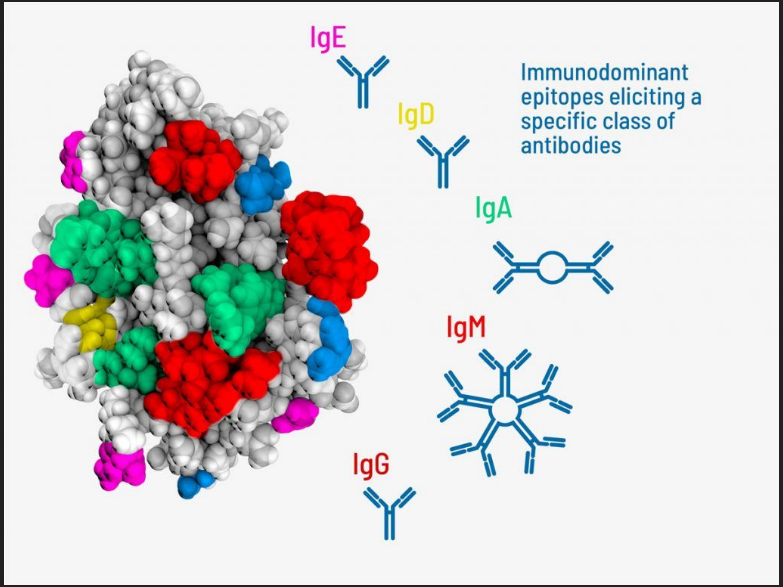
Discontinuous Epitope Feature Extraction, and Analysis

Yatindra Indoria (20CS30060)
Srishty Gandhi (20CS30052)
Devendra Palod (20CS10024)

Introduction

The development of reliable epitope prediction tools is a critical task in immunology and has significant implications for understanding disease pathogenesis, designing vaccines, and developing immune-based cancer therapies. However, experimental methods for epitope mapping are often time-consuming, expensive, and limited by technical challenges. To overcome these limitations, computational methods have been developed


IgE
IgD
Immunodominant epitopes eliciting a specific class of antibodies
IgA
IgM
IgG

Literature

In the method of S¨ollner and Mayer (2006), each epitope is represented using a set of 1487 features extracted from a variety of propensity scales, neighborhood matrices, and respective probability and likelihood values. Of two machine learning methods tested, decision trees and a nearest-neighbor method.....

Recently, a study by Xue et al. (2020):

....To address this issue, the authors introduced two methods, BCPred and FBCPred, for predicting linear B-cell epitopes and flexible length linear B-cell epitopes, respectively, using string kernel-based support vector machine (SVM) classifiers. Both methods demonstrated high predictive performance on independent datasets, with FBCPred outperforming BCPred due to its ability to handle epitopes with flexible lengths.

