Published in final edited form as:
Chem Biodivers. 2010 May ; 7(5): 1040–1050. doi:10.1002/cbdv.200900324.

Network-based Modeling of the Human Gut Microbiome

Ammar Naqvi[c),d)], **Huzefa Rangwala**[a,c,d)], **Ali Keshavarzian**[e)], and **Patrick Gillevet**[b),c),d)]
Patrick Gillevet: pgilleve@gmu.edu
[a)]Department of Computer Science, George Mason University, Fairfax, VA, 22030 USA

[b)]Department of Environmental Science and Policy, George Mason University, Fairfax, VA, 22030 USA

[c)]Microbiome Analysis Center, Manassas, VA, 20110 USA, phone: 703-993-1057

[d)]Bioinformatics and Computational Biology Department, Manassas, VA, 20110 USA

[e)]Department of Medicine, Section of Gastroenterology, Rush University Medical Center, Chicago, IL, 60612 USA

Abstract

In this paper we used a network-based approach to characterize the microflora abundance in colonic mucosal samples and correlate potential interactions between the identified species with respect to the healthy and diseased states. We analyzed the modelled 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. This study has demonstrated, for the first time, an approach that differentiated the gut microbiota in Alcoholic subjects and Healthy subjects using topological network analysis of the gut microbiome.

Introduction

It has been suggested that the human gut contains between 500 and 1000 species with fewer than 50 species making up the bulk of the microbial biomass [1]. This microbial community has been defined as the Human gut microbiome and we define the microbial interactions with the host as the *Metabiome*. Furthermore, the microbiome composition varies significantly between healthy individuals raising the question whether there is a core microbiome that provides core functionality. In diseased states, this core microbiome may be altered shifting the functionality of the Metabiome [1][2].

In this project, we investigate the microbiome with respect to Alcoholic Liver Disease (ALD), 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 [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. In our studies, we analyzed the microbiome composition in mucosal samples identified using the first two variable regions of the 16S rRNA as this region contains taxa specific signatures that allows identification of the taxa down to the species level. Using the

Correspondence to: Patrick Gillevet, pgilleve@gmu.edu.

Authors' contributions
AN developed the code, scripts and programs, while HR and PG contributed to the theoretical aspects of the study. HR and PG also provided support by assisting in the analysis. AK provided clinical expertise, developed the experimental framework for the study, and provided the mucosal biopsies. All authors wrote, read the manuscript, and approved the manuscript for final submission.

identified species, we followed a network-based approach to model the correlations between the different microbes. We were 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 ALD patients.

Results and Discussion

We plotted the network representation for the diseased classes (Figure 1). We observed that the healthy group network (Figure 1a) is more dense with respect to the other networks as reflected by the larger number of taxa that co-occur suggesting that the Healthy subjects have a more robust network that can respond or adapt easier to environmental change.

Network Topology

In Table 1 we present the average correlation coefficient and the average diameter for the five patient-derived networks. In Figure 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 that potentially have interaction with more taxa. From Figure 2 we observed the difference between the “Healthy”, “Sober” and “Alcoholic” classes. 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 computed the intersections and differences amongst the edges or taxa relationships across the different networks (Figure 3).

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.

We also computed a core network, which presents the common correlations between nodes, and superimposed this on the Healthy Network (Figure 3b). This network has 68 vertices and 326 edges and may potentially represent the “core” microbiome relationships in our gut. We can see that there are significantly many edges (1057) and vertices (112) involved in the non-core part of the network.

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 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, as computed by the RGD-distance measure.

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 modelled 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.

Experimental Part

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 that 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 collected by AK at Rush University Medical Center in Chicago, Illinois from a total of 51 middle-aged male and female patients [13]. We define sober patients are those patients that were alcoholics and have stopped drinking due to adverse health effects. In Table 2 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 (5′-AGAGTTTGATCM TGGCTCAG-3′) and 355R (5′-GCTGCCTCCCGTAGGAGT-3′) 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 identified 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"*. We discovered that a word size of 60 was the most efficient and accurate for the identification of signatures in the rRNA and associating them with a taxa. 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 modelled a networks for each the five patient types. An undirected graph $\mathbf{G = (V,E)}$ was used to represent the potential correlations between the different taxa within patient groups. The set of vertices $\mathbf{V}$ represents the set of identified taxa, and an edge $\mathbf{E_{i,j}}$ exists between vertices $\mathbf{V_i}$ and $\mathbf{V_j}$ if both these taxa were found together above a defined threshold in the reads obtained from a patient's sample. In this study, we defined that threshold as the average abundance for that taxa in each class. The edge $\mathbf{E_{i,j}}$ indicates a potential correlation between the bacterial species. We can also compute the weight of the edge $\mathbf{E_{i,j}}$ 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 pursue the analysis of 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 1(a)–(e) shows the network models for the five patient classes – Healthy, Alcoholics (+), Alcoholics (−), Sober (+), and Sober (−), respectively.

