CONSIDER FOR TALK

11th Annual SEA Symposium Abstract

James Madison University

Harrisonburg VA

Corresponding Faculty Member: Steven Cresawn (cresawsg@jmu.edu)



Elise M Rasmussen

A Novel Approach to Improving Automated Bacteriophage Genome Annotation Utilizing Machine Learning

Elise M Rasmussen, Steven G Cresawn

Genome annotation tools such as Glimmer and GeneMark use sophisticated mathematical techniques to model the characteristics of genes, however the quality of these models is fixed from the point at which they are created. They fail to adapt to newly available genome sequences or the refinements in annotations provided by expert human reviewers.  
  
In contrast machine learning utilizes algorithms and statistical modeling to solve problems by relying on learned patterns. It has emerging applications in numerous fields including bioinformatics. Machine learning can be supervised or unsupervised. In supervised machine learning a subject area expert guides the algorithm to the appropriate conclusions. Supervised machine learning is divided into two major processes: regression and classification. Regression is used to predict a continuous output from a given input. Classification predicts the category the data belongs to based on the provided input parameters. It is utilized for predicting discrete responses and was selected over regression as the more appropriate method for gene prediction. After developing a model to predict genes, the model was then compared to expert human-generated annotations or those produced by hidden Markov modeling-based approaches such as Glimmer or GeneMark.  
  
A neural network was created using the TensorFlow machine learning toolkit and the Python programming language. Input parameters for the model included gene length, direction, direction of upstream and downstream genes, distance to the preceding upstream stop codon, and a frequency table of dinucleotides within the coding sequence. The model was trained using nearly all SEA-PHAGES quality-controlled, protein-coding genes from phages that infect Actinobacteria.