CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)	Project Number
Katherine F. Zhang	**S1621**

Project Title

Deep Learning Analysis of Human Gut Microbiome Metagenomic Data with Applications in Geolocation and Disease Prediction

Abstract

Objectives/Goals

Previous studies of microbiome metagenomic datasets have relied on linear models (PCA, SVM, and RF, etc) and known species and biomarkers in reference databases and ignored about 50% of the reads. I used deep learning directly on DNA kmer abundances to study the human gut microbiomes.

Methods/Materials

DNA sequences from 1030 human microbiomes in four large microbiome metagenomic datasets (HMP, MetaHit, T2D, RA) were first preprocessed into 5-mer counts per sample and then L1 normalized into relative abundances, which were used as features for both unsupervised and supervised learning.

Autoencoder was used on HMP to find whether there is nonlinear structure in the kmer data by comparing best nonlinear model against the linear model.

For supervised learning, the kmer relative abundances were normalized to have zero mean and unit std across training samples. Then, autoencoder was used to pretrain the model, after which its decoding layers were replaced by the final softmax layer for classifying the microbiomes by continent, country, or diseased/healthy.

Results

Analysis of PCA and autoencoder modeling on the microbiome data clearly suggests that there is nonlinear structure. Additionally, supervised learning showed that using only DNA kmer relative abundances as features, we can predict with near-perfect Area Under the Curve (AUC) the continent (0.998) and country (0.989) origins of the microbiome samples while it was previously thought that differentiating between American and European samples would be difficult. The same supervised learning techniques also predicted IBD (0.947) and T2D (0.759) with AUCs exceeding state-of- the-art published results.

Conclusions/Discussion

Using deep learning directly on raw DNA kmer abundances in the microbiome is a very effective approach for studying the human microbiomes, and it can potentially enable scientists to take advantage of unknown organisms as well as new genotypes in the microbiome.

Summary Statement

I showed that deep learning on human gut microbiome metagenomic DNA kmers provided better predictions on both geolocation and diseases such as IBD and T2D than previously published results, which used only linear models on known organisms.

Help Received

Dr. Garud (UCSF Gladstone), Dr. Greenblum (Stanford Medicine), and Professor Pollard (UCSF Gladstone) provided the original ideas and advice. I implemented all the Python code and performed all the experiments as well as analysis.