Network Statistics and Operations

We computed several global and local network properties to distinguish between the different patients. In Table 1 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.

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 more efficient in determining small motifs within biological networks [7] in contrast 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] models 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.

Acknowledgments

AN, AK, and PG were supported by the NIH grant number RO-1 AA 013745. HR is supported by a bioengineering startup grant provided by the Volgenau School of Information Technology and Engineering at George Mason University and NSF project number IIS-0905117. We would like to thank John Greffenstette for his initial discussion on the use of Network Analysis to study the Human Microbiome.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. Nature. 2007; 449:804–810. [PubMed: 17943116]
2. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Science. 2005; 307:1915–1920. [PubMed: 15790844]
3. O'Connor PG, Schottenfeld RS. N Engl J Med. 1998; 338:592–602. [PubMed: 9475768]
4. Maher BS, Marazita ML, Zubenko WN, Spiker DG, Giles DE, Kaplan BB, Zubenko GS. The American Journal of Drug and Alcohol Abuse. 2002; 28(4):711–731. [PubMed: 12492266]
5. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Genome Research. 2003; 13:2498–2504. [PubMed: 14597658]
6. Brohee S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y, van Helden J. Nucl Acids Res. 2008; 36:W444–451. [PubMed: 18524799]
7. Wernicke S, Rasche F. Bioinformatics. 2006; 22:1152–1153. [PubMed: 16455747]
8. Milenkovic T, Lai J, Przulj N. BMC Bioinformatics. 2008; 9:70. [PubMed: 18230190]
9. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P, Cordum H, Van Brunt A, Kim K, Fulton RS, Fulton LA, Clifton SW, Wilson RK, Knight RD, Gordon JI. PLoS Biol. 2007; 5:e156. [PubMed: 17579514]
10. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Nat Rev Micro. 2008; 6:776–788.
11. Spear G, Sikaroodi M, Zariffard M, Landay A, French A, Gillevet P. The Journal of Infectious Diseases. 2008; 198:1131–40. [PubMed: 18717638]
12. Ronaghi M. Anal Biochem. 1996; 242:84–89. [PubMed: 8923969]
13. Gillevet, Patrick M.; Mutlu, Ece A.; Rangwala, Huzefa; Sikaroodi, Masoumeh; Naqvi, Ammar; Engen, Phillip A.; Kwasny, Maryto; Lau, Cynthia; Keshavarzian, Ali. Alcoholism: Clinical and Experimental Research. 2010 (submitted).
14. Wilson K, Blitchington RB. Appl Environ Microbiol. 1996; 62:2273–2278. [PubMed: 8779565]
15. Altschul S, Gish W, Miller W, Myers E, Lipman D. Journal of Molecular Biology. 1990; 215:403–410. [PubMed: 2231712]
16. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. Nucleic Acids Res. 2008 10.1093.
17. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. Nucleic Acids Res. 2005; 33:D294–D296. [PubMed: 15608200]
18. Milo R, Shen-Orr SS, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002; 298:824–827. [PubMed: 12399590]
19. Newman MEJ. SIAM Review. 2003; 45:167– 256.
20. Schreiber F, Schwobbermeyer H. Bioinformatics. 2005; 21:3572–3574. [PubMed: 16020473]
21. Erdos P, Renyi A. Publicationes Mathematicae. 1959; 6:290– 297.
22. Molloy M, Reed B. Random Structures and Algorithms. 1995; 6:161– 180.
23. Barabasi AL, Albert R. Science. 1999; 286:509– 512. [PubMed: 10521342]

24. Penrose, M. Random Geometric Graphs. Oxford University Press; 2003.
25. Przulj N, Higham D. Journal of the Royal Society Interface. 2006; 3:711– 716.
26. Shen-Orr SS, Milo R, Mangan S, Alon U. Nature Genetics. 2002; 31:64– 68. [PubMed: 11967538]

