Capstone 2 Milestone Report
By Scott W. Lew

Introduction

Proteins are biological polymers made from 20 different amino acids that perform many of the biological processes necessary for life. The important functional roles of proteins include: catalysts for essential biochemical reactions, hormone regulators , neutralizers of pathogens for the immune system, propagators of electrical signals for the nervous system, and many other important tasks. Each protein is also located in a specific compartment or location inside a cell. Knowing the subcellular location of a protein is an essential part of understanding the protein's function.

The goal of this project is to predict what a protein does and where it is located in a cell using different machine learning classification models. In this case, counts of smaller subsequences within the protein sequence are used as model inputs for classification prediction. For this study, 11 protein functions and 7 cellular locations are predicted with Machine Learning models.

Data sets are constructed from files downloaded from the site uniprot.org, a repository of protein sequences from different organisms with different functions. The data consists of protein sequences for different functional classes and from different cellular locations.

This project could be useful to a researcher in a biotechnology company or academia who is interested in determining a novel protein's function and or cellular location. Moreover, machine learning models that predict function and location can serve as an alternative method to traditional methods that compare protein sequence similarity such as BLAST (basic local alignment search tool).

Problem Solving Methodology/Approach

Proteins in this project will be classified by their function and location in a cell with supervised learning classification algorithms. Protein sequences with labels will be used as datasets for Supervised Learning. These protein sequences will be treated essentially as text and analyzed using a Count Vectorizer approach where the count of substrings of different lengths is utilized. Then, the Count Vectors for each protein sequence will be used as inputs to train different machine learning models based on different algorithms such as Naive Bayes, Support Vector Machines (SVM), and

Logistic Regression. Machine learning classification models will then be compared to determine which model has the best accuracy and F1 scores.

Deliverables: Deliverables will include documentation in the form of Google Docs and pdf files, Python code in the form of Jupyter Notebooks, and a slide presentation.

Description of Dataset

The dataset used for prediction of 11 different functions consists of 380,082 protein sequences with appropriate labels.

And the dataset used for prediction of 7 different cellular locations consists of 700,483 protein sequences with appropriate labels.

Datasets

From the website www.uniprot.org, data was collected by downloading multiple FASTA files, which are standard text-based files used in biological sciences for storing protein or nucleic acid sequences. Protein sequences are essentially strings consisting of different combinations of 20 letters found in the English alphabet, where each letter represents an amino acid.

Data Wrangling and Data Cleaning

Characters such as ‘u’,’b’,’x’ are sometimes found in protein sequences to indicate unknown amino acids, and are replaced by the letter ‘g’ from each sequence to facilitate analysis.
Duplicate protein sequences in each dataset were eliminated using pands.drop_duplicates method. However, only protein sequences that were 100% identical were removed from each dataset.

Exploratory Data Analysis and Initial Findings

The most common pentapeptide,5 amino acid substrings, in the function classes and the most common hexapeptides, 6 amino acid substrings, in the location classes were determined using the Natural Language Tool Kit (NLTK) library functions. NLTK is a suite of programs and libraries created for analysis of text and human language data.

