DO NOT CONSIDER FOR TALK

2021 SEA Symposium Abstract

Purdue University

West Lafayette IN

Corresponding Faculty Member: Kari Clase (kclase@purdue.edu)

A Machine Learning Approach to Bacteriophage Function Call Annotation

Yug Rao, Emily Kerstiens, Kari Clase

Bacteriophages are a class of virus that can infect and replicate in bacterial cells, and are the most abundant organism on Earth. There are many applications of recent interest in utilizing bacteriophages in applications such as genetic engineering for antibacterial properties. Two major steps required to understand novel bacteriophages are start site selection and functional annotation. Whereas many algorithms exist to determine candidate start sites for a draft bacteriophage, there are no comparable function prediction tools. This problem is currently solved by tools such as NCBI BLAST, which compare a target gene sequence with every other gene in every bacteriophage – a process that lacks scalability. As more bacteriophages are discovered and annotated, this limitation will only be exacerbated with time. In recent years, powerful machine learning architectures which can derive nonlinear relationships in high dimensional data have been introduced and popularized. Machine learning models such as the Recurrent Neural Network (RNN) or Long Short-Term Memory network (LSTM) are often used to derive meaning for natural language processing, where large strings are parsed to gain insight about their message. In this research, a method is proposed to train a neural network to parse gene base pair sequences to derive the function of the input gene. By doing this, the function prediction would be almost instantaneous, as opposed to the long runtime in BLAST results. It would also be much easier to find functional relationships between smaller phams.