Protein Function Prediction

1. Finalized team with preliminary project idea
   1. Members: Lara Brindisi, Dhanashree and Yumeng Li
   2. Mentor: Michael Darcy
   3. Preliminary project idea:
      1. predict the biological function of a protein through the Kaggle CAFA 5 Protein Function Prediction competition
      2. develop project to differentiate the kinases from prokaryotes (specifically probiotic strains) from eukaryotes or pathogenic or non-pathogenic bacterial kinases, which will be helpful to design new drugs specific to eukaryotic diseases or infections (pathogens).
2. Data gathering and defining stakeholders + KPIs
   1. Data gathering: training and test datasets provided through the competition <https://www.kaggle.com/competitions/cafa-5-protein-function-prediction/data>
   2. Defining stakeholders: Dhanashree will fill this in
   3. Key Performance Indicators (KPIs):
      1. Accuracy of the model in differentiating kinases
      2. \*Come back to this after data cleaning and preprocessing
3. Data cleaning + preprocessing
   1. Each of us will explore the data, write out an explanation for each file, plan out data cleaning and processing
   2. Familiarize with Git- <https://www.atlassian.com/git>
   3. Use version control
4. Exploratory data analysis + visualizations
   1. Distributions of variables, looking for outliers, etc. Descriptive statistics.
   2. Each of us will start this process using Jupyter Notebook and uploading to the Git Repository
5. Written proposal of modeling approach
   1. Neural network?