Improving Machine Learning Models of the Human Gut Microbiome

Kelly L. Sovacool

Changes in the taxonomic composition and metabolic activity of human microbiomes have been observed in several diseases including colorectal cancer (CRC) and *Clostridioides difficile* infection (CDI). Taxonomic composition is commonly defined by amplicon sequencing the 16S rRNA gene and clustering sequences into Operational Taxonomic Units (OTUs). The OTU abundances can then be used to train supervised machine learning (ML) models for tasks such as classifying samples as CRC or normal, or predicting the severity of CDI outcomes. Such models have the potential to improve the early detection of CRC, inform clinicians on which CDI patients may be most at risk of experiencing a severe case, or more generally contribute to our understanding of how the gut microbiome changes during disease states. However, there are a number of challenges to realizing the full potential of ML models of the human gut microbiome. First, current *de novo* OTU clustering methods produce high quality OTUs, but OTU assignments may change when new data are added. When deploying ML models trained on OTUs, external validation sets need to have the same OTUs as the data used for initial model training, which is not currently possible without reference-based methods that produce lower quality OTUs. Second, efforts to find consistent changes in taxonomic composition of microbiomes between normal and dysbiotic states have found mixed success, in part because interpersonal variability in taxonomic composition sometimes exceeds the variability between disease states. Variability of microbiome composition between individuals with the same disease status may be explained by functional redundancy, where different microbial species carry out the same functions and thus can replace each other with little effect on the overall function of the community. This proposal aims to 1) develop a new OTU clustering method to enable new data to be fit to existing OTUs, 2) train OTU-based ML models to predict CDI severity on a large dataset and apply the new clustering method for validation with an external dataset, and 3) incorporate functional composition from metagenomic data to improve the performance of ML models for classifying CRC cases.

To fill the need for a stable clustering method that produces high quality OTUs, we are developing OptiFit, a new algorithm that assigns sequences to existing *de novo* OTUs created by OptiClust. OptiFit is being refined and tested to determine its performance compared to OptiClust using publicly available datasets from human gut, mouse gut, marine, and soil microbiomes. The OTU quality and execution speed will be compared between OptiFit and OptiClust for all four datasets across 100 different random seeds. We hypothesize that OptiFit clusters reference-based OTUs at nearly the same quality as OptiClust and with faster execution speeds, allowing researchers to fit new data to existing OTUs for ML validation and deployment.

Previous studies have built OTU-based ML models to distinguish CDI cases from controls in order to demonstrate the importance of the gut microbiome in CDI and to identify the most important microbial features contributing to model performance. We now have access to a large dataset of amplicon sequences and clinical features from about 4,000 stool samples of CDI cases, diarrheal controls, and non-diarrheal controls (R21 cohort). OTU-based ML models will be built to not only distinguish CDI cases but also to predict the severity of infection in CDI cases based on clinical lab results. The OptiFit algorithm will then be used to cluster an external dataset (ERIN cohort) to the OTUs from the R21 cohort, allowing the external dataset to be used as a validation set. A model trained on such a large dataset and validated with external data could be used to help clinicians identify cases that may become severe in order to inform treatment strategies.

Previous studies have built OTU-based ML models to classify stool samples as normal or cancerous to serve as a less invasive diagnostic tool for CRC than colonoscopy, but have only achieved modest performance. Incorporating the known functional potential of the microbiome from metagenomic data may help account for functional redundancy and improve the performance of OTU-based models in classifying CRC, predicting CDI severity, or other microbiome modeling problems. We have whole metagenomes from colorectal carcinoma, adenoma, and normal stool samples. We will first identify known gene functions present in the samples and characterize the functional redundancy within and between disease states, then train ML models with known gene functions, OTUs, or both as features to classify samples into disease states. We hypothesize that models trained with known gene functions as features would outperform models built on OTUs alone.