Using Predictive Modeling to Classify Protein Localization Sites in *E. coli*

Dylan Bragdon, Hayden Telson

University of California, San Diego

Abstract

Within any modern society, public health is a key aspect that must be maintained in order to create an environment in which people can thrive without being subjected to diseases that could shorten life spans or cause discomfort. Once such disease causing bacteria is *Escherichia coli*, otherwise known as *E. coli*. While many forms of this bacteria are harmless, or even beneficial, other strains can cause serious gastrointestinal distress. Using the *E. coli* Data Set provided by the UCI Machine Learning Repository [1], we set out to find the best method of generating a predictive model for *E. coli* protein localization sites. To do this we used a three different predictive modeling techniques: K-nearest neighbors, linear discriminant analysis and logistic regression. We then compared each model's prediction accuracy in order to assess which model was able to best predict the location of key *E. coli* protein localization sites.

Introduction

With an ever growing focus on public health, the search to understand and neutralize potential threats to our society's well being is crucial. By understanding the protein localization sites for *E. coli*, we can identify the proteins that make up the bacteria and expand our understanding of the organism as a whole. In order to analyze these sites, having a well selected modeling technique is a necessary place to start. We aim to explore this idea by testing a number of analysis and modeling techniques in hopes of finding one that could be used in the future by the scientific community to understand the behavior of the e coli bacteria. This is already something being analyzed in the microbiology field as seen in [2]. With its potential impacts on the health community and further scientific studies, this analysis of modeling statistics for *E. coli* protein localization offers a multitude of benefits for the understanding of *E. coli* as a whole. We predict that at least one of the following three classification algorithms will help us predict protein localization in *E. coli*: Multinomial Logistic Regression, Linear Discriminant Analysis, and K-Nearest-Neighbors.

Methods

We acquired our dataset from [1]. The dataset consists of 337 samples. Each data point consists of a different *E. coli* sequence. Data points are also associated with a class, or site where the protein is localized. We decided to approach the problem from multiple angles, using a variety of classification algorithms. In total, we used Multinomial Logistic Regression, Linear Discriminant Analysis, and K-Nearest-Neighbors, along with cross-validation, to generate a predictive model for the site of *E. coli* protein localization. For each approach, we used a training size of .75 of the original dataset. Thus, .25 of the dataset was used for testing.

Table 1: Predictors from Original UCI Data Set

Abbreviation	Description
MCG	McGeoch's Method for Signal Sequence Recognition
GVH	Von Heijne's Method for Signal Sequence Recognition
LIP	Von Heijne's Signal Peptidase II Consensus Sequence Score
CHG	Presence of Charge on N-Terminus of Predicted Lipoproteins
AAC	Score of Discriminant Analysis of Amino Acid Content of Outer Membrane and Periplasmic Proteins
ALM1	Score of the ALOM Membrane Spanning Region Prediction Program
ALM2	Score of ALOM Program After Excluding Putative Cleavable Signal Regions From Sequence

Results

Logistic Regression Approach:

We first used multinomial logistic regression to generate a predictive model for *E. coli* protein localization. Our group ended up utilizing five of the seven total predictors above. After performing cross-validation, our group found that LIP and CHG (both binary predictors) had no significant effect on either the training or testing accuracy of the model. Thus, the predictors of our Multinomial Logistic regression model were MCG, GVH, AAC, ALM1, and ALM2. Before cross-validating our results, our group received a testing accuracy of 0.73. After performing k-folds cross validation with 10 folds, the model's testing accuracy lowered to a value of 0.61.

Linear Discriminant Analysis:

The predictors that were used for LDA were the same as those used for the previous approach. We found that, for LDA as well, LIP and CHG did not significantly improve the performance of the model at predicting *E. coli* protein localization. After training and testing the model on the same random sample as in the Logistic Regression approach, we received a training accuracy of 0.89 and a testing accuracy of 0.84. After cross-validating these results, we obtained training and testing accuracies of 0.87 and 0.80 respectively.

K-Nearest Neighbors:

For this approach, we picked an arbitrary 20 nearest neighbors to start. With this approach, we received a test accuracy of 0.83. This score suggests that our KNN model outperforms the Logistic Regression approach, but fails to perform as well as LDA. After performing cross-validation on this model, we received a training accuracy of 0.85, as well as a testing accuracy of 0.75. Our cross validation results are summarized in Table 2.

