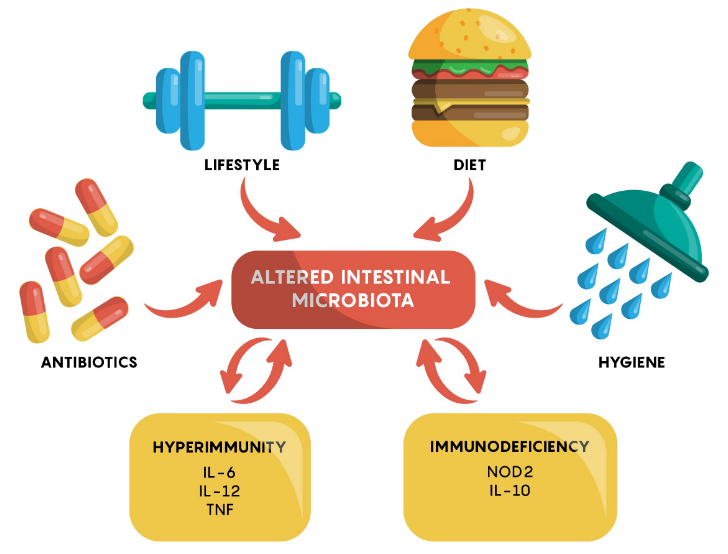
**Correlation Network Analysis of Gut Microbiota in Healthy Humans and Prediabetic Humans**

Hugo H. Corado

***Abstract* – Over the course of the last several years, the gut microbiome has been demonstrated to have a key role in the influence of remote organs and immune functions. Due to these influences, the function of the gut microbiota has been linked to different chronic diseases such as gastrointestinal inflammatory and metabolic conditions. The gut microbiome has also shown an affiliation with neurological, cardiovascular, and respiratory illnesses. Due to vast amount of disease and illnesses that the gut microbiome has been associated with, an increase in effort to understand the development of the microbiome in humans has been made to increase the amount of knowledge available. The end goal is to develop treatments and therapies that can be used to restore health or even prevent a disease before it comes to fruition.  
 Previous studies have shown that there is difference in the gut microbiome composition between healthy people and those that have been diagnosed with Type 2 diabetes. In order to understand how the gut microbiome changes while the disease in developing, a more robust investigation can be taken into the gut microbiome composition of someone that has been deemed as at risk of developing Type 2 diabetes (prediabetic). Different abundances in phylum, classes, orders, families, and genus have been observed between healthy, prediabetic, and T2D gut microbiota. Using abundance data generated by studies can help in developing more robust insights into how the composition of the gut microbiota changes depending on whether someone is healthy, at risk of diabetes, or has already been diagnosed with Type 2 diabetes.**

1. **Introduction**  
    A deficit in the production of insulin from the pancreatic beta cell or insulin resistance are the reasons that a human develops Type 2 diabetes (T2D). The disease stems from consequences that come from genetic and environmental factors, see Figure 1. The gut microbiome plays a key role in organizing the physical and chemical components of the intestinal barrier with various other systems in the body [4]. The diabetes is the most common metabolic disease and targeting the gut microbiome through treatments such as fecal transplantation could be considered since there have been differences found in between the gut microbiota of healthy human and T2D humans [5].  
    When looking at the gut microbiome in healthy humans the most abundant phyla found are the Firmicutes and Bacteroidetes. The most abundant classes in healthy human are the Clostridia and the Bacteroidia. Meanwhile the most abundant orders in healthy human are the Clostridiales and the Bacteroidales [2]. In prediabetic humans the most abundant phyla are the Bacteroidetes and firmicutes. The most abundant class in prediabetic human is the Chloracidobacteria. When compared to prediabetic and healthy humans, T2D humans had increased abundances of the Clostridia class and the lower abundances of the Bacteroidia class [3].  
     
   *Figure 1*. [6]. This image shows the different factors that can affect the gut microbiome in a person. These factors along with the genetics of a person can upkeep or alter the composition of the gut microbiome of a unique individual.

**Motivation**  
 Other studies have shown that there is a difference in the gut microbiota composition between healthy humans and T2D humans [1]. Being able to take a more robust look in the gut microbiome of individuals that are at risk of diabetes (prediabetic) could help in understanding the biological changes that occur after obtaining T2D.  
 The motivation behind this report is to use correlation networks built by the phyla, class, order, family, and genera found in healthy and prediabetic humans. Correlation networks are used to visualize the measure of how vertices are related to one another. The edges in a correlation network are derived by the correlation analysis that is indicated by the coordinated behavior between vertices across the data set.

1. **Materials and Methods**

**Data Collection**  
 The abundance data for healthy humans was collected from study done to create a baseline human gut microbiota profile that could be used in a clinical setting to improve existing treatments or develop new therapies. The dataset has abundance values for 8 phyla, 18 classes, 23 orders, 38 families, and 59 genera across 98 samples [2]. Participants used to gather this data were given instructions on how to record the dietary intake and provide samples over a seven-day period. The DNA in the fecal samples was extracted via an isolation kit. The DNA was then diluted and amplified. Lastly, the DNA was cleaned and sequenced. The file came in the form of a Microsoft Excel Workbook.  
 The abundance data for humans that have been diagnosed as prediabetic was collected by the Integrative Personal Omics Profiling study. The dataset has abundance values for 6 phyla, 12 classes, 12 orders, 21 families, and 44 genera across 856 samples [7]. Each individual used collection kits that collected the stool twenty-four hours before each baseline visit. The sample is then frozen and processed [1]. This file came in the form of comma-delimited file.