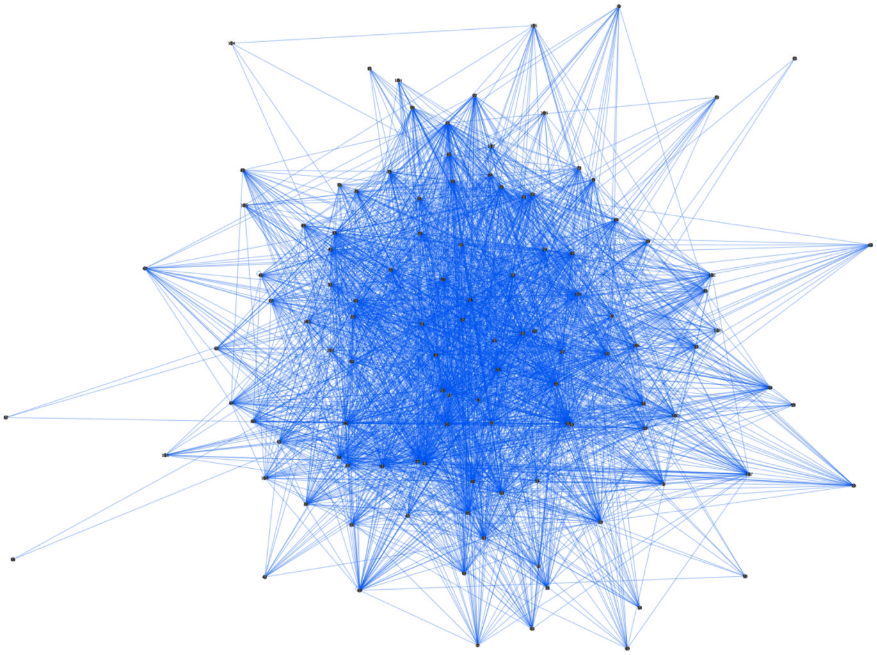


Figure 1a

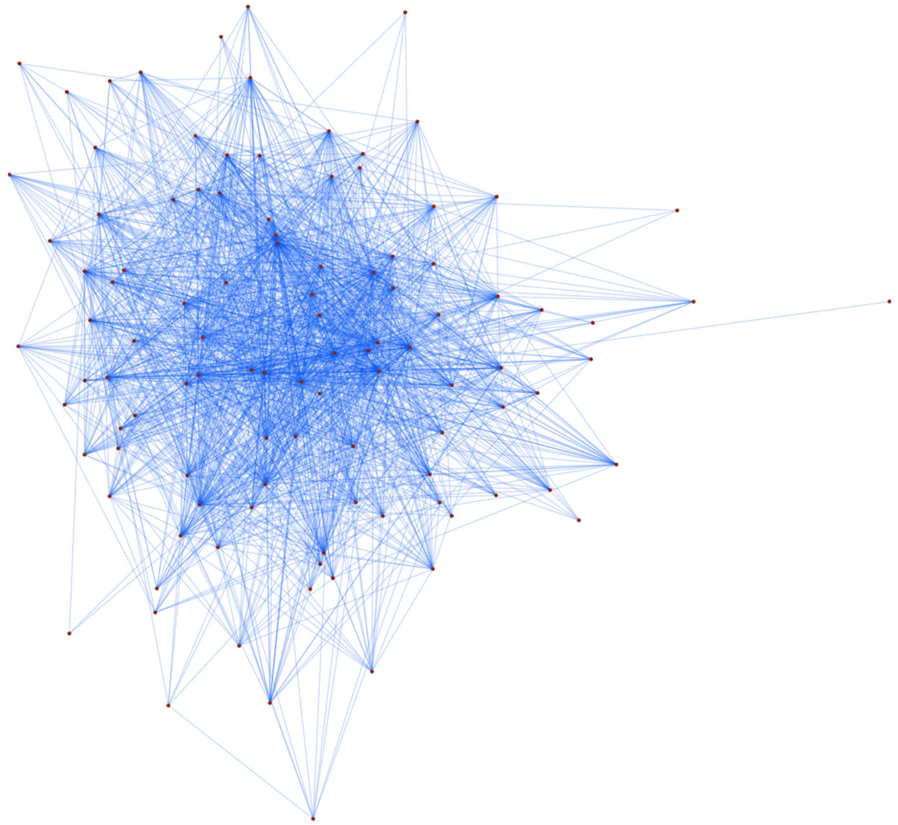


Figure 1b

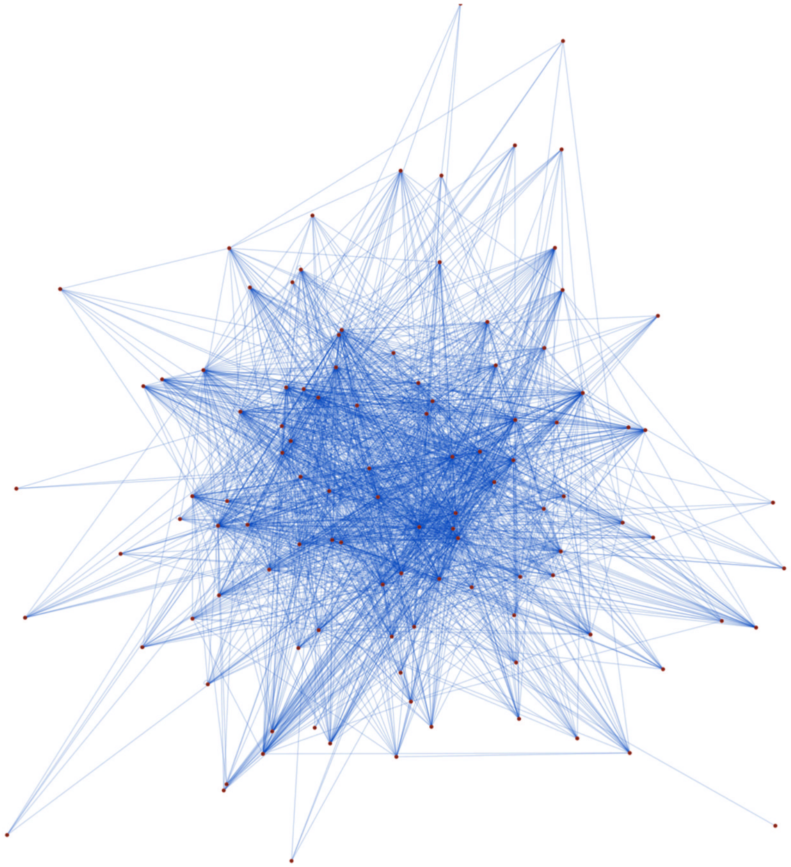


Figure 1c

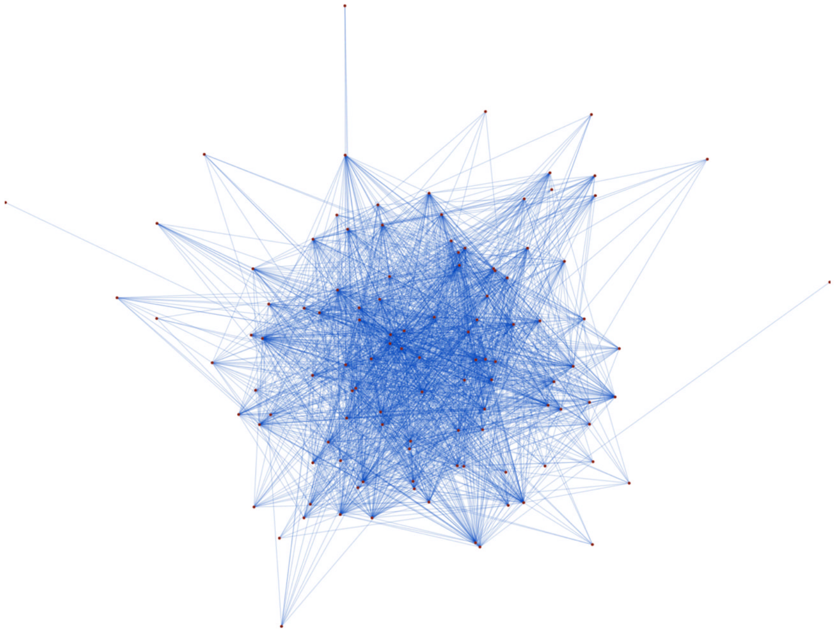


Figure 1d

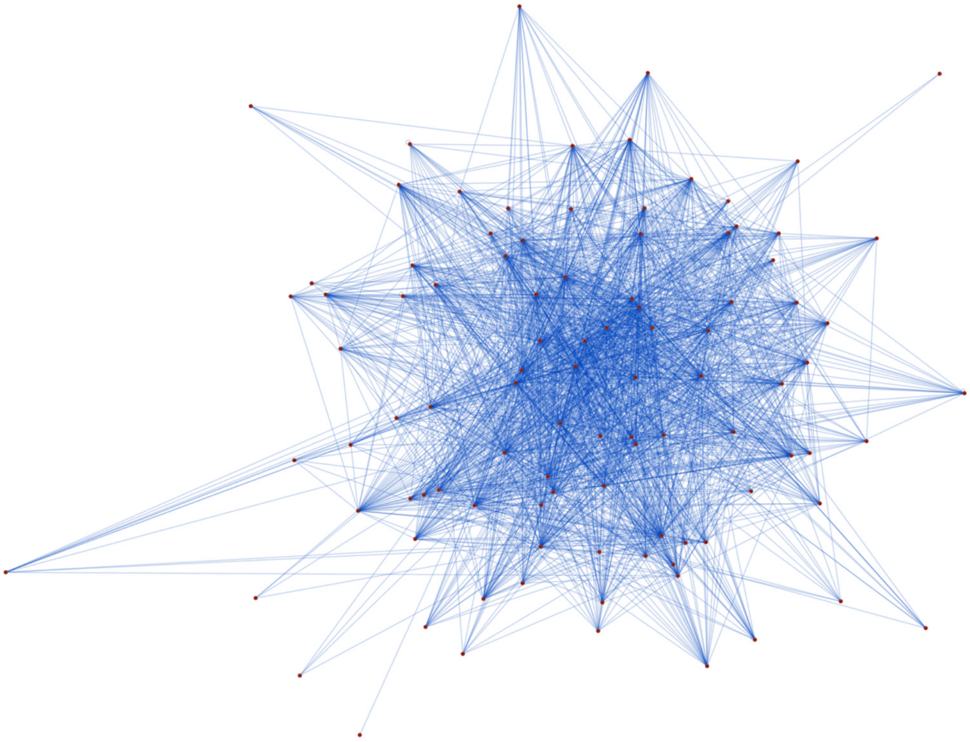


Figure 1e

