DREAM Research Final Report

Samiha Ajaj

Summer-Fall 2023

1. Introduction

My research was focused on the study of metagenomics. Metagenomics involves the examination of complete nucleotide sequences, extracted and examined from all the organisms found in a collective sample, commonly microbes. It is frequently used to study a group of microorganisms, such as those in soil. This research focused on metagenomics, and analyzed ten soil samples to study the nucleotide sequences of microbial communities. The research aims to develop computational methods to explore the metabolic potential of soil microbes in restored and pre-restoration tallgrass prairies. By identifying bacterial composition, community structure, and discrepancies among samples, the study seeks to fill knowledge gaps about soil communities and understand how microorganisms thrive in different soil types. The ultimate goal is to advance sustainable agriculture and enhance plant success by manipulating soil microbiomes.

While studying the microbial communities, my responsibility entailed developing a data structure designed to intake soil sample results and generate output detailing the percentage of various bacteria and archaea present in each soil sample. Additionally, the data structure was designed to identify the specific taxonomic classification to which each bacterium or archaeon belonged.

2. Related Work

Metagenomics, a subject of study for several decades, has played a pivotal role in advancing scientific, healthcare, and agricultural domains. During the course of my research, I consulted scholarly works such as Cuadros-Orellana et al., Fungal Genom Biol 2013, 3:2, which delved into the diversity of fungi in the environment. Additionally, I referred to a piece titled , "Using Metagenomics to Connect Microbial Community Biodiversity and Functions", exploring the impact of microbial communities on ecosystems, both in terms of adverse and beneficial effects, and the consequential alterations in the ecosystem.

3. MetaPhlan (Metagenomic Phylogenetic Analysis)

In collaboration with Dr. Ramaraj, we utilized MetaPhlan, a bioinformatics tool, to extract the microbial communities present in each soil sample. MetaPhlan is a specialized program utilized in metagenomics and microbiome research. Its fundamental function is to analyze DNA sequences derived from microbial communities across diverse samples, including soil, and identify the taxonomic composition of these microbial communities. The following text file exemplifies the results obtained through the use of MetaPhlan.

```
#SampleID       Metaphlan_Analysis
#clade_name     NCBI_tax_id     relative_abundance      additional_species
k__Bacteria     2       95.40545
k__Archaea      2157    4.59455
k__Bacteria|p__Proteobacteria   2|1224  53.9728
k__Bacteria|p__Actinobacteria   2|201174        22.09769
k__Bacteria|p__Nitrospirae      2|40117 8.1279
k__Archaea|p__Thaumarchaeota    2157|651137     4.59455
k__Bacteria|p__Bacteria_unclassified    2|      4.07235
k__Bacteria|p__Acidobacteria    2|57723 2.8542
k__Bacteria|p__Bacteroidetes    2|976   1.69763
k__Bacteria|p__Verrucomicrobia  2|74201 1.31587
k__Bacteria|p__Chloroflexi      2|200795        0.74883
k__Bacteria|p__Candidatus_Saccharibacteria      2|95818 0.44363
k__Bacteria|p__Gemmatimonadetes 2|142182        0.06022
k__Bacteria|p__Planctomycetes   2|203682        0.01427
k__Bacteria|p__Proteobacteria|c__Alphaproteobacteria    2|1224|28211    20.55886
k__Bacteria|p__Proteobacteria|c__CFGB52626      2|1224| 18.47906
k__Bacteria|p__Actinobacteria|c__CFGB52786      2|201174|       13.86218
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria    2|1224|1236     9.75989
k__Bacteria|p__Actinobacteria|c__Actinomycetia  2|201174|1760   7.70522
k__Bacteria|p__Nitrospirae|c__CFGB38382 2|40117|        5.34446
k__Archaea|p__Thaumarchaeota|c__Nitrososphaeria 2157|651137|1643678     4.59455
k__Bacteria|p__Bacteria_unclassified|c__CFGB44871       2||     4.07235
k__Bacteria|p__Proteobacteria|c__Betaproteobacteria     2|1224|28216    3.04494
k__Bacteria|p__Nitrospirae|c__Nitrospira        2|40117|203693  2.78344
k__Bacteria|p__Acidobacteria|c__CFGB14597       2|57723|        2.63033
k__Bacteria|p__Proteobacteria|c__Deltaproteobacteria    2|1224|28221    1.93041
k__Bacteria|p__Verrucomicrobia|c__CFGB2153      2|74201|        1.31587
k__Bacteria|p__Bacteroidetes|c__CFGB24244       2|976|  0.89708
k__Bacteria|p__Chloroflexi|c__CFGB49770 2|200795|       0.74476
k__Bacteria|p__Candidatus_Saccharibacteria|c__Candidatus_Saccharibacteria_unclassified  2|95818|        0.44363
k__Bacteria|p__Bacteroidetes|c__Sphingobacteriia        2|976|117747    0.39198
k__Bacteria|p__Actinobacteria|c__Thermoleophilia        2|201174|1497346        0.36676
k__Bacteria|p__Bacteroidetes|c__Chitinophagia   2|976|1853228    0.2919
```

