B-cell Epitope Analysis and Prediction | Brainstation

Rafael Almazan

June 25th, 2023

Introduction:

B-cells are the immune cells in the body that are responsible for releasing antibodies that target specific antigens. These B-cells identify peptide regions of a protein, binding to it and releasing the specific antibodies. These antibodies inhibit the proteins that they are specified for. These peptide regions are called B-cell epitopes. Detecting B-cell epitopes have crucial impact for vaccine development since vaccines work by inducing an immune response from the immune system. Designing vaccines for viruses that use these epitopes ensure their efficacy and trigger a stronger immune response (Ahmad, 2016). In this analysis, we have information on the amino acid sequence of the peptides, the sequence of their parent protein, and many other peptide and protein features. Of these peptide features, we have the "methods for predicting antibody epitopes" (IEDB) which we will refer to as the methods. These were originally taken from the Immune Epitope Data Base (IEDB) and are identified as the "Chou Fasman", "Emini", "Kolaskar and Tongaonkar", and "Parker" values, named after authors of academic papers. These methods are further explained within the Exploratory Data Analysis Jupyter Notebook. For now, know that these are methods that aid in epitope prediction. The data is fit to multiple different classification machine learning models such as Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and various ensemble learning models. A Recurrent Neural Network architecture was also run, to pick up on patterns within each epitopic peptide sequence. An attention-based LSTM neural network has also been proven to have great performance on epitope prediction (Noumi 2021). However, no code on their attention-based model was provided to the public and this project will aim to replicate their experiment with a clearer implementation.

Exploratory Data Analysis:

To start, we will define our target class. Our target class will be whether a peptide region of the protein will bind to a B-cell and induce an antibody response. The distribution of the target class is imbalanced. There are 10864 peptides that are classified as non-epitopic, and there are 4032 peptides that are classified as epitopic. With this, the epitopic peptides only make up about 28% of the data and may cause complications in the model-building section of this project. Taking a look at the distribution of the peptide methods, It is shown that these are normally distributed with some outliers in the Emini graph (Figure 1).

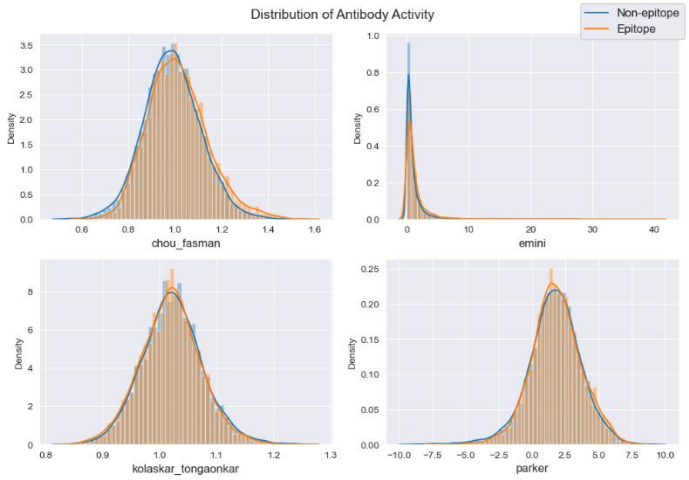

Figure 1.

These distributions also show no difference between the two target classes. This shows that the two target classes are very indistinguishable, even with the features being methods specifically

for epitope prediction (Figure 1). This is a challenge we will face when attempting to predict these epitopes. A PearsonR correlation test was ran between each feature and the target variable.

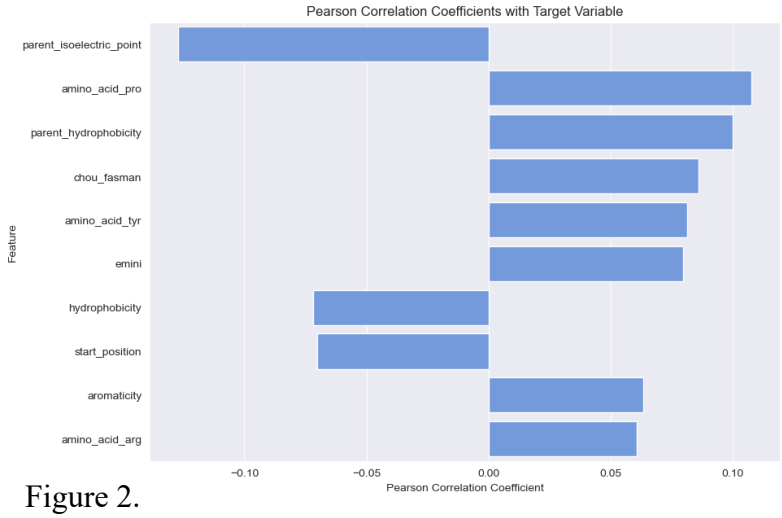

Figure 2.

