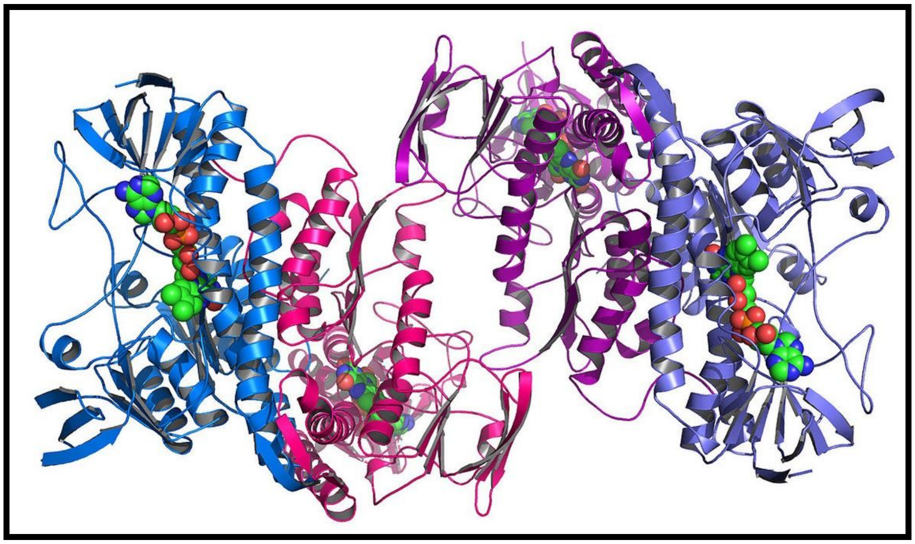
Capstone Project 2: Protein Function & Cellular Location Prediction

By Scott W. Lew



SPRINGBOARD DATA SCIENCE CAREER TRACK

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 www.uniprot.org, a repository of protein sequences from different organisms with different functions 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). Using these Machine Learning methods, a researcher could save time and energy in determining a novel protein’s function and or cellular location.

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.

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. These FASTA files were processed and the data was stored in a csv file for later use.

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 the 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.