Figure 1. Network Representation of Gut Microbiome

Figure 1 depicts the Network representation for the (a) Healthy (b) Alcoholics (+) (c) Alcoholics (−) (d) Sober (+) (e) Sober (−) microbiomes. These networks are visualized using the Cytoscape network modeling tool.

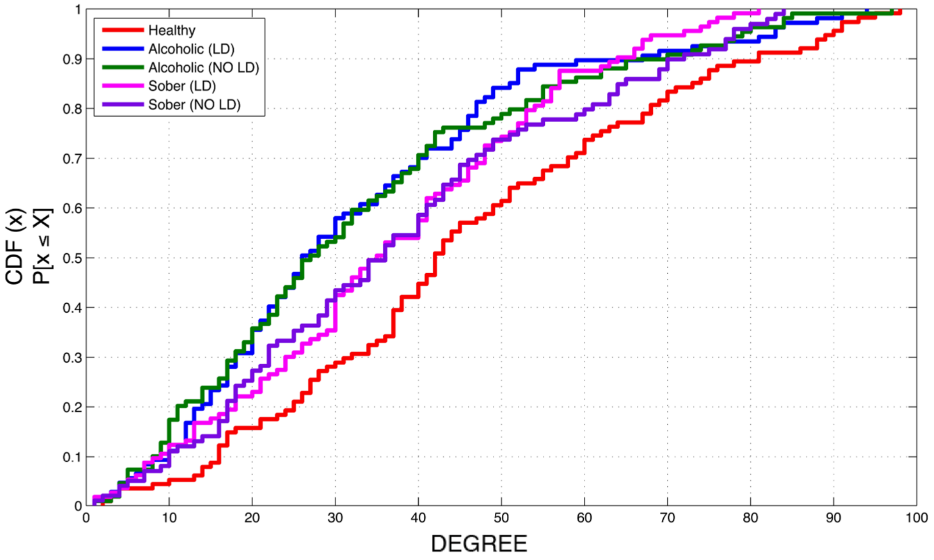