Data Exploration

1. Epitope aggregate dataset.
2. Dengue, Chikungunya, and Malaria dataset.

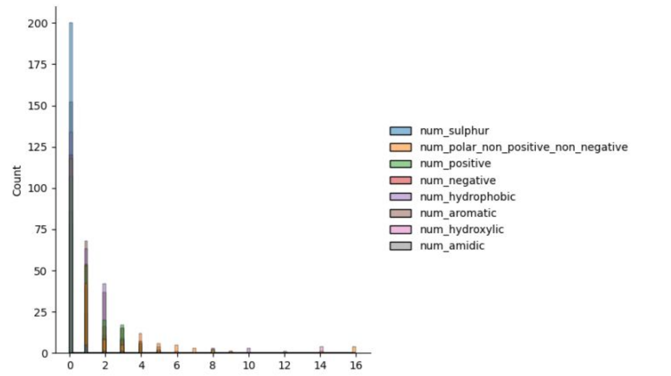
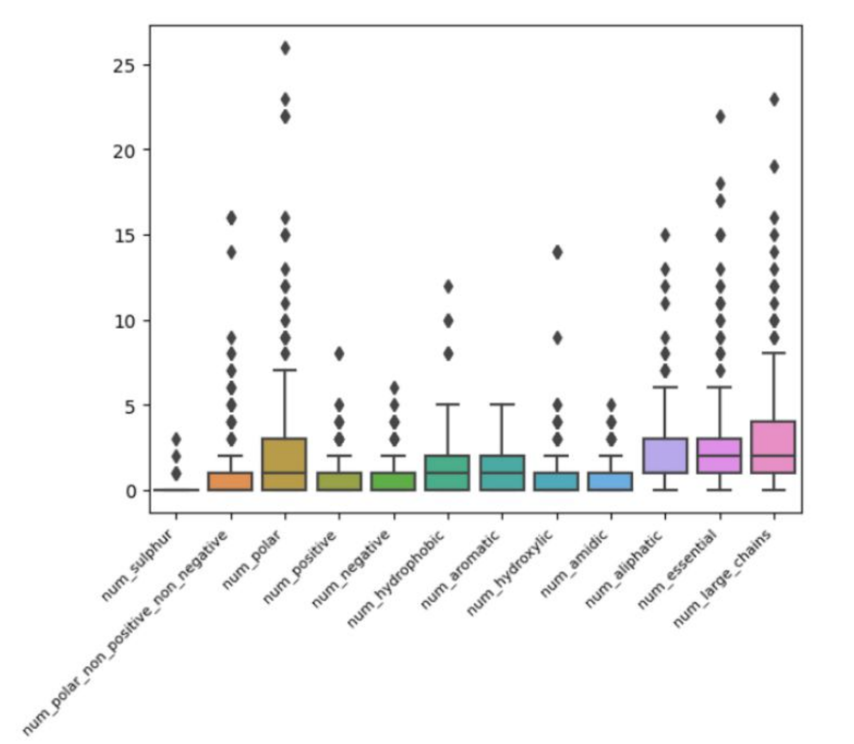


Fig. 6. count plot for the disease Malaria



Methodology

1. We extract 20 custom features from amino acids (like number of large side chains, and SASA). We use a 560 feature AAIndex dataset along with it, but give comparable results with the custom selection.
2. We then analyze the correlation between the features extracted, and due to high correlation we apply PCA.
3. To gauge the effectiveness of the features extracted we run a series of classifiers over 'real', and 'random' epitopes.

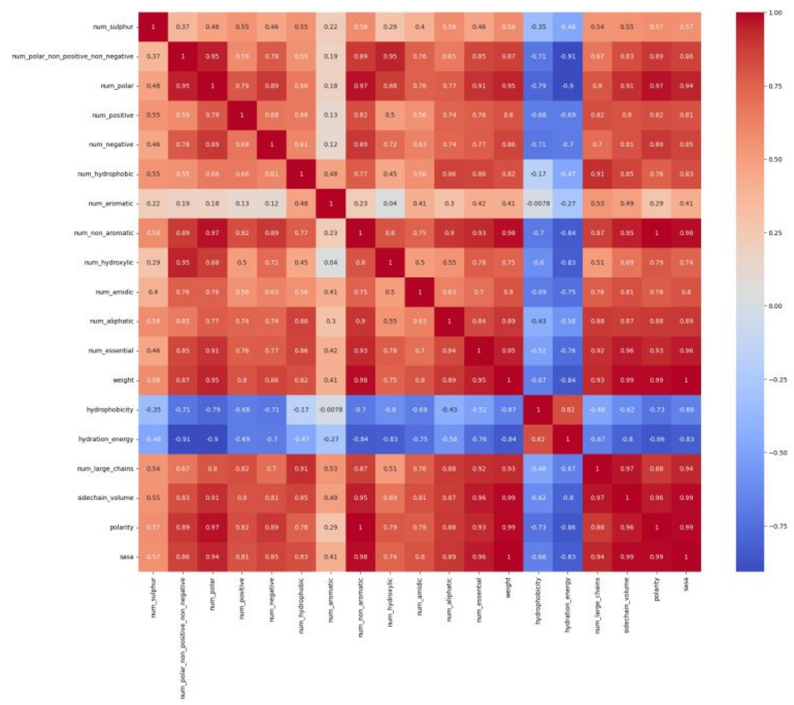


Fig. 11. heat map for the disease malaria

Results

We have obtained 80%+ accuracy in almost all runs.

