Brad Fortunato

BF550

Professor Kirill Korolev

Project 1

Introduction:

Zhou et al. of *Correlation between Either Cupriavidus or Porphyromonas and Primary Pulmonary Tuberculosis Found by Analysing the Microbiota in Patients 'Bronchoalveolar* Lavage Fluid set out to find links between the diversity (or lack thereof) of lung microbiota in cases of chronic pulmonary tuberculosis (TB). Applying concepts developed through the rise of systemic biology, it is suggested that cases of chronic infectious diseases often point towards a type of twisted equilibrium between the host microbiota and the pathogen, allowing said pathogen to remain. Zhou et al. wished to apply nucleotide sequencing in divulging the composition of the diseased lung microbiome, as it has been done previously with other areas (i.e., linking obesity to differing diversities in the intestinal microbiome).

A sample of 32 patients presenting chronic TB were enrolled, with chest radiography being the most important selection factor; patients had to display cases featuring one normal lung and one diseased lung. Bronchoalveolar fluid was obtained from both sides as a marker for microbiota, with the normal sample being denoted as Group A and the diseased sample denoted as Group B. Further, 24 healthy patients were chosen as the control group, denoted as Group H. Zhou et al. went on to perform parallel pyrosequencing of bacterial 16s rDNA amplicons in the V3 region. Subsequent sequencing data was binned into FASTA files, aligned, and then used calculate the evenness and Shannon entropy indices.

Methods and Analysis:

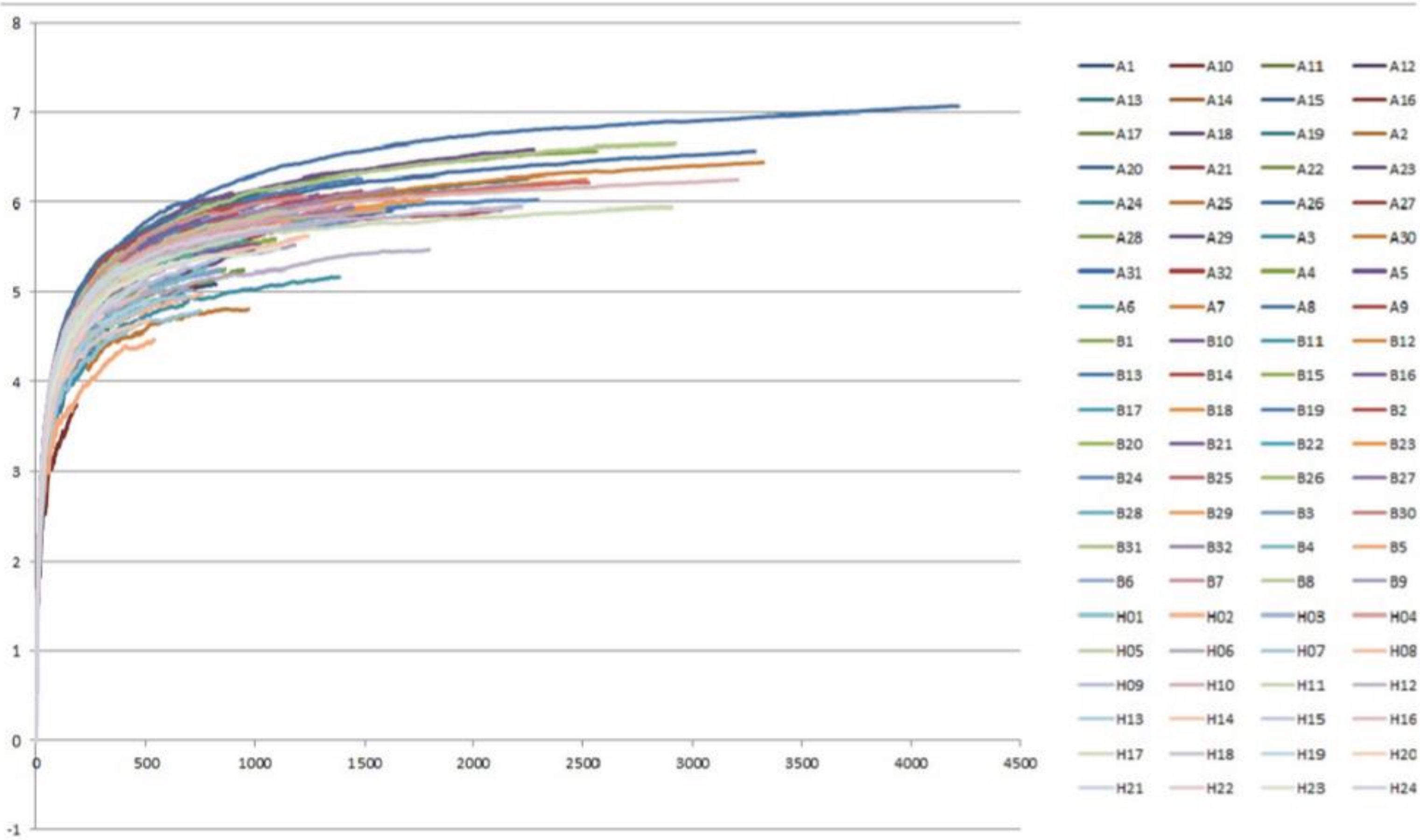
Raw data was obtained in FASTA format from the data repository DRYAD. Data was analyzed the following ways;