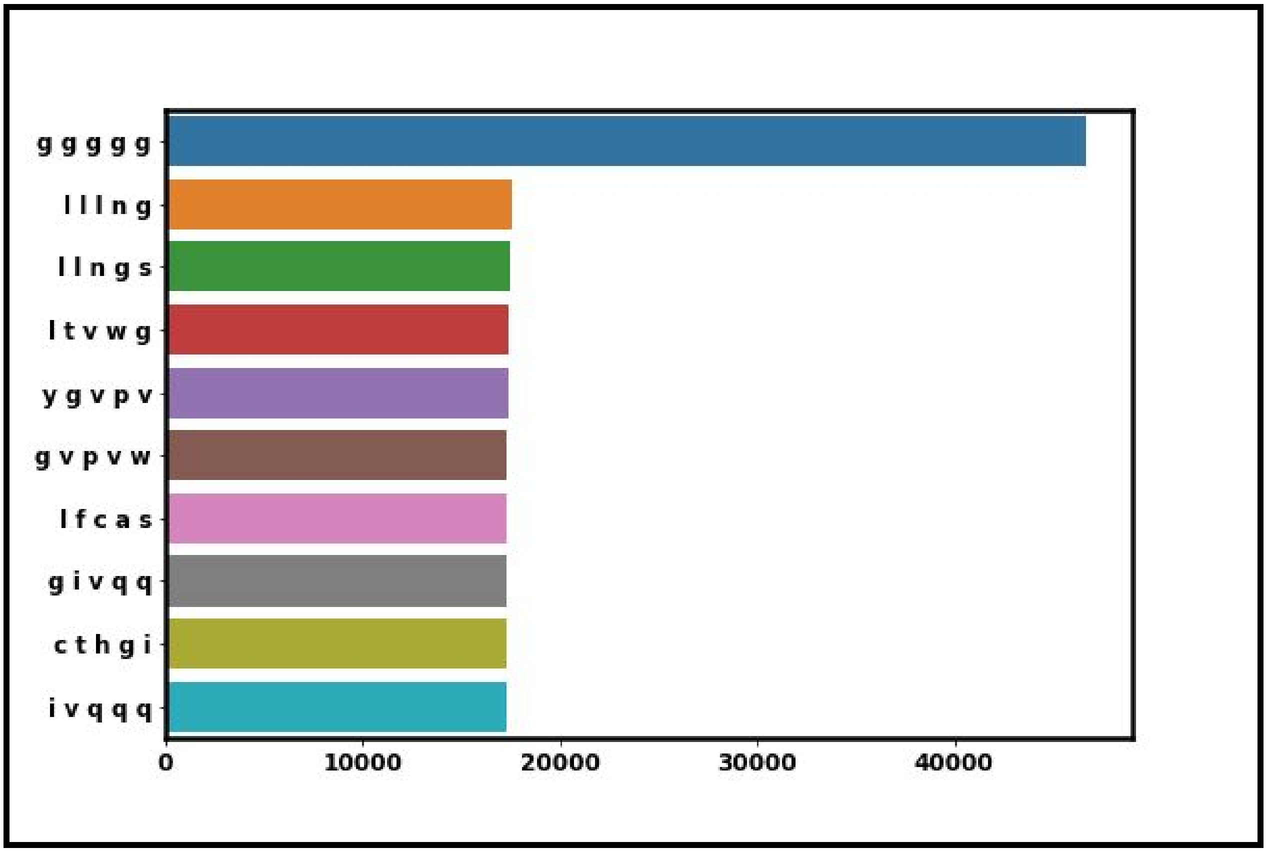


Figure 1 Top 10 Pentapeptides found in all 11 functional classes

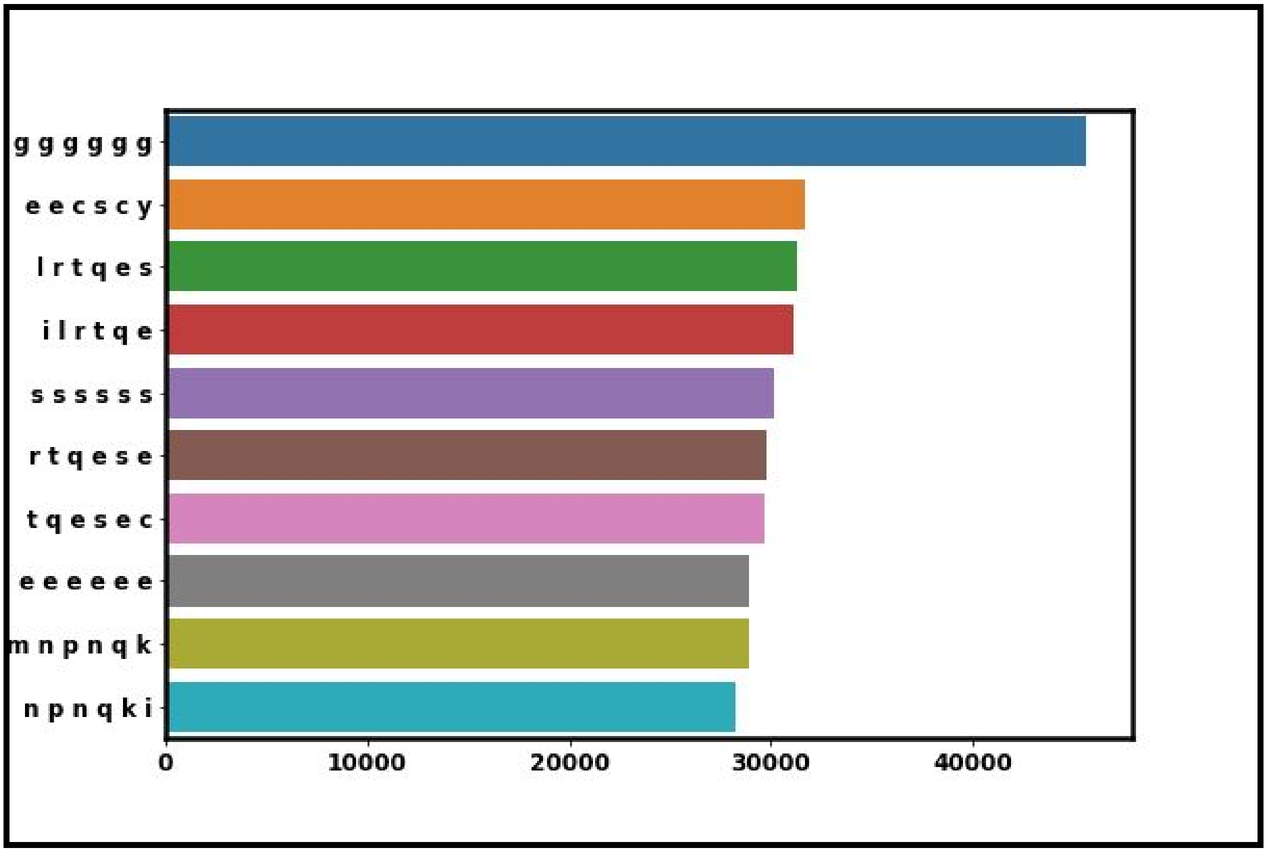


Figure 2 Top 10 Hexapeptides found in all 7 Cellular Location Classes

