Erdos Institute Spring 2024 NLS (Nuclear Localization Signal) Prediction Project

Team Members: Scott Auerbach, Ukamaka Nnyaba, Ming Zhang, Yingyi Guo, Hemaa Selvakumar, and Çisil Karagüzel

Datasets: There are two datasets that contain nuclear localization signals: one from ~2008 from the Rost Lab in Munich and another from a team in Japan (published 2016). The latter dataset was found in the Supplementary Information of this paper: <https://www.sciencedirect.com/science/article/pii/S2352340915003613>. The first dataset is saved in our project repository and is available upon request.

Stakeholders: Medical researchers (NLS mutations have been implicated in pathogenesis of certain diseases (source: <https://iubmb.onlinelibrary.wiley.com/doi/10.1002/iub.194>) as well as general molecular biologists who are trying to understand the mechanisms of their processes of interest.

Goal: Build prediction tool using position-specific scoring matrix with SVD for optimization, language-based modeling (NLP), and 3-D structures (using query to BLAST tool or some other database). The data visualization is intended to show the potential or lack thereof of each amino acid’s contribution to a possible NLS throughout the length of the protein sequence.

KPIs (Key Performance Indicators):

* AUC and ROC scores greater than or equal to 0.8
* Ability to predict nuclear localization signals using validation set (Rostlab dataset)
* Data visualization will directly display the potential contribution of each amino acid to a NLS
* Successfully integrate factoring in 3-D structure of a protein to assess potential of being part of a NLS
* Superior performance in terms of accuracy and precision compared to other available NLS prediction tools