COVID-19/SARS/B-CELL EPITOPE PREDICTION

RIYA AGRAWAL

JAMES ALFANO

OVERVIEW

- Introduction
  - Project Goal
- Background Knowledge
  - What is an Epitope?
- Methods
  - The Data
  - The Models
  - The Metrics
- Results
- Conclusion
  - Future Work
- Questions

INTRODUCTION

Epitopes, which are found on antigens, can play a key role in the design of vaccines and in the prevention, diagnosis, and treatment of diseases

Employing computational techniques may improve the accuracy of epitope identification and may reduce the time and potentially the cost of identifying epitopes

INTRODUCTION

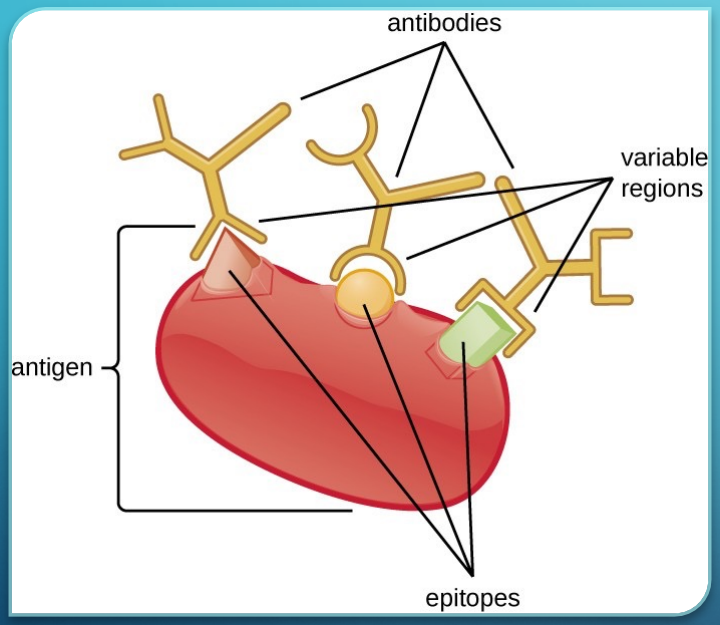
- Within machine learning, researchers have found success using various approaches for epitope identification
  - Traditional machine learning models
    - Support Vector Machines for B-cell epitope Prediction
  - State-of-the-art deep learning classifiers specific to epitope prediction
    - DeepNetBim for human leukocyte antigen (HLA)-peptide binding prediction
    - Epitope3D for conformational epitope prediction

INTRODUCTION: PROJECT GOAL

The purpose of this project is to predict, using binary classification, whether an amino acid peptide exhibits antibody inducing activities

Specifically, this project seeks to predict whether B-cell, SARS, and COVID-19 amino acid peptides are positive or negative for antibody inducing activities

BACKGROUND KNOWLEDGE: WHAT IS AN EPITOPE?



- An antigen is any substance that causes the host's immune system to produce antibodies to fight off an unrecognized substance.
- An epitope is the region of the antigen that is recognized by the host's immune system and upon which the antibody binds

BACKGROUND KNOWLEDGE : WHAT IS AN EPITOPE?

If epitope regions of an antigen can be predicted and then mapped, then this information can be used to design and develop vaccines that will prompt the immune system to produce antibodies to a specific antigen

While the scope of this project focuses on B-cells, SARS, and COVID-19, the knowledge gleaned from the results may have broader applications to other epitopes

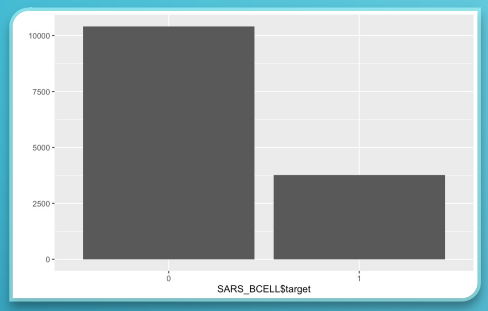
METHODS: THE DATA

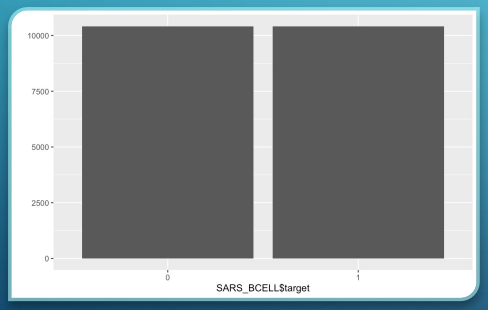
- The dataset (from Kaggle) contains three files each corresponding to different peptides
  - B-cell
  - SARS
  - COVID-19
- Each file contains no missing values
- B-cell and SARS files were combined for training and validation
  - 11, 712 samples for training
  - 3, 195 samples for testing
- COVID-19 data set is used for future prediction
  - The ability/extent to which COVID-19 exhibits antibody inducing activities has yet to be to be identified