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 illustrates the similarity comparison between hydrolase and transporter sequences.

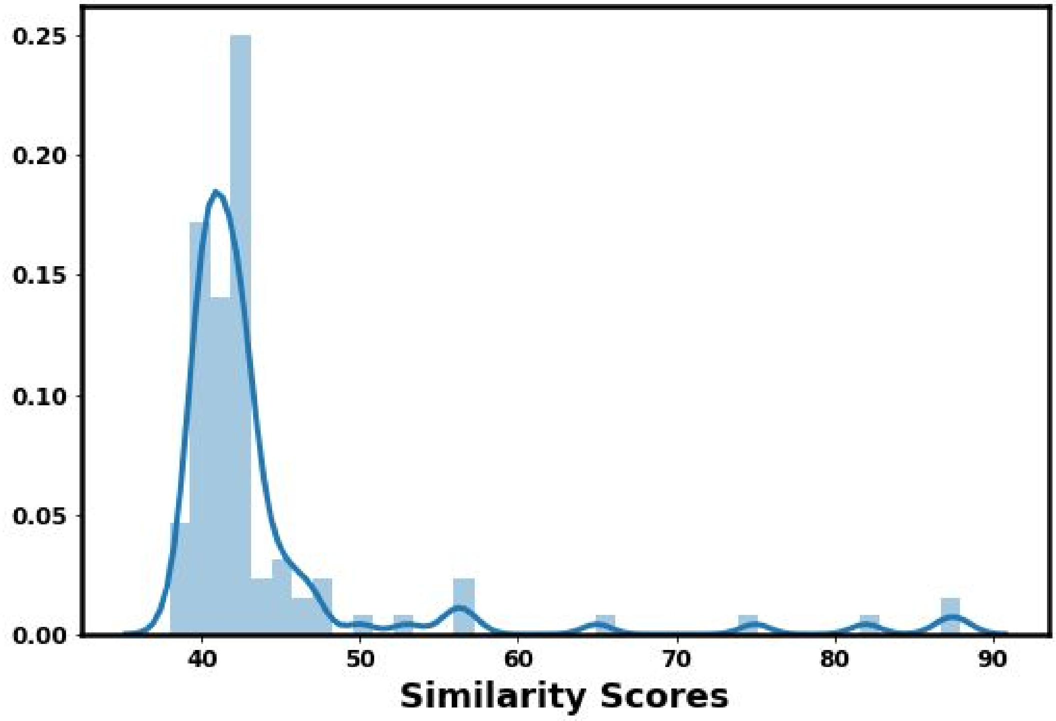
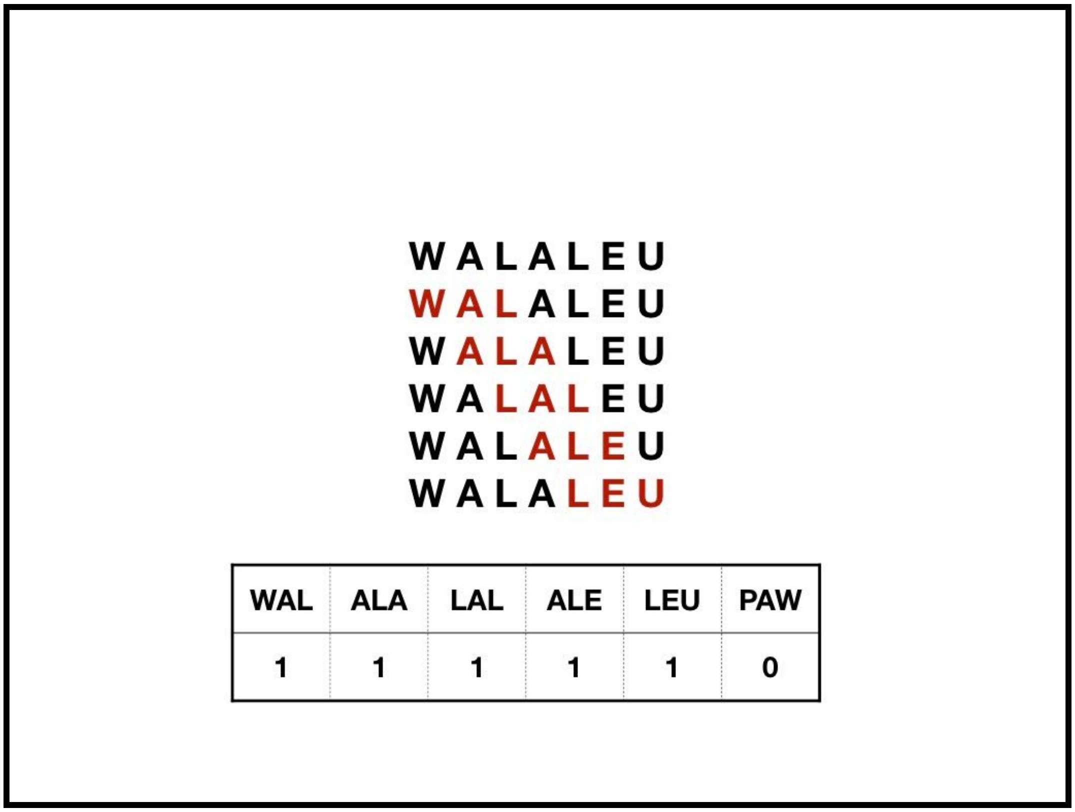


