ViGAR: Viral Genome Annotation with Recurrent Neural Networks - Facilitating Drug Therapy Research for Thousands of Understudied Viruses

Gupta, Ritvik (School: Johnston High School)

Globally, numerous living organisms suffer from viral infections. Genome annotation is the process of mapping coding-sequences (CDS) of a genome. CDS are the portions of the genome that are translated into mRNA. An infected host cell will utilize this mRNA to synthesize the viral protiens. Understanding CDS locations within the genome aids researchers to unravel the virus's biology, identify the proteins it creates, and discover potential drug therapies. NCBI GenBank serves as a national repository for genomes. Presently, GenBank contains 83,909 viral genomes, yet only 38,567 of them are annotated. 45,342 viral genomes remain unannotated in GenBank, leaving a significant void in our comprehension of various viruses and their subsequent drug therapies. ViGAR (Viral Genome Annotation with Recurrent Neural Networks), helps address this gap. ViGAR utilizes a 2 layer stacked Bidirectional LSTM (bLSTM) network. Each bLSTM layer contained 256 units in both directions. Layer normalization was done between layers. The output layer was a 2 neuron TimeDistributed Dense layer with sigmoid activation. 14,819 genomes were downloaded from NCBI RefSeq to train the model. Data was preprocessed using one-hot encoding. ViGAR was evaluated on 3000 random samples. ViGAR excels at annotating CDS on viral genomes, achieving approximately 96% recall, 93% F-1 score, and 91% precision on test data. Additionally, ViGAR can annotate genome sequences virtually instantly. This research shows LSTMs can achieve higher accuracy and speed than traditional genome annotation models. ViGAR will facilitate fast and accurate drug therapy research for thousands of unannotated viral genomes.