TEAM: The Phospho Force

Executive Summary:

The project on Protein Function Prediction (Differentiating Kinases for Targeted Drug Discovery) addresses a critical challenge in drug development by distinguishing between kinases present in prokaryotes, precisely probiotic strains, and those in eukaryotes or pathogenic bacteria. Kinases, a protein essential for cellular signaling, are promising targets for drug discovery due to their involvement in various biological processes. However, the lack of differentiation between kinases across different organisms hinders the development of specific drugs for eukaryotic diseases and infections caused by pathogens.

The primary objective of this project is to develop robust models that can accurately predict the functions of proteins based on their sequences. We collected protein sequences from Kaggle.com (CAFA 5 Protein Function Prediction) to use as valuable training data for the models.

The methodology employed in this project involves several key steps.

- Firstly, the collected protein sequences we preprocessed and feature-engineered to extract relevant information from the data. Some examples of pre-processing include filtering to select only kinases, filtering to select the most frequent GO terms, and randomly subsampling the observations.
- Next, state-of-the-art machine learning techniques, such as deep learning algorithms, are employed to train the models using the protein sequence data. The classification models we chose to trial included random forest, support vector machine, and Keras neural network.
- The models are fine-tuned and optimized to improve their performance in predicting protein functions using different encoding or numeric representation strategies, sub-sampling sizes, layer choice, etc.

Results

- Accuracy was used to assess model performance.
- The Keras neural network performed best with an accuracy of 10-11%. We were able to run this model on our personal computers before the deadline by using k-mer numeric representation, selecting more efficient layers, and reducing the number of observations and classes by extracting the relevant kinase data.

Future steps:

- This model can continue to be improved if allowed more time or computational memory.
- To comprehensively understand the distinctive characteristics and kinase variations among prokaryotes, eukaryotes, and pathogenic bacteria.
- **Comparative proteomic study:** To compare kinase sequences and regulatory elements, enabling the identification of distinctive patterns.
- **Structural biology techniques:** We will examine kinase protein structures, including active sites and binding pockets.
- **Kinase profiling:** To analyze known kinases in prokaryotic strains, eukaryotes, and pathogenic bacteria to identify shared and unique features.

In conclusion, the Differentiating Kinases for Targeted Drug Discovery project represents a promising opportunity to advance drug development. By establishing strategies to differentiate kinases in prokaryotes from those in eukaryotes or pathogenic bacteria, the project opens doors for designing particular drugs for eukaryotic diseases and infections caused by pathogens. The resulting advancements in targeted drug discovery will have far-reaching implications, including personalized therapies, improved.