Figure 3 Fuzzy Wuzzy Maximum Similarity scores for 100 hydrolase proteins with respect to transporter sequences. The majority of hydrolase proteins in the sample have at maximum 40% sequence similarity with a transporter sequence. However, there do exist hydrolase proteins in the sample that have more than 70% sequence similarity with a transporter sequence.

Count Vectorizer Approach

A protein sequence is represented by a string of letters where each letter represents one of the twenty naturally occurring amino acids. As shown in Figure 4, essentially, the string of a protein sequence can be broken up into smaller substrings of a desired size,peptides,which is analogous to breaking a stream of text into words. Then, these smaller amino acid substring sequences,peptides, are converted into mathematical vectors with the one-hot-encoding method, where each contiguous sequence of 3-6 letters, amino acids, in the protein sequence are counted. In the example shown in Figure 4, tripeptides made of three amino acids are counted.



WAL	ALA	LAL	ALE	LEU	PAW
1	1	1	1	1	0

Figure 4 Sliding Window Peptide Count Approach for Counting Tripeptides of 3 amino acids in a hypothetical protein

Thus, peptides,substrings, found in the protein sequence are assigned a 1 while those peptides not found in the sequence are assigned a 0.

Machine Learning

Classification

In this project, supervised data, protein sequences with labels, are used to build a model that predicts a label of other protein sequences. And, labels were created by having numerical numbers assigned to each functional class and cellular location category using the SciKit Learning label encoding method found in the sklearn.preprocessing module. Prior to building supervised learning classification models, the dataset is split into Training and Test datasets.

