**PREDICTING SIGNAL PEPTIDES**

Sheng Han, Feride Eren

**Abstract**

Prediction of signal peptides requires computational approach because they are not responsive to the current crystallization methods. 4 different prediction algorithms are used in this project which are Bernoulli NB, Multinomial NB, Logistic Regression and Linear Support Vector Machine. A stretch of hydrophobic amino acids is both present in signal peptides and transmembrane helices which makes it challenging for signal peptide and transmembrane helix predictors to correctly name the stretches of hydrophobic residues near the N-terminal methionine of a protein sequence. The results for each prediction method…….

**Author Summary**

Not sure about this part

**Introduction**

Signal peptides are extra peptide extension that contains 16–30 amino acids added to the N-terminus of the protein. They function by helping the transport mechanism within the cell to bring it to its specific destination which protein is delivered within the cell. It is removed while the protein is translocated across the endoplasmic reticulum membrane.

One of the most important topics to investigate in bioinformatics is structure and function of membrane proteins. Especially transmembrane proteins control very important function in organism (Yu and Zhang, 2013). To put simply, flow of information and substances in and out of cell is controlled by transmembrane proteins which are also important drug targets (Reynolds et al., 2008). 70% percent of the drug target that are known or being researched are transmembrane proteins (Yildirim et al., 2017)

Confusions at predicting transmembrane protein can occur especially at signal peptides’ strongly hydrophobic and not part of the mature protein region is misclassified as a membrane-spanning portion of a transmembrane protein (Reynolds et al., 2008). These kinds of inaccurate analysis also happen when a transmembrane protein with a membrane-spanning segment near the N-terminus is often misclassified as having a signal peptide.

In this project, it is aimed to train and test a signal peptide classifier based on the data provided and two proteome sets chosen, Drosophila melanogaster and Mus musculus.

Prediction models used in this project are Bernoulli NB, Multinomial NB, Logistic Regression and Linear Support Vector Machine.

**Methods**

**Results**

**Discussion**

**References**

Yıldırım, M., Goh, K., Cusick, M., Barabási, A. and Vidal, M. (2007). Drug—target network. *Nature Biotechnology*, 25(10), pp.1119-1126.

Yu, B. and Zhang, Y. (2013). A Simple Method for Predicting Transmembrane Proteins Based on Wavelet Transform. *International Journal of Biological Sciences*, 9(1), pp.22-33.