**Data Formatting**

The healthy human abundance dataset came in the form of a Microsoft Excel Workbook. A comma-delimited file was made from this workbook via Excel itself that resulted in the dataset being in the same format as the prediabetic dataset. In both datasets, each phylum, class, order, family, and genus had its own column. For example, “phylum\_Actinobacteria” was its own column. Each row consists of a sample ID and provides abundance values for each respective column. The Python programming language with the Pandas package was used to group together every phylum, class, order, family, and genus into their own respective data frames, which made it easier to process. Due to the low samples size in both abundance datasets, no further filtering had to be made.

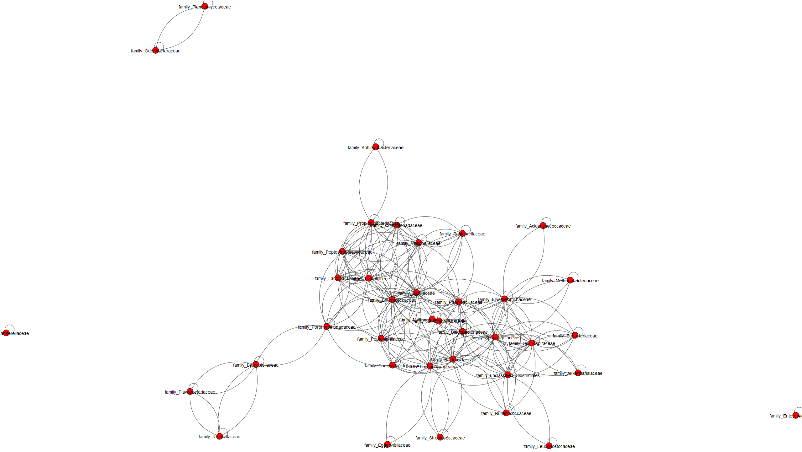
**Network Generation**  
Several steps were taken using the Python programming language to create a correlation network for the phyla, classes, orders, families, and genera, respectively. The SciPy Stats package was used to determine the beta distribution of the data frames. Since each of the data frames are the same size, this distribution only had to be calculated once.  
For every unique data frame, the Pandas package was used to calculate the correlation values between each of the columns within the data frame. A new data frame was then created in which column 1 and column 2 were composed of unique phyla, class, order, family, or genus, while column 3 held the correlation value between the two. For example, a row would have “phylum\_Firmicute” in column 1, “phylum\_Actinobacteria” in column 2, and the third value in the row consists of the correlation value between the two.  
 The SciPy Stats package was used to determine the p-value based on the correlation value of the third column mentioned in the newly made data frame. The Statsmodels package is then used to determine the adjusted p-values based on an alpha value of 0.05 using the Bonferroni method. The same package was used to determine whether that correlation value is True or False based on the adjusted p-value.

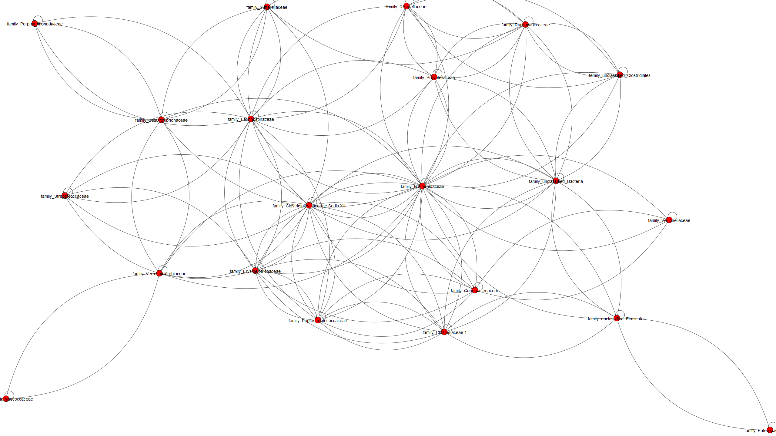
Each of the new data frames that was made for the phyla, classes, orders, families, and genera were then filtered to only keep the rows that resulted in the correlation value being determined as True and were saved to their own, respective, tab-delimited file.  
 The last step to generate the correlation networks involved using the iGraph package. This package made it very simple to make the correlation networks since all that is needed is to pass the new tab-delimited files this package. The package also comes with network property tools that can be used to determine values such as betweenness, edge betweenness, diameter, and clusters amongst other properties. The Kamada-Kawai force-directed algorithm was used as the layout of each of the correlation networks that was made and was then saved a 2560 by 1440 pixel, PNG file.