Figure 2. Degree Distribution Graph

Graph of the Cummulative Distribution Function for the five patient classes.

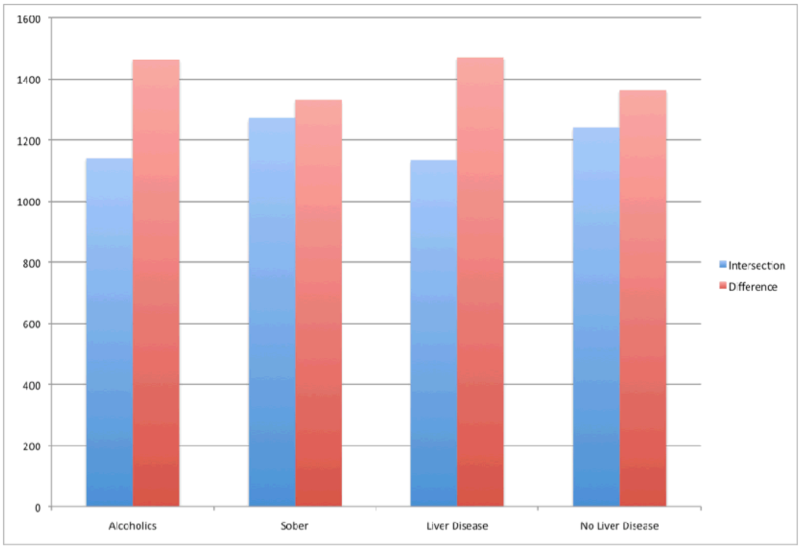

1600
1400
1200
1000
800
600
400
200
0
Alccholics
Sober
Liver Disease
No Liver Disease
Intersection
Difference