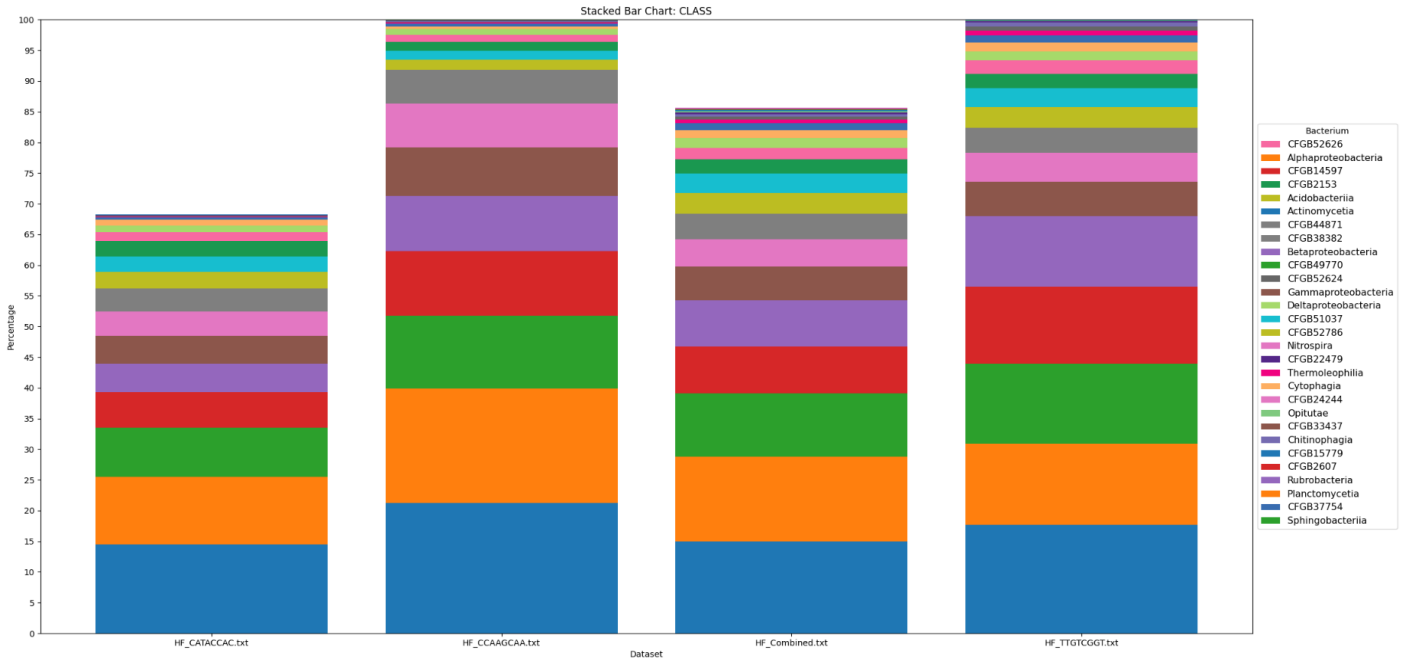
While inspecting the text file, it is apparent that MetaPhlan has broken down the bacterial composition at each taxonomic level. Notably, the yellow underlined letters, K, P, and C correspond to Kingdom, Phylum, Class, Order, Genus, and Species. However, this image does not include representations for Order, Genus, and Species. The last value on each line represents the percent of each bacteria and archaea within the sample. Referencing the squared red box and arrow, it is evident that bacteria constitutes 95.40545% and archaea accounts for 4.59455% of this specific sample.

Acknowledging MetaPhlan's capacity to extract taxonomic levels and percentages, my devised data structure will take that information and create a bar chart to visualize the results.

4. Data Structure for Visualization

I crafted a data structure using the Python programming language. Throughout the development of this data structure, I recognized the necessity of using a text file as input and producing outputs detailing the taxonomic levels and corresponding percentages of bacteria and archaea. To achieve this objective, I created an assortment of for loops, if statements, and other conditional constructs. The program systematically examines the text file, parsing through each line and assessing whether it satisfies any of the predefined conditions within the program. For instance, in the context of generating a bar chart for sample HF with a focus on the Class level, the program scans each line, identifies the condition associated with the presence of "C_" in the line, and subsequently extracts and outputs the microbial percentage corresponding to the Class level.

Figure demonstrating four different sample types of HF at the taxonomic level Class.



Following this, the program generates a key that lists the names of various bacteria and archaea, assigning distinctive colors to each. These colors, along with their respective percentages, are depicted in the stacked bar. The *x-axis* of the graph specifies the visualized sample, while the *y-axis* indicates the corresponding percentage.

