Protein Function Inference via Artificial Intelligence: Predicting Cancer-Related Gene Functions

Strauss, Charles (School: Los Alamos High School)

Understanding gene function is important for many reasons; in particular, one could examine the gene of a pathogen to find a weakness, or discover a gene in a human related to cancer. Annotating (discovering) the function of every gene without currently known functions remains a grand challenge in biology. Genes code for proteins, and protein-protein interaction can be measured. However, working backward from patterns of protein-protein interaction to function is an unsolved problem. In favorable cases, one can manually infer function, suggesting it could be automated. I hypothesize that an artificial intelligence can classify gene functions from protein interaction data. I compared two unsupervised learning approaches, evaluating their ability to successfully annotate twenty chosen Cancer-related proteins in the human genome. This is a multi-class classification problem because each gene can have multiple functions according to the Gene Ontology dataset. I aimed for classification with low false positives, which is challenging because most genes do not have a given function. The two unsupervised learning methods used were nearest neighbor classification of principal components derived from protein-protein interactions and nearest neighbor classification of data compressed by an autoencoder neural network. I assessed the performance using precision and recall curves. I found that proteins with similar interactions have similar functions and this is useful for annotation, however different methods of unsupervised learning performed differently in functional prediction.

Awards Won: