DIRECT Capstone Project Title and Short Abstract

Binary Classification of Protein Expression from DNA and Amino Acid Sequences Using a Long-Short Term Memory/Convolutional Neural Network

Project Sponsor: Jon Rue (Novo Nordisk)

Project Team: Jay Rutherford (Chemical Engineering), Josh Smith (Chemical Engineering), Christopher Nyambura (Chemical Engineering)

In medicine and biology, there are many applications requiring expression of synthetic proteins, such as biopharmaceuticals and proteomics. To find effective expression systems, high throughput experimentation is used to determine expression and solubility. Predicting properties of synthetic sequences using high performance computing can have a major impact on rapid screening of nucleotide sequences resulting in reduced usage of resources during optimization of biologic products made from transfected eukaryotic or prokaryotic expression platforms. Previous efforts to predict protein properties *in silico* typically rely on features derived from the protein nucleotide and amino acid sequences. In this work, we predict protein expression and solubility directly from the nucleotide or amino acid sequence using a neural network architecture that combines convolutional and long-short term memory layers. Our approach simplifies the process of protein property screening *in silico* and outperforms previously reported models on a classification task for an industrial-scale peptide expression data set. By predicting expression and solubility computationally, high throughput experiments can be focused on sequences with higher likelihood of successful expression, possibly resulting in faster process validation of large scale protein production.