BIO-FI: LEVERAGING THE POWER OF BIOLOGICAL BIG DATA FOR ADVANCED ANALYTICS AND MODELING OF CHEMICAL AND BIOLOGICAL THREATS

Learning Microbial Representations With Hilbert Curve Visualizations

Camilo Valdes Lawrence Livermore National Laboratory **Giri Narasimhan** Bioinformatics Research Group, Florida International University **Michael Morrison** Lawrence Livermore National Laboratory **Jennifer Clarke** University of Nebraska-Lincoln **Crystal Jaing** Lawrence Livermore National Laboratory

Microbiome samples can be examined by means of low-cost, high-throughput DNA sequencing, which is followed by the creation of microbial community abundance profiles, where the sequenced "DNA reads" are mapped against a collection of reference genomes from databases such as Ensembl, or NCBI's RefSeq. The abundance profiles report what genomes from the reference collection are present, and what are their quantities.

Analyzing and understanding these profiles can be a challenge since the data they represent are complex. Particularly challenging is their characterization, which is an important task because it can help us understand its functional and compositional characteristics under different biological and environmental conditions. This has practical applications in the process of identifying microbial pathogens in cohorts of microbiomes using supervised machine learning (ML) technologies for associating features with spatial locations, characterizing diseased samples from healthy, treated samples from untreated, and much more.

In this work we present a method for characterizing microbiomes using "microbiome maps" in a custom ML model based on a convolutional neural network and vision transformer architectures. We use a technique called the Hilbert curve visualization (HCV) to create colorful 2D images from abundance profiles. In the maps, color and location provide functional features for the model: locations represent a genome (if whole-genome sequencing), or a set of OTUs (if 16S sequencing); and color can represent their relative abundance in the sample.

Space-filling curves offer an intriguing scheme for characterizing microbiomes in ML models for their ability to preserve positional information. Different linear orderings produce different Hilbert curve visualizations, which result in clusters of related microbes along neighboring regions. This creates abundance "hotspots" that can identify very abundant microbial groups, and used as important features in ML tasks.

We discuss our classification model by analyzing the strain-level abundances of 44K genomes in 328 samples from the Human Microbiome Project. We also evaluate our models with microbiome samples collected from the International Space Station (ISS) and compare the performance of our image-based classifier with that of other well-known classifiers that only use tabular data ("Multi-Layer Perceptron", "Support Vector Machine", "Random Forest", "Naïve Bayes", and "AdaBoost"). When using the tabular-numerical profile, the "Multi-Layer Perceptron" performed the best with 85% accuracy, and "Naïve Bayes" came in second with 84%. "AdaBoost" performed the worst with 51%. In contrast, our classifier had an accuracy of 94%.

Our work highlights the importance of microbiome maps by showing their value goes well beyond visual representations, as they are a powerful asset for ML workflows: they can visualize the dynamicity of thousands of microbial genomes and their corresponding relative abundances, while at the same time embedding the context of microbial relationships and their taxonomies. This embedding creates distinctive features that can be used to characterize sample cohorts, establish dominant taxonomic clades, and identify pathogens.

Individual maps can be interactively explored using the free Jasper software available at "www.microbiomemaps.org", which integrates with resources such as Ensembl, GenBank, and UniProt.