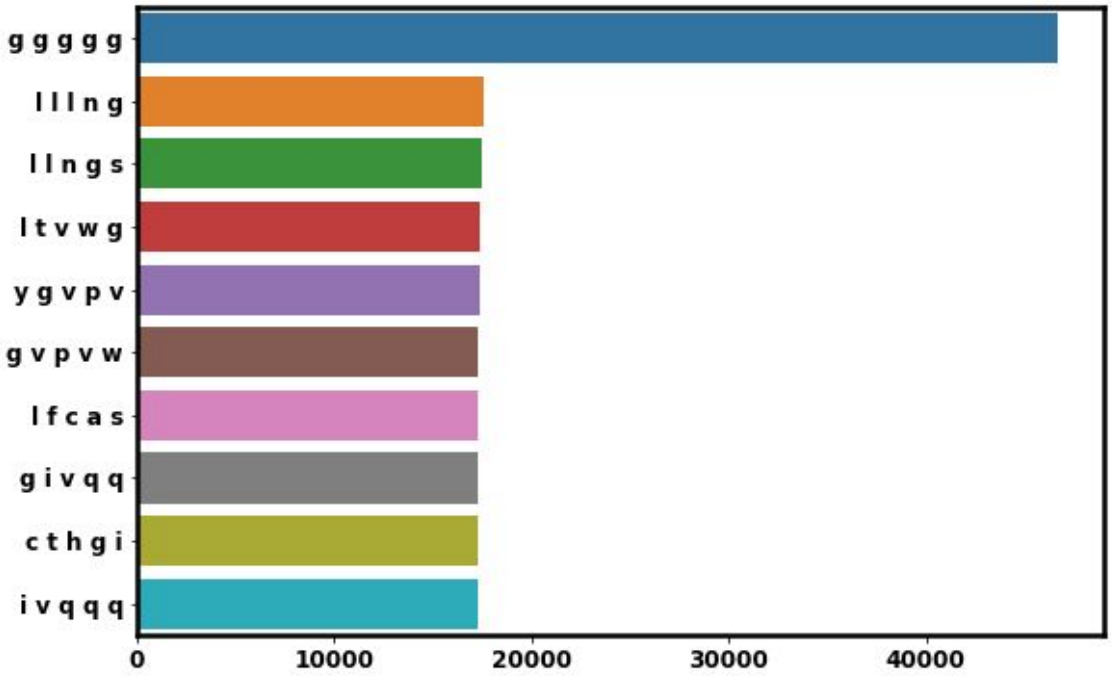

Figure 1 Top 10 Pentapeptides found in all 11 functional classes

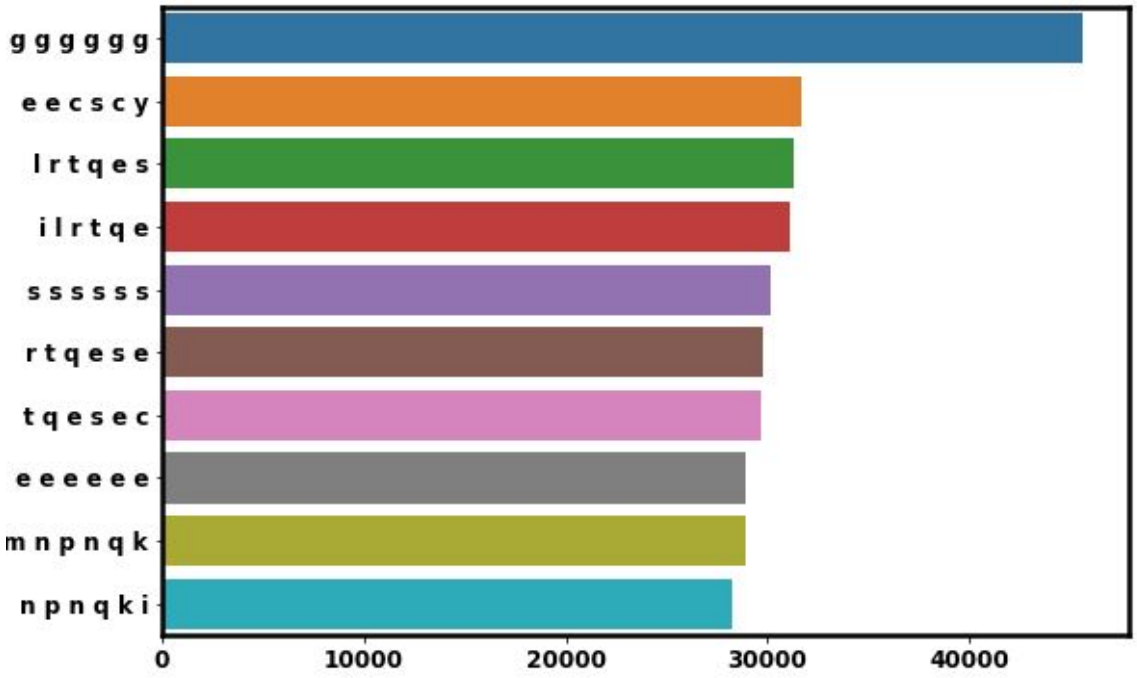

Figure 2 Top 10 Hexapeptides found in all 7 Cellular Location Classes

Fuzzy Wuzzy Analysis:

In biochemistry, proteins often have more than one function. Indeed, the protein's ability to perform one function is often a prerequisite for performing another different function. For instance, a class of proteins known as ABC transporters are responsible for transportation of molecules in and out of cells. But ,in order to transport molecules, the ABC transporter protein must first perform a chemical reaction known as hydrolysis, which breaks down the bonds of the molecule adenosine triphosphate (ATP). Thus, ABC transporter proteins are both transporters and enzymes. FuzzyWuzzy is a python library for string matching and comparison. FuzzyWuzzy analysis can compare two strings and calculate the similarity between the two strings. In this exploratory analysis, a sample of 100 proteins from the hydrolase class were compared with all 37968 proteins found in the transporter class using FuzzyWuzzy to calculate the similarity of each pair of sequences. Based on this analysis, the majority of the hydrolase proteins in the sample had on average a 40% sequence similarity with any of the transporter protein sequences. However, there were a few hydrolase sequences in the sample that

had more than 70% sequence similarity with a transporter sequence that were most likely ABC transporters. As demonstrated by the analysis above, there exist some protein sequences that are common to two separate functional classes, and can be classified as either category. This is an inherent problem in classifying certain protein classes that is not due to mislabelling, but rather an inherent feature of proteins in biology.

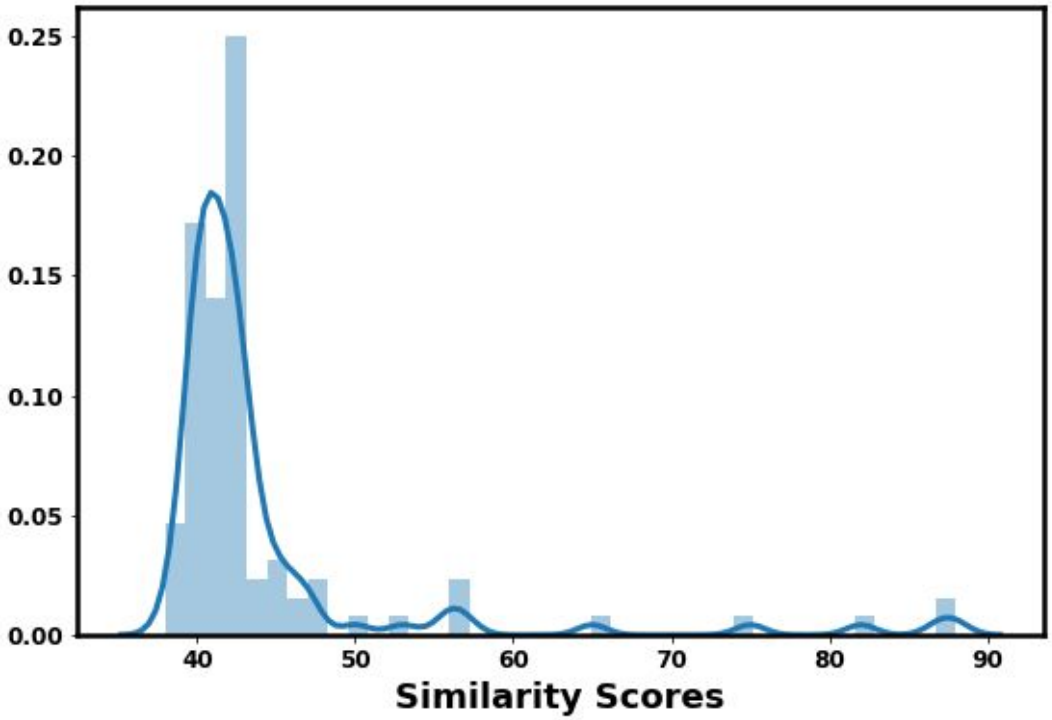

Figure 3 Fuzzy Wuzzy Maximum Similarity scores for 100 hydrolase proteins with respect to transporter sequences. The majority of hydrolase proteins in the sample have less than 50% sequence similarity with a transporter sequence. However, there do exist hydrolase proteins in the sample that have more than 70% sequence similarity with a transporter sequence.