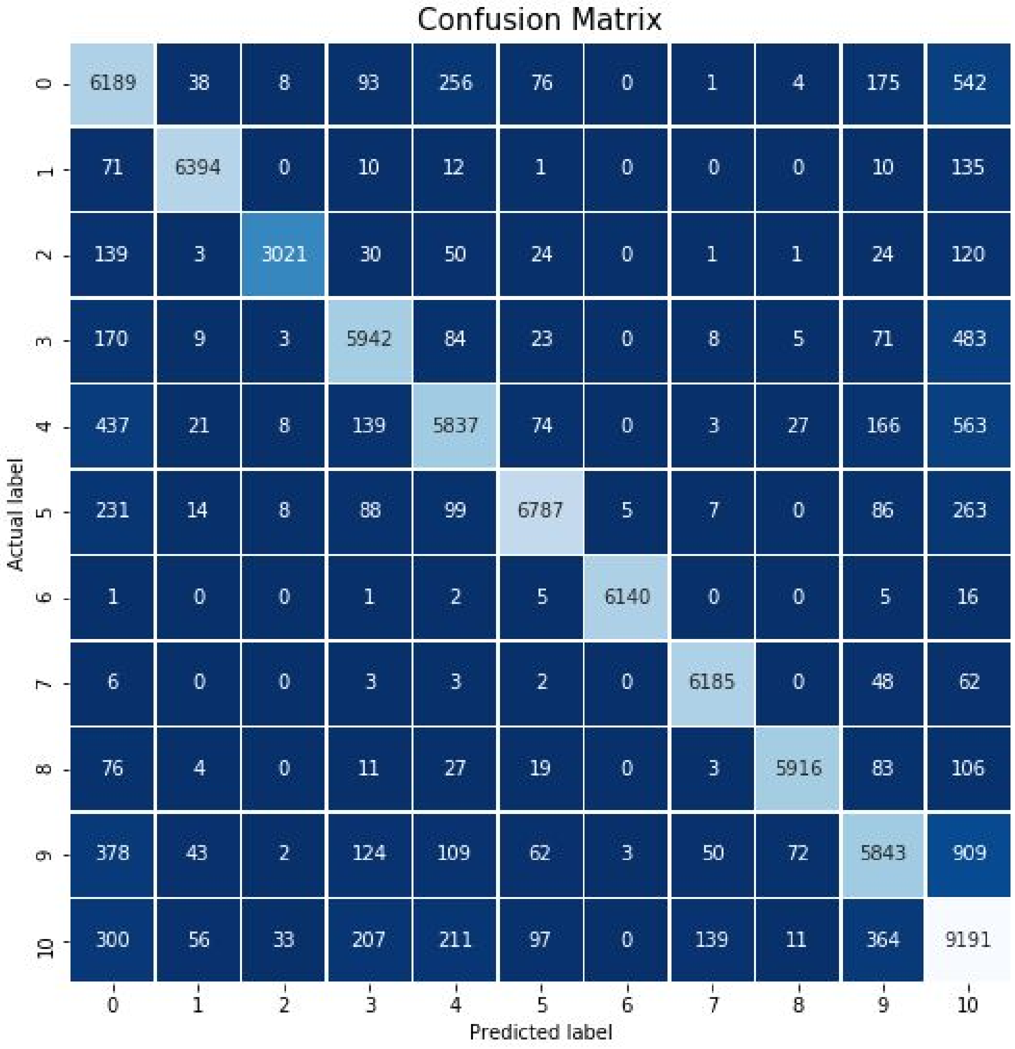
For this project, 11 functional classes and 7 location classes were predicted using different machine learning classification models. The functional classes include: DNA binding, transport, oxidoreductase, hydrolase, transferase, ligase, immune system, GPCR, lipid binding, iron-sulfur-cluster, and isomerase.

And, the 7 cellular location classes that were predicted include: nucleus, endoplasmic reticulum (ER), mitochondria, cytoplasm, Golgi apparatus, lysosome, and plasma membrane.

Logistic Regression, Stochastic Gradient Descent (SGD) Classifier, Multinomial Naive Bayes (MNB), and passive aggressive supervised learning models/algorithms were all employed to predict protein classes. Logistic Regression was used as a baseline classifier, and this algorithm/method achieved ~86% accuracy in predicting function and ~91% accuracy in predicting cellular location. However, Logistic Regression was also the slowest of all the classification models used: training lasted for more than 4 hours on the training data set. SGD Classifier is a linear classifier optimized by the Stochastic Gradient Descent. In addition, Majority Voting Models that were made up of the above mentioned models were used for protein classification. In Majority Voting, a collection of classifiers is used ,and a prediction of a class is determined by selecting the class predicted by the majority of the classifiers in the collection.

