Introductions

Microbiome

Microbiomes are like small people in a crowded city. They live in different places. Inhabitant species normally include bacteria, fungi, archaea, eukaryotes, and viruses. For example, they can help us to digest better in the digestive system and provide resistance to protect us from getting sick in the immune system. They also exist in the plant root system, which helps plants take nutrients and tolerate tension. These species also live in environments as part of natural cycles. Their composition can be influenced by diet, environment, and medication. In such a dynamic nature, they are always evolving, and humans still need to learn more about their influence on health and the environment.

Challenge of microbiome data

Examining microbiome data is not straightforward. Lots of species and genes de­mand our attention, so directing our focus on significant ones is imperative. The data demonstrates the proportional presence of each type, not their precise quantities. Therefore, unique statistical techniques are required for proper comprehension. A large number of zero values are normally found in data. This means that no microbes are present. It may be because of the way we collected the data. This will assist ensure that our findings are correct. There is no universally accepted method for dealing with this type of data. Different strategies produce different outputs. This demonstrates that we must all use the same methodologies in microbiome research.

16S rRNA sequence

16S rRNA, found in bacteria and archaea, represents a conserved gene that is valuable for microbiome research. The gene was sequenced by scientists to detect and categorize microorganisms, especially those which had never been cultivated. Whereas 16S rRNA sequencing offers great taxonomic specificity at the level of species and taxa within a given sample, its effectiveness diminishes with decreasing taxonomic granularity owing to sequence homology. Additionally, biases introduced by PCR amplification at every step may compromise accurate diversity as well as compositional assessments. In spite of these limitations, however, 16S rRNA sequencing still remains an affordable method suitable for large-scale studies on microbiomes.

Operational Taxonomic Units OTUs

Operational Taxonomic Units (OTUs) are groups of closely related 16S rRNA gene sequences. These sequences resemble fingerprints for several bacteria. Scientists utilize 16S rRNA sequencing to investigate bacteria. This approach generates a large amount of data, with each sequence representing a single bacterium. Grouping comparable sequences into OTUs facilitates the management of a large number of species. It simplifies data analysis for academics. Scientists assume that identical 16S rRNA sequences indicate that the microbes are related. However, the definitions of OTUs can change. A modest difference in similarity level can result in distinct groupings of sequences. This can have a nonlinear effect on the outcomes.

Phylogenetic Tree

A phylogenetic tree is a chart that reveals how different species are connected through evolution. These links trace back to a single ancestor, showing how the current species came to be. To build this chart, we usually use genetic sequence data. Each branch on the chart represents a unique line of heredity. The root of the tree stands for the first species, the one from which all others have descended. For creating these trees, some different methods were used to create the phylogenetic tree. These include those based on distance and simple methods known as parsimony. They can also use likelihood methods. This looks at how similar different genes are. But also, how they might be different.

**Operational Taxonomic Units OTUs**

In microbiome research, OTU counts represent the frequency with which 16S rRNA gene sequences cluster into Operational Taxonomic Units (OTUs). These OTUs serve as proxies for bacterial species as closely related sequences are grouped together. This approach Eases the analysis of large information produced through 16S rRNA sequencing by simplifying the vast number of individual bacterial sequences into manageable units.

**Phylogenetic tree**

The phylogenetic trees represent ancestor information between the species that finally lead to one joint ancestor. These trees had been based upon genetic information where each stem represents a different but separate evolutionary path. The evolutionary past is assessed with the help of sequence similarities and differences, using sequence distance, parsimony, and maximum likelihood.

For this simulation study that would be mentioned in Part 3, we will use the OTU counts computed from the read counts of the COMBO dataset (Wu et al.) Here lies an added value: that feature of unique phylogenetic distances, as is the case with other simulators, is absent here.

in microbiome research, OTU counts track how often 16S rRNA gene sequences cluster into groups called Operational Taxonomic Units (OTUs). These OTUs represent bacterial species, as closely related sequences are grouped together. This simplifies the massive amount of individual bacterial sequences generated by 16S rRNA sequencing into manageable units, making analysis easier.

Phylogenetic trees show the evolutionary relationships between species, tracing back to a common ancestor. Each branch in the tree represents a distinct evolutionary path. Sequence similarities and differences are used to assess evolutionary history through methods like sequence distance, parsimony, and maximum likelihood.

For a simulation study in Part 3, we'll use OTU counts from the COMBO dataset. This approach differs from other simulators by not utilizing unique phylogenetic distances.