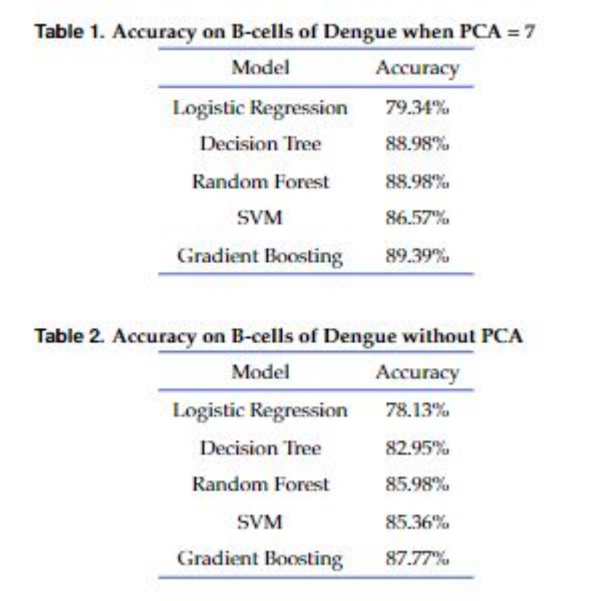
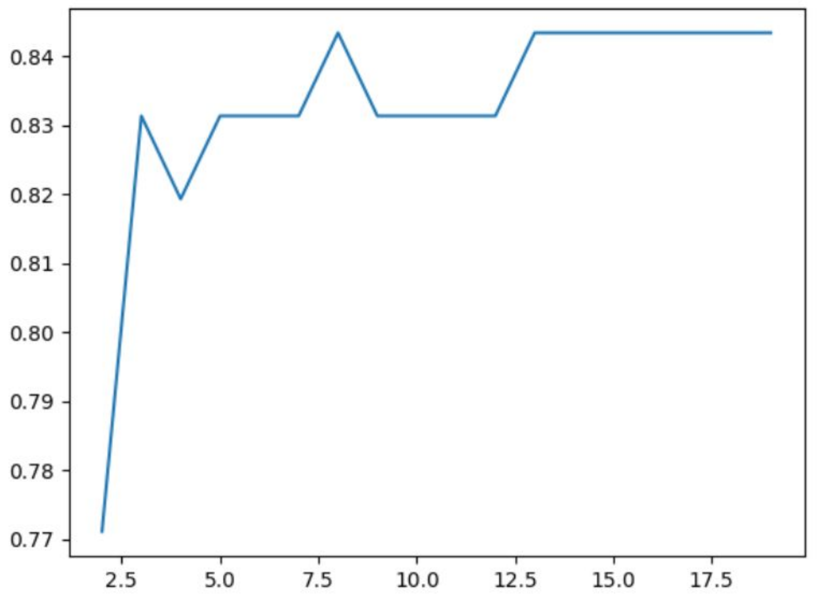


Table 1. Accuracy on B-cells of Dengue when PCA = 7

Model	Accuracy
Logistic Regression	79.34%
Decision Tree	88.98%
Random Forest	88.98%
SVM	86.57%
Gradient Boosting	89.39%

Table 2. Accuracy on B-cells of Dengue without PCA

Model	Accuracy
Logistic Regression	78.13%
Decision Tree	82.95%
Random Forest	85.98%
SVM	85.36%
Gradient Boosting	87.77%



Conclusion

- there is enough physicochemical information in aggregate-based features of an epitope.
- features calculated like this pose a high level of redundancy. Classification accuracy (followed by PCA) does not go downhill even after halving the number of features
- The features can be useful in aggregation with structural features.

Further Work

- Use the features of the base protein.
- incorporate thermodynamic stability, antigenicity, secondary structure fraction,Allergenicity, among other features
- Implement a sequential deep learning model which makes prediction over each residue on whether or not the residue will constitute the epitope.

MOST IMPORTANTLY! RELEASING A PyPi PACKAGE.