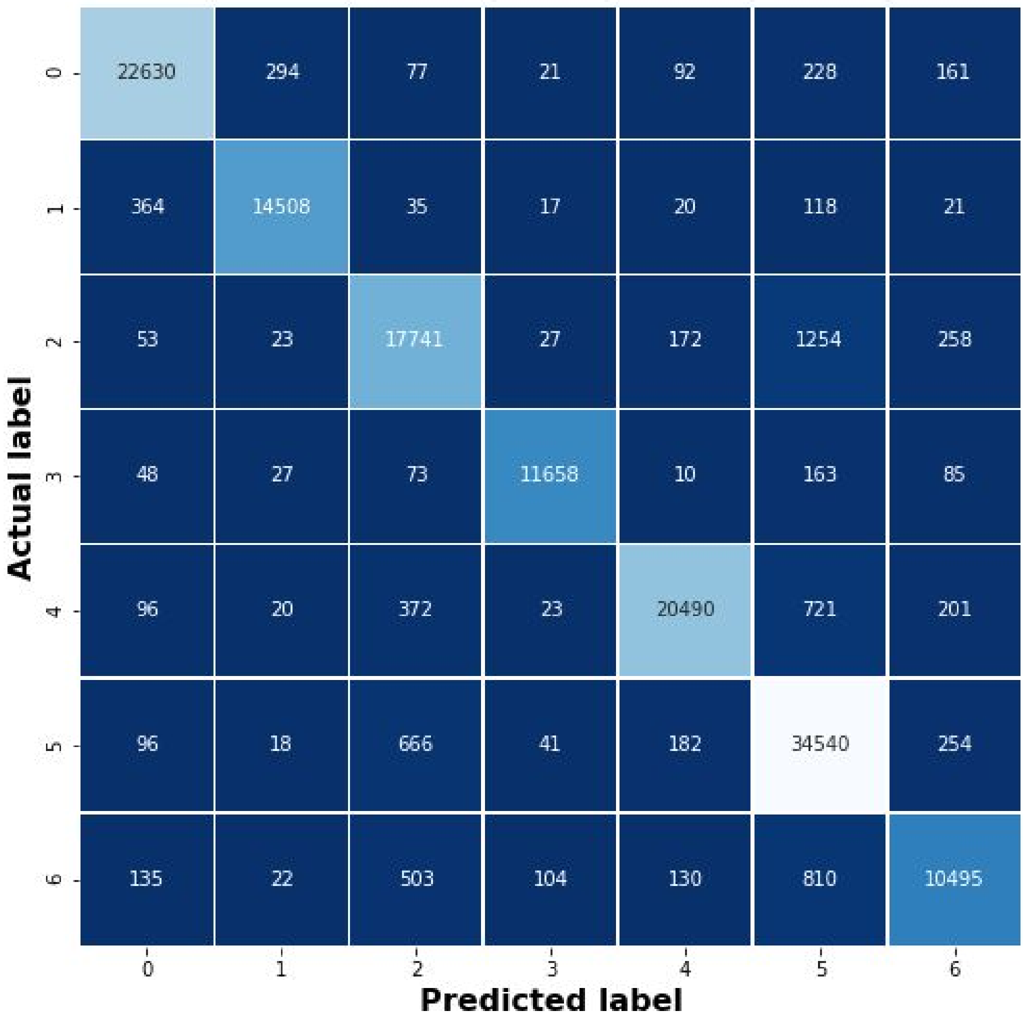
Metrics for Classification

Confusion Matrices for Protein Classifiers



	0	1	2	3	4	5	6	7	8	9	10
0	6189	38	8	93	256	76	0	1	4	175	542
1	71	6394	0	10	12	1	0	0	0	10	135
2	139	3	3021	30	50	24	0	1	1	24	120
3	170	9	3	5942	84	23	0	8	5	71	483
4	437	21	8	139	5837	74	0	3	27	166	563
5	231	14	8	88	99	6787	5	7	0	86	263
6	1	0	0	1	2	5	6140	0	0	5	16
7	6	0	0	3	3	2	0	6185	0	48	62
8	76	4	0	11	27	19	0	3	5916	83	106
9	378	43	2	124	109	62	3	50	72	5843	909
10	300	56	33	207	211	97	0	139	11	364	9191

