Building Tools for Modeling Disease States of the Human Gut Microbiome

Kelly L. Sovacool

Specific Aims

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). OptiClust, a *de novo* OTU clustering method, has been shown to produce higher quality OTU assignments than other methods. OTU assignments from *de novo* clustering may change when new sequences are added to a dataset. However, one may wish to incorporate new samples into a previously clustered dataset without re-clustering all sequences, such as when comparing across datasets or deploying machine learning (ML) models with OTUs as features. To fill this need, the Schloss Lab is developing OptiFit, a reference-based clustering method that assigns sequences to existing *de novo* OTUs. OptiFit is being refined and tested to determine its performance compared to OptiClust, with the ultimate goal of applying it to test new microbiome data against deployed ML models.

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 could be built to not only distinguish CDI cases but also to predict the severity of infection in CDI cases based on clinical lab results. Additionally, an external dataset (ERIN cohort) could be used as a validation set to test the best model built on the R21 cohort. Permutation feature importance would identify the most important microbial features contributing to model performance, which could help clinicians in identifying cases that may become severe to inform treatment strategies.

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. 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 metagenomics data may help account for functional redunancy and improve the performance of OTU-based models in classifying CRC, predicting CDI severity, or other modeling problems.

Aim 1: Determine the performance of OptiFit in clustering reference-based OTUs against *de novo* OTUs.

Hypothesis: OptiFit clusters reference-based OTUs at nearly the same quality as OptiClust and with faster execution speeds.

A. *De novo* cluster databases into OTUs, then cluster datasets to the database OTUs with OptiFit.
B. *De novo* cluster half of the sequences in each dataset, then cluster the remaining sequences to the *de novo* OTUs with OptiFit.
C. Compare the OTU quality of the two OptiFit strategies above versus that of *de novo* clustering the datasets with OptiClust.

Aim 2. Build ML models with OTUs and clinical features to predict the severity of CDI and identify important micriobial features.

Hypothesis: OTU- and clinical-based ML models can be used to predict the severity of CDI.

A. Cluster sequences from the R21 cohort into *de novo* OTUs.
B. Build models using the R21 dataset to predict the severity score of CDI cases and identify the most important features.
C. Use OptiFit-clustered OTUs from the ERIN cohort as a validation test set for the best R21-based model.

Aim 3. Explore the impact of functional redundancy on model performance.

Hypothesis: Incorporating functional potential along with OTUs as features improves the performance of disease prediction models.

A. Identify known gene functions in CRC and CDI metagenomes.
B. Characterize the presence of functional redundancy within samples.
C. Build ML models using OTUs, known metagenomic functions, or both as features and compare performance.