MetaLyzer: A Novel Analyzer for the Metagenomic Bacteria Using Deep Learning

Sharma, Cameron (School: Mills E. Godwin High School)

Annually, the environmental bacteria cause food-borne illnesses that affect over 50 million persons worldwide. Over half the pharmaceuticals utilize bacterial products. Less than 1% of the environmental bacteria have been studied in-vitro because they do not survive outside their natural habitats such as the biofilms. Metagenomics is the study of the genomes collected from such environmental samples. MetaLyzer identifies bacteria in the metagenomic samples by analyzing the 16S rRNA gene, which is highly conserved. It is a deep-learning based biostatistical model incorporating convolution neural network and support vector machine in an unsupervised learning framework. The model was trained on the samples that were randomly bootstrapped from the pool of 16S rRNA protein and DNA sequences retrieved from the GenBank. The model parameters were finetuned iteratively. The dataset comprised 3,186,451 gene sequences from 1,384 bacterial genera. The model was tested against the five most prevalent genera in the dataset. It successfully identified Salmonella (95.17% accuracy, 81.34% sensitivity and 75.09% specificity), Xanthomonas (95.55%, 75.99% and 83.70%), Escherichia (96.97%, 88.13% and 82.13%), Burkholderia (94.17%, 69.04% and 73.91%) and Mycobacterium (92.8%, 74.29% and 71.56%). The model was further used for differentiating between K12 and O157:H7 variants of E. coli which are commensal and pathogenic strains, respectively. In a first of its kind, MetaLyzer addresses the unmet need for analyzing the environmental bacteria inexpensively and help identify bacterial products for the next generation drug discovery. In future, the model could be extended to other organisms that have conserved genes such as beta-globins in the vertebrates.