Fig 5 Confusion Matrix for Voting Classifier of 11 Protein Functions



	0	1	2	3	4	5	6
0	22630	294	77	21	92	228	161
1	364	14508	35	17	20	118	21
2	53	23	17741	27	172	1254	258
3	48	27	73	11658	10	163	85
4	96	20	372	23	20490	721	201
5	96	18	666	41	182	34540	254
6	135	22	503	104	130	810	10495

Fig 6 Confusion Matrix for Voting Classifier of 7 Cellular Locations

A potential problem with protein classification is the overlap between two classes representing proteins that can be classified in either category. In a confusion matrix , the off-diagonal elements generally indicate incorrect classifications made by a model while diagonal elements indicate correct predictions. However, in the function prediction model, off-diagonal elements can also indicate an overlap between 2 category classes rather than a prediction error. As mentioned before, there exist proteins that can be classified in more than one category, and therefore, the model is not necessarily making an incorrect prediction in some cases. This is an inherent attribute of some proteins and is unavoidable.

Cross Validation on Both Models

The Majority Voting Classifier used to predict a protein's cellular location had an average accuracy of 94% as demonstrated by 5-fold cross-validation where the data was divided into 5 subsets.

And, the Majority Voting Classifier used to predict a protein's function had an average accuracy of 89% as demonstrated by 5-fold cross-validation methods.

F1 Scores for Models

The F1 score is an overall measure of a model's performance that takes into account both precision and recall, and is often considered a better metric for a model than accuracy.

Precision is defined as : True Positive/ True Positive + False Positive.

And, recall is defined as: True Positive/ True Positive + False Negative.

The formula for the F1 Score is:
F1 = 2 X precision * recall/precision + recall.

An F1 score of 1 is considered perfect, while an F1 score of 0 indicates the model is a failure. As indicated in the tables below, both classification models are relatively good. For instance, the lowest F1 score for the functional classification is 0.84 and the highest is 1.00 for one of the function categories. In this case, the model had a F1 score of 1.0 when predicting the iron-sulfur cluster proteins.

TABLE I: F1 Scores for Protein Function Prediction Model

Class Category	Precision	Recall	F1 Score
0-Hydrolase	0.82	0.83	0.83
1-Immune System	0.96	0.97	0.97
2-Isomerase	0.91	0.97	0.94
3-Ligase	0.88	0.90	0.89
4-Transferase	0.82	0.86	0.84
5-Oxidoreductase	0.90	0.94	0.92
6-Iron-Sulfur Cluster	1.00	0.99	1.00
8-GPCR	0.98	0.96	0.97
9-Lipid Binding	0.96	0.98	0.97
10-Transport	0.81	0.82	0.81
12-DNA Binding	0.85	0.77	0.81

TABLE II: F1 Scores for Protein Cellular Location Prediction Model

Class Category	Precision	Recall	F1 Score
0-ER	0.96	0.97	0.97
1-Golgi	0.96	0.97	0.97
2-Cytoplasm	0.91	0.90	0.91
3-Lysosome	0.97	0.98	0.97
4-Mitochondria	0.93	0.97	0.95
5-Nucleus	0.96	0.92	0.94
6-Plasma Membrane	0.88	0.89	0.89

Conclusion And Future Directions

The function and cellular location of a novel protein can be predicted using only the protein’s amino acid sequence. Here, two separate machine learning models, built using a peptide count, can predict the function and the cellular location of a novel protein with good accuracy. In the future, it might be possible to extend these methods by predicting other functional and cellular location classes after training with additional data. These machine learning models are an alternative to other established methods like BLAST (basic local alignment search tool) which require having other protein sequences that are highly similar to the unknown protein. It is hoped that machine learning methods like the ones described here will be of use to researchers either in academia or in the biotechnology industry. At the very least, these Machine Learning models can point the researcher in the right direction in determining function and location of a novel protein.