- Files of Group A, Group B and Group H were divided into three lists; *filesA, filesB, filesC.*
- Iterate through each file in each list, going line by line searching for any occurrences of amplicons, and then adding said amplicon to list *count*. Then creating a dictionary containing each amplicon as the key and each count of said amplicon being the value. Several assumptions were made in this analysis;
    - As stated in Zhou et al., detection of amplicons only occurred in sample sequences around 200 bp long. So in turn, my algorithm only searched for amplicons in sequences > 185 bp long.
    - As stated in Zhou et al., amplicon length was found to be around 200-220 bp long. Using this information, the algorithm would iterate through the length of the line, gradually growing in size in the range between 200-220 bp.
        - For example; Search line for line[0:200] → line[0:220], line[1:200] → line[1:220], etc.
        - The range was eventually decided to be between range(4,1000). This decision will be explained more in depth in the next section.
    - What exactly are amplicons? The Wikipedia description denotes them as most commonly being direct repeats and inverted repeats. So knowing this, the algorithm search each line for occurrences of both.

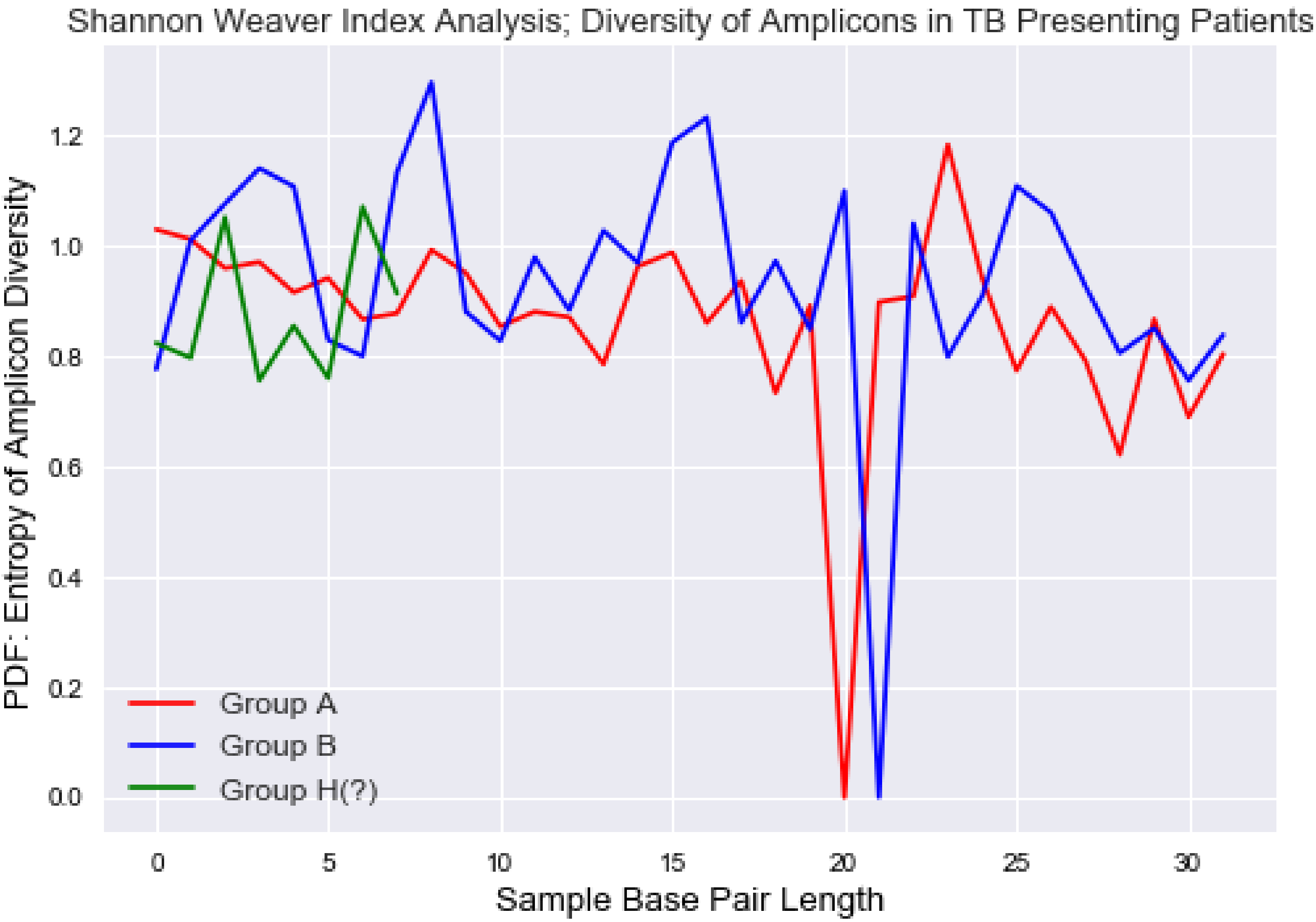
    - Final count is given an additional occurrence to make up for first amplicon detected.
- This analysis was repeated for each list of files, eventually with each count value for each line of each file group being added to lists *platoA*, *platoB* and *platoC* respectively.
- Using *pandas* and *scipy*, entropy (Shannon Weiner Diversity Index) was found for each list of amplicon counts per line of each file. Each instance of entropy was appended to list *dataA*, *dataB* and *dataC* respectively.
- Finally, each entropy data set was plotted on a probability density curve.

