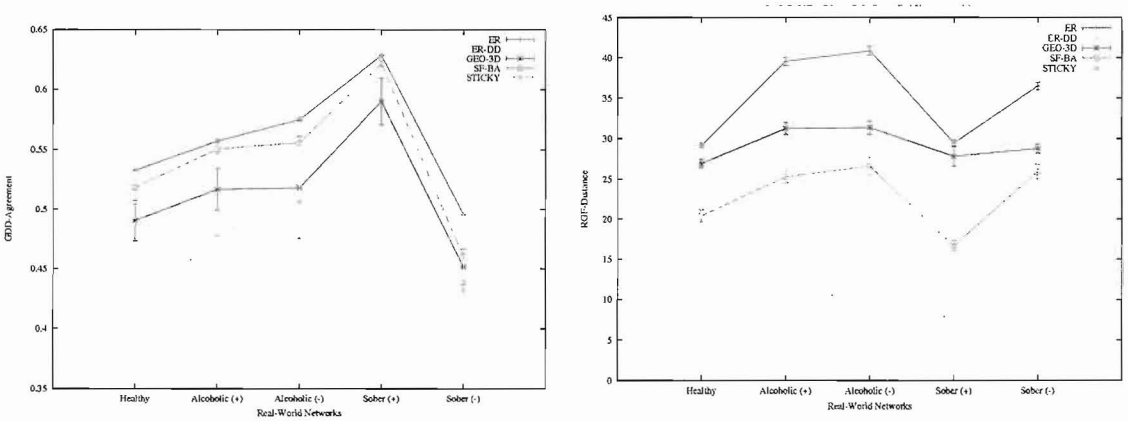


Figure 3.4: Statistically significant 4-vertex motifs detected across the five network using the FANMOD motif finding algorithm.



(a) (b)

Figure 3.5: Local metrics (a) GDD-Agreement and (b) RGF-Distance comparing the patient-derived networks against the ER, ER-DD, GEO-3D, SF-BA, and STICKY graph models

3.8 Tables

Table 3.1. General Dataset Statistics.

Class	# Patients	# Reads	#Reads per Patient	#Taxa Identified
Healthy	10	8058	805	114
Alcohol (+)	8	13025	1628	107
Alcohol (-)	9	10016	1112	109
Sober (+)	11	13694	1244	113
Sober (-)	13	19240	1480	99
Total	51	64033	1255	116

Table 3.2. Global Network Properties.

Class	Total Interactions	Average Diameter	Average Clustering Coefficient
Healthy	2604	1.608	0.636
Alcohol (+)	1726	1.709	0.626
Alcohol (-)	1781	1.719	0.632
Sober (+)	2052	1.711	0.538
Sober (-)	1867	1.638	0.688

CHAPTER 4

Summary

4.1 Summary

The studies presented in the thesis are a step towards the broader objectives and goals of the Human Microbiome Project. It has been proven that in diseased individuals the interface between microbial and human cells is transformed potentially allowing different types of interactions to occur that are directly correlated or associated with a persons state of the health. In this particular study we implemented a network-based approach and discovered possible distinct interactions or relationships amongst different bacterial taxa related to a persons class or clinical health state (ie. healthy, diseased). Furthermore, prior to our network approach, we have successfully introduced a novel analysis pipeline for the taxonomic identification of 16S rRNA sequences obtained from multi-tag pyrosequencing samples. Our results have lead us to affirm that there is a means to classify and diagnose individuals based on their microbiome properties.

4.2 Future Direction

Nevertheless, there is a need for more implementation to confirm our conclusions. In the future, the study needs to incorporate and be validated on other diseased datasets. More importantly, the focus should be concentrated on actual transcription that occurs in

the bacteria. For example, add a deeper layer to the networks that include the actual metabolic pathways and metabolites that are present. This type of depth will give us greater insight on the types of interactions that are present in a person's microbiome.

CURRICULUM VITAE

Ammar S. Abbas Naqvi. With the advent of Bioinformatics and Computational Biology, Ammar has had the opportunity to combine his social conscience with his intellectual ambitions. He finds it very gratifying to be a part of solutions that work as a catalyst for human advancement. He was a part of the first graduating class in the field at Rochester Institute of Technology and now is upon completion with his Masters of Science at George Mason University. In addition to his educational background he has also worked in the field as a researcher at the University of California at San Diego, the Salk Institute of Biological Studies, and at George Mason University. Furthermore, he plans on continuing his education at the doctoral level in order to further contribute to the field at a higher level of sophistication. He hopes to combine his knowledge of programming with his passion to study biology in order to produce the most applicable and practical scientific tools, which will be vital in furthering science and research.