Executive Summary for NLSeer Nuclear Localization Signal Prediction Project

Team Members: Scott Auerbach, Ukamaka Nnyaba, Ming Zhang, Yingyi Guo, Hemaa Selvakumar, and Cisil Karaguzel

Overview: Nuclear localization signals are fragments of a protein sequence that help direct a protein's movement to the nucleus and have been implicated in human diseases and play a major role in many biological functions. In this project, we have implemented several techniques to construct a model specifically designed to predict NLS signals based on the sequence of a protein. These are the position-specific sorting matrix (PSSM), convolutional neural networks (CNN) coupled with XGBoost to optimize performance, and long short-term memory recurrent neural networks (LSTM-RNN). The objective is to be able to generate probabilities of having NLS for any given protein as well as parse through the sequence to identify amino acids that could be part of a NLS motif.

Stakeholders: Life science researchers specializing on molecular mechanisms centered around the nucleus and pathologists working to decipher pathologies of certain diseases.

KPIs (Key Performance Indicators):

- AUC and ROC scores greater than or equal to 0.8
- Ability to predict presence of nuclear localization signals using datasets composed of proteins with and without known NLS sites
- Data visualization will directly display the potential contribution of each amino acid to a NLS
- Comparable or superior performance in terms of accuracy and precision compared to other available NLS prediction tools

Methodology and Results:

PSSMs were generated for amino acid sequences extracted from NLS databases as well as more general databases containing several types of sorting signals. Two types of deep learning methods were employed: convolutional neural networks (CNN) and a long short-term memory recurrent neural network (LSTM-RNN). PSSM data was pre-processed with one-hot encoding. Our results were as follows:

- PSSM as an input for several classification methods had an average accuracy of ~70%
- PSSM + CNN + XGBoost had an accuracy of 91.4% with the main dataset. When tested on a different dataset, the accuracy dropped to ~74%.
- When using a LSTM model, the overall accuracy increased to 85.4%.
- Simple web application capable of NLS prediction and generating graphs indicating possible NLS residues via Logomaker package.

Future Directions:

3-D structures will also be incorporated into the prediction pipeline via node embeddings of graphs generated by Networkx, which in turn are generated from contact maps. We also plan on using a specialized natural language processing model for proteins called ProtBERT.