Figure 3a

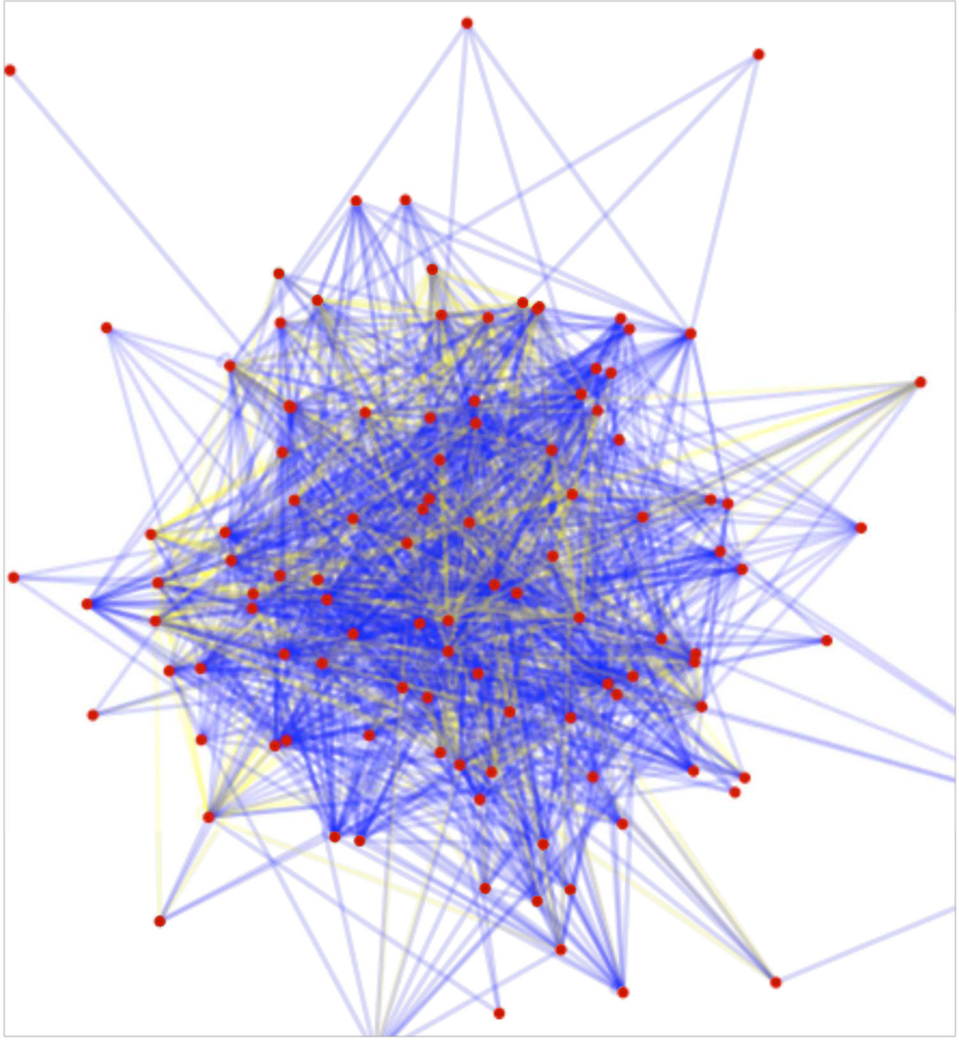
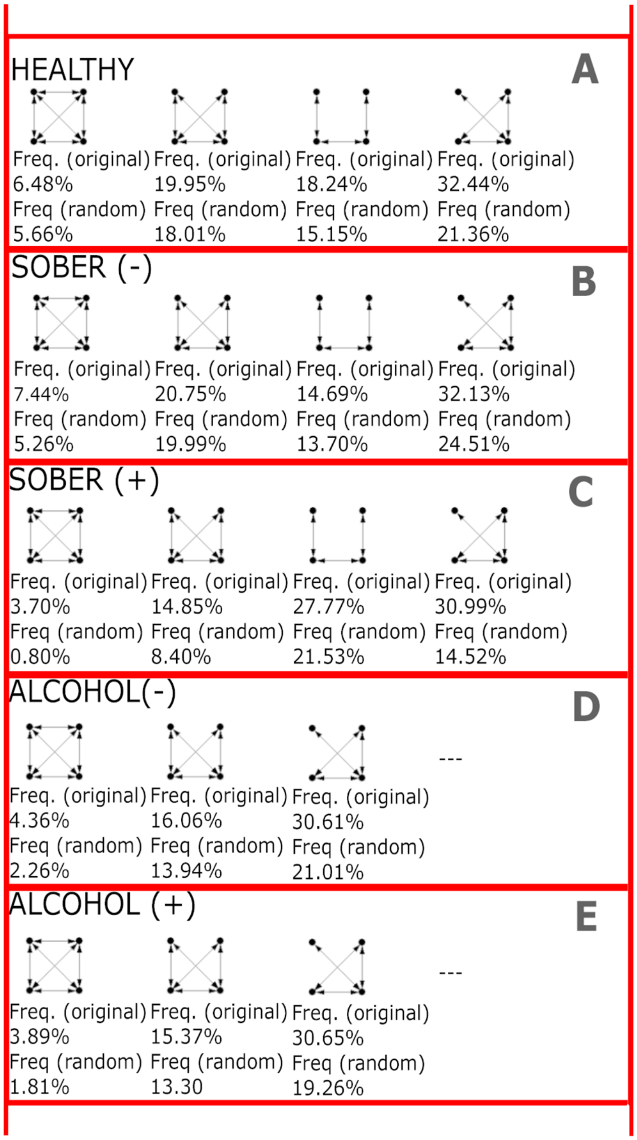