<u>Results and Discussion:</u>

The graph I was attempting emulate;



My graph;



Very obviously, I was not able emulate Zhou et al.'s graph. As I moved further through my analysis, it became clear that what appeared to be a simple(r) looking graph would be quite difficult to emulate. I believe this to be due to varying reasons;

- Zhou et al. did not use Python (at least directly) to analyze the FASTA data as far as I could tell. They performed several different forms of analysis;
  - "Individual sequences were aligned using Aligner tools, and aligned sequence files for each sample were processed using complete-linkage clustering with distance criteria".
  - "We used the Uclust algorithm to cluster all of the sequences with a cut-off value of 97%; after clustering, we used the representative sequence of each type as the

    - operational taxonomic unit (OTU) and re-corded each OTU sequence representing the number of sequences and the classification information."
    - "Shannon diversity was estimated using Estimate S Win 8.20 software."
- It seems as if Zhou et al. were able to determine multiple points of entropy per line, not per group, as shown in their graph. This points to calculating entropy for every line in every file. How they were able to do this I'm not sure, as when I searched for amplicons, I often times was only able to locate 1 or 2 examples per line
    - This may have been different if I was searching in lines shorter than 185 bp.
    - Further, I had to shift my range(200, 220) to range(4,1000), as I was detecting no amplicons in my earlier range. Through tinkering I found that there were seemingly no amplicons above 25 bp. This has mostly likely something to do with my analysis and the differences in software we used to infer information from the FASTA files.
- Besides this, not too much information was given on the steps of the analysis. Some more information detailing use of the software tools and how said tools worked to parse through the data would have been appreciated.
- Further, the if the complete DRYAD data was available, I could not locate it. The .zip file featuring the data contained all the samples for both Group A and Group B, but seemingly none for Group H. I utilized the FASTA files denoted by an "N" in their file name, which I'm assuming stands for "Normal". That being said, there were only 8 of said files in comparison to the 32 of Group A and Group B.
- Lastly, while Zhou et al's graph is pretty, is noticeably lacking *x* and *y* axis labels. The graph is described as a collection of "Shannon-Weaver index curves", but through my research I could not find any other examples of this type of graph, at least by the name given.

In the end I believe I bit off more than I could chew. Nevertheless, I believe there are some similar conclusions involving microbial diversity in TB patients that can be interpreted from both graphs. As quoted from the paper; "According to our research, significant differences can be observed in the respiratory tract microbiota of healthy people when compared with TB patients." Looking at my graph, you can note a greater degree of differences in amplicon diversity in the Group B curve vs the Group A curve, lending to a greater presence of exotic microbes in the TB effected lung and less homogeneity.

Overall, this project was a learning experience. I am intensely interested in the study of our microbiome, and hopefully after learning a bit more in depth in data analysis I would eventually be able to return to this study and attempt at creating the figures in the future.