Assignment 2

Supervised Machine Learning and Classification (randomForest)

The genera Bifidobacterium and Collinsella are found in the gut. While Collinsella is a pathogen that causes gut leakage and interferes negatively with lipid metabolism in the liver and intestine, Bifidobacterium on the other hand is beneficial to the gut (Gomez-Arango et al 2018, O'Callaghan and Douwe van 2016). They both belong to the phylum Actinobacteria, but the frequency of dietary fiber consumption controls their abundance (Gomez-Arango et al 2018, O'Callaghan and Douwe van 2016).

16S ribosomal RNA has frequently been employed successfully for the taxonomic classification of bacteria. Fiannaca and co-workers (2018), built a bacteria taxon classification technique for 16s rRNA (approximately 469 bp), using a k-mer length of 3<k<7 reaching 91% accuracy. I aim to obtain a classifier that identifies the genera Bifidobacterium and Collinsella using 16s RNA genes less than 800 bp, with up to 90% accuracy.

I obtained 16s RNA genes of 500 – 1000 bp of both genera from NCBI. After data wrangling and cleaning, I settled to build a classifier with genes of 536 – 802 bp so that I could have sufficient sequences for my training and validation data.

I observed that the accuracy of the classification increased with k-mer length, 74% accuracy for 1-mer length, and approximately 87% accuracy for 4-mer length. Also, for k-mer 4, an ntree of 50 resulted in 83% accuracy while an ntree of 100 resulted in ~ 87% accuracy. Beyond ntree of 100 up to 500, accuracy only increased by 1%.

Unfortunately, I could not build a classifier past the kmer-4 length. Rstudio returned an error message of not being able to subset past a particular number of columns.

In conclusion, the random forest algorithm for building a classifier model worked well. Though I aimed at achieving 90% accuracy, I only succeeded at approximately 87% accuracy. Probably the dataset used was not enough to improve accuracy. Only had about 1000 sequences. Hopefully, I would like to increase the data set if I intend to work on classifiers.

References

Fiannaca A, La Paglia L, La Rosa M, Lo Bosco G, Renda G, Rizzo R, Gaglio S, Urso A. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinformatics* 2018;**19(7):198**. https://doi.org/10.1186/s12859-018-2182-6

Gomez-Arango LF, Barrett HL, Wilkinson SA, Callaway LK, McIntyre HD, Morrison M, Dekker Nitert M. Low dietary fiber intake increases Collinsella abundance in the gut microbiota of overweight and obese pregnant women. Gut Microbes. 2018;9(3):189-201. doi: 10.1080/19490976.2017.1406584. Epub 2018 Mar 13. PMID: 29144833; PMCID: PMC6219589.

O’Callaghan A and Douwe van S. Bifidobacteria and their role as members of the human gut microbiota. Frontiers in Microbiology 2016;7