Improving machine learning models for microbiome analysis and democratizing data science along the way

Kelly L. Sovacool

The human microbiome plays an important role in maintaining health. Changes in the taxonomic and functional composition of the gut microbiota have been implicated in numerous diseases including colorectal cancer, *Clostridioides difficile* infection (CDI), and others. Thus, the gut microbiome is a promising source of biomarkers for disease diagnosis and prediction. Machine learning (ML) approaches can leverage large datasets to gain insights into associations between the microbiota and disease. Here, we present a new algorithm that improves microbiome analysis for ML applications, apply ML to predict severity of CDI, and introduce resources that empower data scientists to go from the basics of coding to applying ML for reproducible research.

Assigning amplicon sequences to operational taxonomic units (OTUs) is an important step in characterizing microbial communities across large datasets. However, a gap in existing OTU assignment methods inhibited the ability of researchers to incorporate new samples to previously clustered datasets, such as when deploying ML models. To provide an efficient method to fit sequences to existing OTUs while maintaining high OTU quality, we developed the OptiFit algorithm, an improved implementation of reference-based clustering. Our benchmarks revealed that OptiFit produces similar quality OTUs as a gold standard method yet at faster speeds. Thus, OptiFit provides a suitable option for users requiring consistent and high quality OTU assignments for ML applications and beyond.

CDI can lead to severe complications including death, with half a million cases annually in the United States. The composition of the gut microbiome plays an important role in determining colonization resistance and clearance upon exposure to *C. difficile*. We investigated whether ML models trained on OTUs from stool samples on the day of CDI diagnosis could predict which cases led to severe outcomes. We trained models to predict CDI severity for four different severity definitions. The models performed best when predicting pragmatic severity, a composite definition of complications due to any cause or confirmed as CDI-attributable via chart review when possible. Our results suggest that while chart review is valuable to verify the cause of complications, including as many samples as possible is indispensable for training performant models on imbalanced datasets. We evaluated the potential clinical value of these models and found similar performance compared to prior models based on Electronic Health Records, although further work is needed to determine the feasibility of deploying such models in clinical practice. These results represent a step toward the goal of deploying ML to inform clinical decisions and ultimately improve CDI outcomes.

Bioinformatics is a kind of data science, an interdisciplinary field integrating computer science, statistics, and domain knowledge. Novice researchers frequently have domain knowledge, but lack other skills necessary to apply data science to their datasets while adhering to best practices in reproducibility. We developed three resources to help democratize data science: a curriculum teaching the basics of Python for data science to young students, a curriculum teaching programming skills for reproducible research, and an R package implementing an ML framework to help novices apply ML responsibly while being customizable for advanced users. These contributions cover a breadth of audience skill levels to help fill gaps in existing resources for data science. In summary, this dissertation advances bioinformatics for microbiome research from the start of data analysis through application, and ultimately toward enabling others to reproduce and extend our work.