METHODS: THE DATA

- The target value is antibody valence
  - Antibody valence is the relative capability of an antibody to bind with an antigen (ability for antibody inducing activities)
- There are 13 Predictor variables used
- Variables that give identity information are not considered for prediction

METHODS: THE DATA





- The target value
  - Originally measured in the dataset as "Positive-High," "Positive-Intermediate," "Positive-Low," "Positive," or "Negative," the variable was transformed by the publishers to be either “Positive” or “Negative” due to an imbalance of data
  - Despite this, the target value was still imbalanced
    - Upsampling was implemented to fix this
    - After upsampling:
      - 16, 372 samples for training
      - 4, 464 samples for testing

METHODS: THE DATA

```
     emini
Min.    : 0.000
1st Qu.: 0.244
Median : 0.551
Mean    : 1.083
3rd Qu.: 1.208
Max.    :40.605
  stability
Min.    :  5.449
1st Qu.: 31.726
Median : 41.948
Mean    : 43.338
3rd Qu.: 49.101
Max.    :137.047
```

- Two predictor variables, stability and emini, were discovered to contain outliers
  - The summary command aided in this discovery
  - The max value is significantly larger than the 3rd quartile value
  - These outliers were removed

METHODS: THE MODELS

4 types of models were implemented

- Logistic Regression
- K-Nearest Neighbor (KNN)
  - K = sqrt(n)
  - K = 5
- Random Forest
- Gradient Boosting Machine (GBM)

All models were tuned with 10-Fold Cross Validation (CV)

Random Forest used 10-Fold CV to tune the hyperparameter “mtry”

METHODS: THE METRICS

- To test model adequacy, 3 main metrics were used
  - Confusion Matrix
  - Accuracy
  - F1 Score

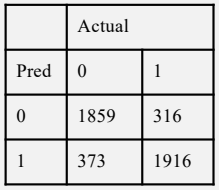
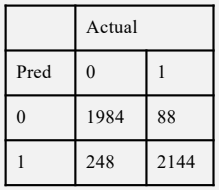
RESULTS

- Evaluation was done on the test data
- The model with the best results is in bold

Model	Confusion Matrix	Accuracy	F1 Score
Logistic Regression	<table><tr><td></td><td colspan="2">Actual</td></tr><tr><td>Pred</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1280</td><td>845</td></tr><tr><td>1</td><td>952</td><td>1387</td></tr></table>	59.0%	58.0%
KNN with k = sqrt(n)	<table><tr><td></td><td colspan="2">Actual</td></tr><tr><td>Pred</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1610</td><td>748</td></tr><tr><td>1</td><td>622</td><td>1484</td></tr></table>	69.0%	70.0%
KNN with k = 5	<table><tr><td></td><td colspan="2">Actual</td></tr><tr><td>Pred</td><td>0</td><td>1</td></tr><tr><td>0</td><td>1611</td><td>380</td></tr><tr><td>1</td><td>571</td><td>1852</td></tr></table>	78.6%	77.7%

RESULTS

- Evaluation was done on the test data
- The model with the best results is in bold

Model	Confusion Matrix	Accuracy	F1 Score
GBM		84.5%	84.0%
Random Forest		92.4%	92.1%

CONCLUSION

- The purpose of this project was to predict, using binary classification, whether an amino acid peptide exhibits antibody inducing activities
  - This task was accomplished with the best model having a 92.4% accuracy and a 92.1% F1 Score

Random Forest outperforms all other models for the given dataset

Balancing the data and hyperparameter tuning gives the best results.

CONCLUSION: FUTURE WORK

- As previously stated, deep learning techniques have shown promise in epitope prediction
  - Applying a neural network to this problem could improve our results
- While our project focused on B-cell, SARS, and COVID-19, our trained models can be used to predict for antibody inducing activities in other epitopes


QUESTIONS?


THANK YOU!