A Novel Computational Model to Predict Subcellular Protein Localizations

Hu, Kevin (School: Sir John A. Macdonald Secondary School)

Approximately 20,000 different human proteins exist to perform specific tasks within a cell’s organelles focused on building, repairing, and giving signals needed for the body to survive. However, the majority of proteins have yet to be catalogued into a classification system due to lack of knowledge of their subcellular localizations. In 2018, consumer medications targeted only 3% of all proteins. These targeted proteins can occur in multiple areas of the body, damaging non-targeted regions in the form of side effects. This project focused on the development of a deep learning algorithm capable of classification of all possible subcellular localizations of unknown proteins based on fluorescent microscope image data. The classification model used a 34-layers-deep residual neural network for transfer learning. Many training parameters were optimized to deal with the image input, and the model performance was validated using 10-fold cross-validation. The resulting framework takes RGBY image input and produces a CSV file with predicted labels for the aforementioned images. The model was tested on 120,000 additional images and reached an F1 score of 79%, resulting in a 10% improvement over the previous state-of-the-art classifier developed by the Human Protein Atlas. Its ease-of-use and robustness are also improved over other technologies such as CellProfiler. This algorithm can facilitate the mapping of the human proteome, which could lead to developing more target-specific medications, nullifying harmful side effects caused by mislocalization of proteins, and allowing for earlier diagnoses of disease such as certain types of cancers and genetic disorders.

Awards Won: