Gene function annotation using few-shot learning

Guy Shur and David Burstein

Functional annotation of microbial genes is a pivotal and challenging component of metagenome analysis. After assembly and gene annotation, the gene functions are predicted to allow inference regarding the functional potential of microorganisms in the sample. These annotations are mostly based on sequence or structure similarity of the genes to increasingly large databases of sequences or sequence profiles, and typically the highest similarity score is the one considered for annotation. Weighing in weak similarities to various genes and the usage of advanced deep-learning approaches can be instrumental in improving gene annotation methodologies, especially when there is no sufficiently good match for a particular gene. In this work, we describe a novel methodology to assign function predictions to genes by using a neural network model that considers similarity scores between a putative gene and a large set of gene families. Our method, which uses Siamese Neural Networks (SNNs), is based on a few-shot learning approach. By leveraging the SNN framework, our method can predict gene function among tens of thousands of possible labels despite many sparsely populated gene families in the training dataset. We applied our approach to develop a tool that we call Functional Annotation in Multiple datasets Using Siamese neural networks (FAMUS), which predicts protein function. When trained using the Kyoto Encyclopedia of Genes and Genomes (KEGG)’s protein family database, our model’s implementation achieves a weighted F1 score of 0.915 on a balanced sample of annotated and un-annotated genes from KEGG’s protein sequence database and is accurate at both predicting whether a proper ortholog exists for an input protein sequence as well as assigning the correct ortholog. FAMUS demonstrates that the SNN framework is an effective and efficient approach to labeling gene sequences, particularly given many thousands of possible labels on a sparse dataset.