5. Presented is the program I designed to visualize the results of the soil samples.

```
import numpy as np
import matplotlib.pyplot as plt

taxaKingdom = {'CLASS':['c__', 'o__'] , 'ORDER':['o__', 'f__'], 'FAMILY':['f__', 'g__'], 'GENUS': ['g__', 's__'], 'SPECIES': ['s__',
't__'] }

def extractTaxa (files, taxa):

    for file in files:
        names.setdefault(str(file), {})
        try:
            file = open(file, "r")
        except FileNotFoundError:
            print("The file "+ file + " does not exist.")
            exit()
        Lines = file.readlines()

        for line in Lines:
            if str(taxaKingdom[taxa][0]) in line and str(taxaKingdom[taxa][1]) not in line and 'Archaea' not in line:
                    sentence = line.split()
                    array = sentence[0].split(str(taxaKingdom[taxa][0]))
                    nameAndBacteria = array[1]
                    last_key = list(names)[-1]
                    names[last_key].setdefault(nameAndBacteria, []).append(float(sentence[2]))


names = {}

userInputfile = input("Enter the 4 files with .txt (seperate by just spaces)\n")
files = userInputfile.split()
taxa = input("Enter the taxa level you want\n")
taxa = taxa.upper()

extractTaxa(files, taxa)

# Create a figure and axis
fig, ax = plt.subplots()

# Get the unique bacterium names from all datasets
all_bacteria_names = set()
for dataset in names.values():
    all_bacteria_names.update(dataset.keys())

# Define a list of 40 distinct and contrasting colors
distinct_colors = [
    '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728', '#9467bd',
    '#8c564b', '#e377c2', '#7f7f7f', '#bcbd22', '#17becf',
    '#1a9850', '#f768a1', '#a6d96a', '#fdae61', '#386cb0',
    '#f0027f', '#666666', '#756bb1', '#542788', '#7fc97f',
    '#1f77b4', '#ff7f0e', '#2ca02c', '#d62728', '#9467bd',
    '#8c564b', '#e377c2', '#7f7f7f', '#bcbd22', '#17becf',
    '#1a9850', '#f768a1', '#a6d96a', '#fdae61', '#386cb0',
    '#f0027f', '#666666', '#756bb1', '#542788', '#7fc97f'
]

# Ensure there are 40 colors in the list
while len(distinct_colors) < 40:
    distinct_colors += distinct_colors[:40 - len(distinct_colors)]

# Initialize x positions for each dataset
x = np.arange(len(names.keys()))

# Loop through each dataset
for i, (name, dataset) in enumerate(names.items()):
    # Initialize the list of percentages for this dataset
    percentages = [dataset.get(bacterium, [0])[0] for bacterium in all_bacteria_names]

    # Sort percentages and corresponding bacterium names in descending order
    sorted_data = sorted(zip(percentages, all_bacteria_names), key=lambda x: x[0], reverse=True)
    sorted_percentages, sorted_bacteria_names = zip(*sorted_data)

    # Initialize the cumulative sum for this dataset
    cumulative_sum = 0
```

```
    # Initialize the cumulative sum for this dataset
    cumulative_sum = 0

    # Loop through each bacterium in this dataset
    for j, bacterium in enumerate(sorted_bacteria_names):
        color = distinct_colors[j % len(distinct_colors)]  # Cycle through distinct_colors
        ax.bar(
            x[i],
            sorted_percentages[j],  # Use the sorted percentages
            label=bacterium,
            color=color,
            bottom=cumulative_sum
        )
        cumulative_sum += sorted_percentages[j]  # Update the cumulative sum

# Set x-axis labels and tick positions
ax.set_xticks(np.arange(len(list(names.keys()))))
ax.set_xticklabels(names.keys())

# Set y-axis limits to 0% to 100%
ax.set_ylim(0, 100)

# Set y-axis ticks in increments of 5
ax.set_yticks(np.arange(0, 101, 5))

# Set labels and title
ax.set_xlabel("Dataset")
ax.set_ylabel("Percentage")
ax.set_title("Stacked Bar Chart: " + taxa.upper())

# Create the legend without duplicates
handles, labels = ax.get_legend_handles_labels()
unique_labels = {label: handle for handle, label in zip(handles, labels)}
x = 10 - (len(unique_labels) - 50) / 20

ax.legend(unique_labels.values(), unique_labels.keys(), title="Bacterium", loc='center left', bbox_to_anchor=(1, 0.5,), fontsize= x)
print(len(unique_labels))
# Show the plot
plt.show()
```

6. Conclusion

Upon analyzing the entirety of the forty sample types and subsequently presenting them visually, a trend emerges: a preponderance of the samples manifests a composition primarily constituted by diverse strains of bacteria, notably of the proteobacteria classification. Remarkably, a subset of the samples exhibits an absence of archaea. A pattern emerges, demonstrating the marginal role played by archaea in the sampled specimens. Majority of the samples included alphaproteobacteria, actinomycetes, among others. This implies a significant influence of proteobacteria in shaping the soil composition related to crop cultivation. Thereby, impacting their growth dynamics, both beneficial and harmful, contingent upon other variables such as sunlight exposure, water intake, and weather conditions. Overall, the research was successful, creating the visual representation of each sample's outcomes, thus, providing a foundational platform for interventions in soil microbes to propel advancements in sustainable agricultural practices.

Presented here is a Google Folder that includes all sample types in text file format, graphical depictions illustrating the outcomes of each text file, and the data structure:
https://drive.google.com/drive/folders/1wQ5_CpDXQx1isKyF3wH_-zHt4-kxZvvP?usp=sharing