Figure 3b

Figure 3. Network Operations

Figure 3a depicts the intersection and difference statistics of union networks (Alcoholics, Sober, Liver Disease, and No Liver Disease) with respect to the Healthy class. Figure 3b depicts the superimposition of the Core Network (yellow) over the Healthy Network in blue.



HEALTHY				A
Freq. (original) 6.48%	Freq. (original) 19.95%	Freq. (original) 18.24%	Freq. (original) 32.44%	
Freq (random) 5.66%	Freq (random) 18.01%	Freq (random) 15.15%	Freq (random) 21.36%	
SOBER (-)				B
Freq. (original) 7.44%	Freq. (original) 20.75%	Freq. (original) 14.69%	Freq. (original) 32.13%	
Freq (random) 5.26%	Freq (random) 19.99%	Freq (random) 13.70%	Freq (random) 24.51%	
SOBER (+)				C
Freq. (original) 3.70%	Freq. (original) 14.85%	Freq. (original) 27.77%	Freq. (original) 30.99%	
Freq (random) 0.80%	Freq (random) 8.40%	Freq (random) 21.53%	Freq (random) 14.52%	
ALCOHOL(-)				D
Freq. (original) 4.36%	Freq. (original) 16.06%	Freq. (original) 30.61%	---	
Freq (random) 2.26%	Freq (random) 13.94%	Freq (random) 21.01%		
ALCOHOL (+)				E
Freq. (original) 3.89%	Freq. (original) 15.37%	Freq. (original) 30.65%	---	
Freq (random) 1.81%	Freq (random) 13.30	Freq (random) 19.26%		

Figure 4. Motif Detection

Table of statistically significant 4-vertex motifs detected across the five network using the FANMOD motif finding algorithm.