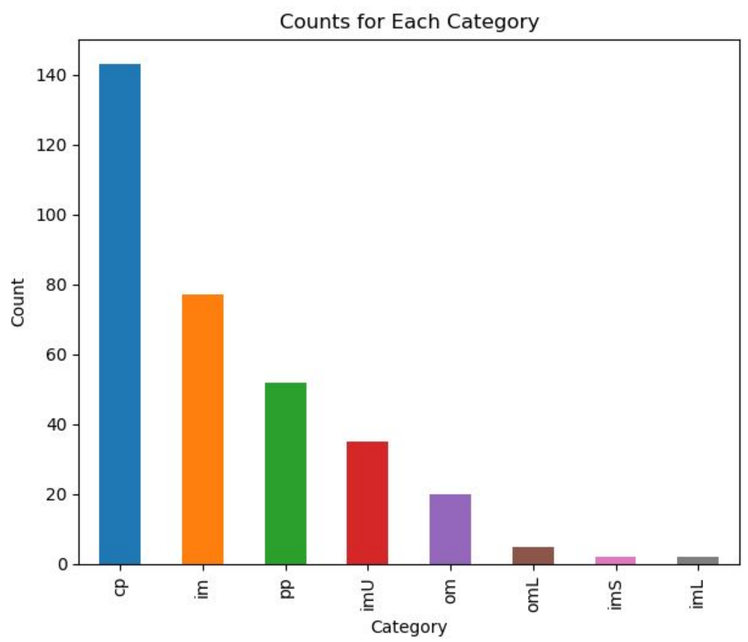
Cross Validation:

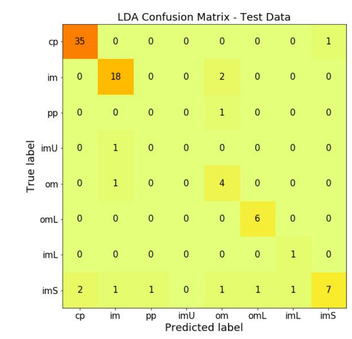
To get a better estimate of the true testing accuracy of each of the above models, we decided to cross validate the results of each model. After running the model with multiple different amounts of folds and seeing no significant effect, we decided on an arbitrary amount of 10 folds.

Table 2: Results from 10-folds cross validation of each model

Model	Training Accuracy	Testing Accuracy
Logistic Regression	0.78	0.61
LDA	0.87	0.80
KNN	0.85	0.75

Figure 1: Distribution of each class (Protein Localization Site) in data set. Notice the imbalance in class distribution.





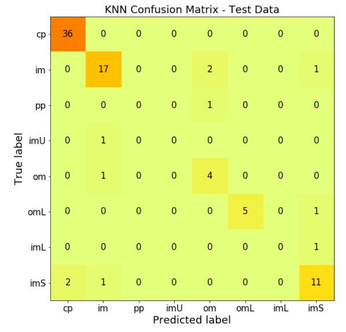


Figure 2: Confusion Matrices for results of each (Non-cross-validated) Classification Method on testing data. Sampling was done once to obtain a 75/25 split between training and testing samples.

Conclusions

After performing cross-validation, it is clear that linear discriminant analysis performed the best at classifying *E. coli* protein localization sites with 80% testing accuracy. While this value isn't terribly low, it may not be confident enough for more rigorous microbiological studies. Our group hypothesizes that LDA results in higher prediction accuracy than linear regression because of LDA's tendency to perform better than logistic regression when the size of the dataset is relatively low. Also, for reasons not explored in this paper, LDA may have performed better on our classification tasks since LDA assumes the explanatory variables to be normally distributed with equal covariance matrices [3]. Perhaps this assumption made by the LDA algorithm closer-models the true distribution of protein localization in *E. Coli*. Our Logistic Regression model's performance may have been sub-par to that of LDA because it did *not* make the same assumptions.

Due to the (relatively) small size of our dataset, all of our predictive modeling methods may be suffering from bias. With more data, we would have a better understanding of the true shape of the data. Thus, all of our models would most likely show improved performance across the board with a larger dataset. To increase prediction accuracy, our group could have also increased the training size for each model. This would have most likely reduced bias and overfitting. With more samples to work with, our model would have been able to train on a dataset that more closely resembles the true distribution of *E. coli* protein localization sites.

As for variance in model performance, our group suspects that the process of cross-validation allowed our group to minimize the observed variance in our samples. But, even with 10 folds, our k-fold approach may not have provided enough flexibility to capture the effect of every single datapoint on the model. In future studies, we advise that a larger number of folds is used, so that it is more likely that each class is being represented equally. Even with 10 folds, there still might be a chance that none of the folds contained samples from lower-represented classes such as imS and imL.

References

[1] https://archive.ics.uci.edu/ml/datasets/ecoli
[2] https://jb.asm.org/content/192/4/912
[3] https://www.stat-d.si/mz/mz1.1/pohar.pdf