**Microbiome Correlation Analysis**  
 Using the iGraph Python package, selecting the central node of each of the correlation networks that made can be done by using the betweenness property that is a part of the package. The summary property provides the number of vertices and edges within the correlation network. The omega property is used to get the number of cliques within the correlation network. Lastly, the cluster property provides the number of clusters and lists them out, diameter provides the diameter of the correlation network, and the transitivity undirected property provides the clustering coefficient of the correlation network.

1. **Results**

For this study, a correlation network was made for the phyla, class, order, family, genera categories found in the healthy human dataset and prediabetic human dataset. While a unique correlation network was made for the phyla, class, order, family, and genus, respectively, this report will be taking a closer look at the family category. This analysis can be forked through a GitHub repository and everything needed to run the code in a Docker container.  
 From the healthy human dataset, there were a total of 38 nodes and 252 edges made, while in the prediabetic human dataset, there were a total of 21 nodes and 129 nodes. This is due to more families being tested for in the healthy human dataset. There were seven cliques observed in the healthy human network while there were five in the prediabetic human network. While the clusters property of iGraph resulted in the healthy human network having six clusters, clusters two through six consisted of a cluster being having only one vertex, so both network in this case, had one cluster. The diameter of the healthy human network is six while the one in the prediabetic human network is only four. The healthy human network has a higher transitivity measure than the prediabetic network meaning that the probability that the families are connected in the healthy human network is higher than that of the ones in the prediabetic network. The family that has the most centrality in the healthy human correlation network is the Bacillaceae, unfortunately this family was not tested for in the prediabetic human dataset but the Bacteroidaceae showed the most centrality in that dataset. In the healthy human network, Bacteroidaceae showed a lower centrality measure than in the prediabetic network. The family correlation networks for healthy humans and prediabetic humans have been visualized and available via Figure 2 and 3, respectively.

  
*Figure 2*. The family correlation network for the healthy human dataset. The main cluster can be observed here and while it cannot really be seen in this image, the bacillaceae family is at the center of this cluster.

  
*Figure 3*. The family correlation network for the prediabetic human dataset. The only cluster can be observed here and while it cannot really be seen in this image, the Bacteroidaceae family is at the center of this cluster.

1. **Discussion**

Differences have been found in the gut microbiome composition of people that been deemed as either healthy, at risk of diabetes, or have been diagnosed with Type 2 diabetes. There are many factors that go into a person developing diabetes both, genetic and environmental. While a person might not have control over the genetic factors, things such as diet and exercise all have a role in the diversity and health of the gut microbiome.

This report looked creating correlation networks for the phyla, class, order, family, and genera of the gut microbiome. There were differences spotted at the family level of the dataset. Investigating the genus categories of the dataset led to the betweenness centrality measure of Bacteroides being much higher in prediabetic humans than in healthy humans. Bacteroides play a crucial role in processing complex molecules into simpler ones in the host intestine. The transitivity measure of the genera is higher in the healthy human correlation network than that of the prediabetic human correlation network, meaning that the correlations found in the bacteria of healthy humans are more probable to occur than those found in prediabetic humans.

The correlation networks and analysis that were done in this report can be reproduced via the use of the Docker container. The GitHub link had been provided for the container and can then be forked and ran in your own environment. Continued studies should be done in which there is a more one-to-one comparison between the healthy human dataset and prediabetic dataset. Future studies should include some form a Type 2 diabetes dataset to create a more robust analysis and comparison.

**References**  
1. Integrated Personal Omics Profiling. (n.d.) Retrieved from: <http://med.stanford.edu/ipop.html>

2. King, C. H., Desai, H., Sylvetsky, A. C., LoTempio, J., Ayanyan, S., Carrie, J., ... & Howell, P.  
(2019). Baseline human gut microbiota profile in healthy people and standard reporting template. *PloS one*, *14*(9), e0206484.

3. Lambeth, S. M., Carson, T., Lowe, J., Ramaraj, T., Leff, J. W., Luo, L., ... & Shah, V. O. (2015). Composition, diversity and abundance of gut microbiome in prediabetes and type 2 diabetes. *Journal of diabetes and obesity*, *2*(3), 1.

4. Navab-Moghadam, F., Sedighi, M., Khamseh, M. E., Alaei-Shahmiri, F., Talebi, M., Razavi, S., & Amirmozafari, N. (2017). The association of type II diabetes with gut microbiota composition. *Microbial pathogenesis*, *110*, 630-636.

5. Ussar, S., Fujisaka, S., & Kahn, C. R. (2016). Interactions between host genetics and gut microbiome in diabetes and metabolic syndrome. *Molecular metabolism*, *5*(9), 795-803.

6. William, J. (2020). The gut microbiome and its impact on the brain. *Medicine LibreTexts*. Retrieved from: <https://med.libretexts.org/Bookshelves/Pharmacology_and_Neuroscience/Book%3A_Neuroscience_(Ju)/04%3A_Emergent_Topics_in_Neuroscience/4.01%3A_The_Gut_Microbiome_and_its_Impact_on_the_Brain>

7. Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., ... & Johnson, J. (2019). Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, *569*(7758), 663-671.

**GitHub Repository**  
<https://github.com/jinbe-808/Senior-Project>