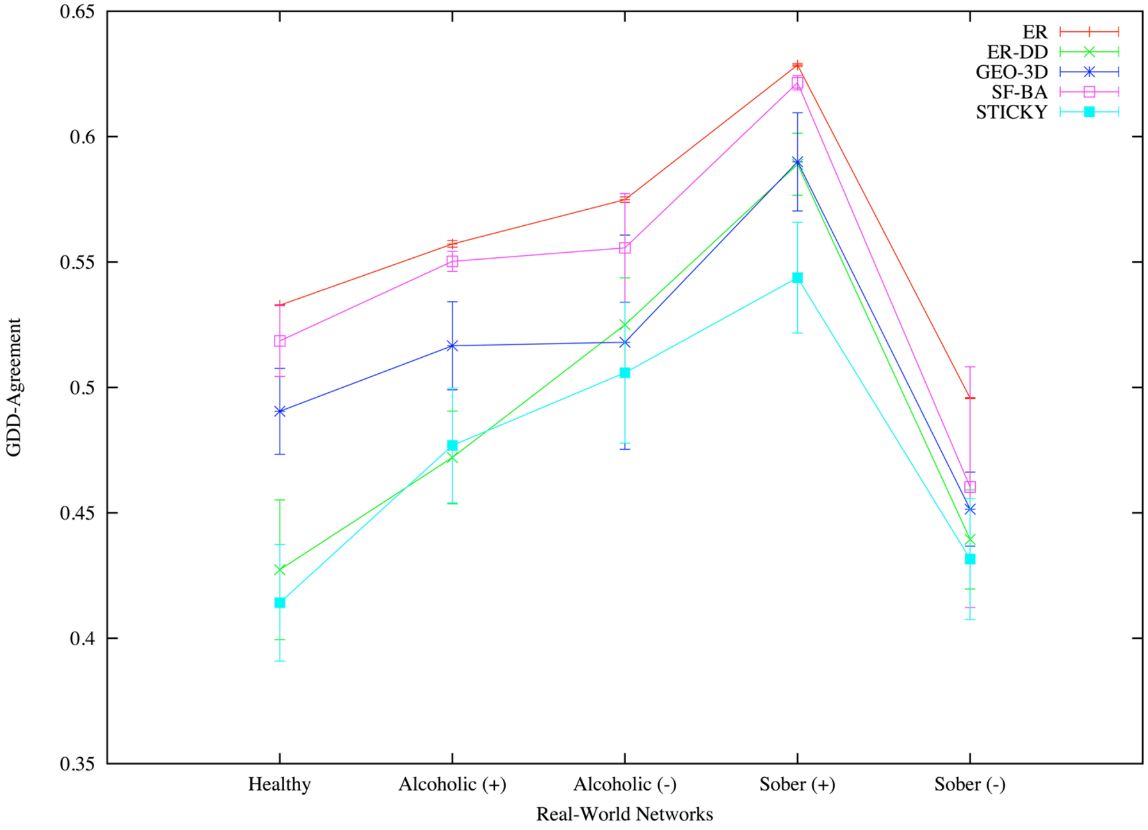

0.65
0.6
0.55
0.5
0.45
0.4
0.35
GDD-Agreement
Healthy
Alcoholic (+)
Alcoholic (-)
Sober (+)
Sober (-)
Real-World Networks
ER
ER-DD
GEO-3D
SF-BA
STICKY

Figure 5a

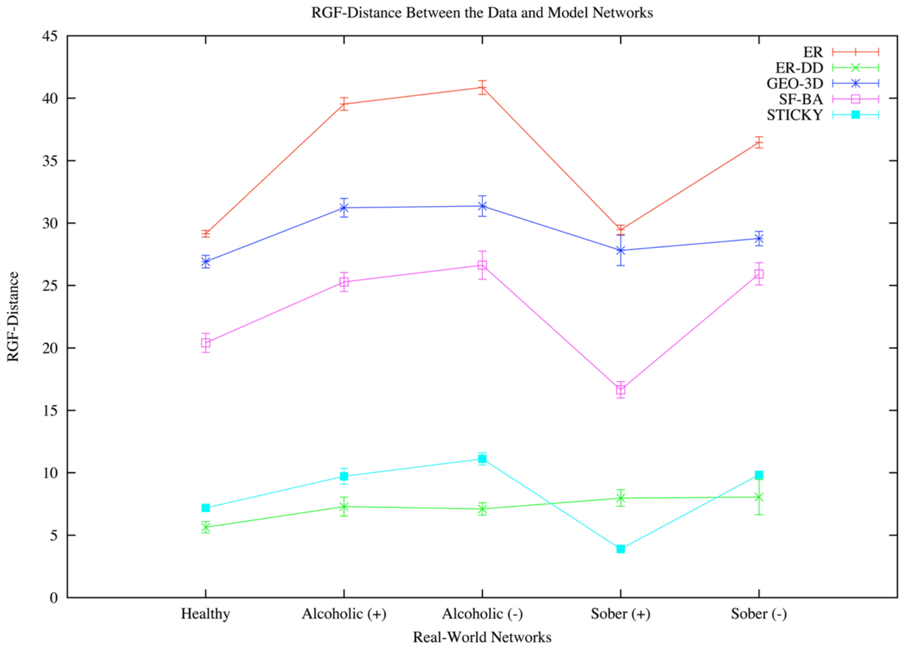


Figure 5b

Figure 5. Local metrics

Graphs of GDD-Agreement (5a) and RGF-Distance (5b) comparing the patient-derived networks against the ER, ER-DD, GEO-3D, SF-BA, and STICKY graph models

Table 1

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

Table 2

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