The correlations of the features to the target class were not very big with the parent protein's isoelectric point being the feature that had the highest correlation magnitude with a value of -0.127 (Figure 2). However, all our correlations seemed to be statistically significant according to the PearsonR test. It is important to remember that statistical significance does not equate to practical significance. The numbers may seem statistically significant but may not have any real-world implications on the reality of the problem. Looking at the Pearson coefficients, we see that the second highest coefficient comes from the presence of a proline amino acid in the peptide's sequence. The presence and abundance of proline within a peptide's sequence may give insight onto how we may differentiate between epitopic and non-epitopic peptides. A proline threshold was defined to be the minimum amount of proline molecules required to meet the threshold criterea. For example, if the proline threshold is 1, then it will include all the peptides that have one or more proline molecules in its sequence. The proline threshold was then graphed alongside the proportions of each target class and the results showed that as the number of proline molecule increase, so does the probability of seeing an epitopic peptide (Figure 3). This suggests that the proline molecules may have an effect on the presence or absence of a B-cell epitope. Proline is unique in that it is naturally very rigid in structure. Due to its rigidity, proline dense areas tend to be more exposed and accessible for immune cells to bind to, which may be the reason for its increased abundance in protein-protein interactions (Kini 1995). The abundance of proline in each peptide will be important to include in feature selection of our models since it definitely gives context to B-cell epitopes.

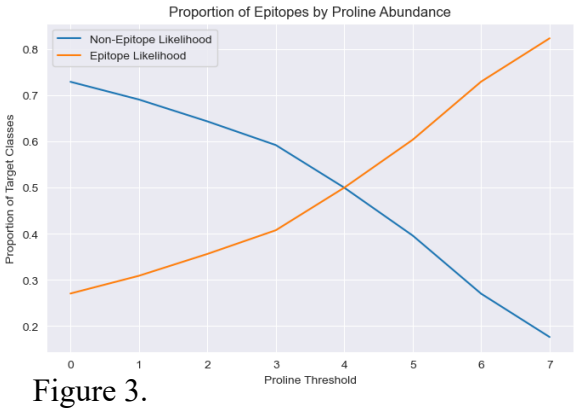

Figure 3.

Machine Learning and Epitope Prediction:

Many different machine learning approaches were experimented on and conducted to find the most optimal machine learning model for the epitope prediction problem space. For all

models, a standard scaler as well as a 70-15-15 train-validation-test split was used. The random state for all models was set to zero to ensure consistency. Since the target class is imbalanced, the f1-score and recall of the epitopic (positive) class will serve as the evaluation metric of our project. Feature selection was done based on a mix of context and performance of our models. A top-down approach was used starting with the methods for epitope prediction (IEDB), followed by the parent protein's features. Two amino acids, Proline and Tyrosine, were also added to the selected features as they were seen as PearsonR coefficients for the target class. To include further context on each peptide within the protein, the start position of the peptide as well as the hydrophobicity and aromaticity were also included in feature selection. Feature reduction through Principal Component Analysis was not done since we do not have many features and it greatly hinders our models' performance. In our final logistic regression, we saw a recall of 9% and an f1-score of 0.17. This poor performance indicates that there is no linearity to our data. The KNN performed better, showing a 41% recall to epitopes as well as a 0.5 f1-score. This is a large jump from the logistic regression, but still not a good performance overall. The SVM with a sigmoid kernel showed the best SVM performance with about 29% recall and 0.29 f1-score. The ensemble learning models seemed to have the best performance. A random forest model with 86 estimators and a maximum depth of 22 showed a 62% recall and 0.7 f1-score which is much higher than any of the other base models seen. Adaptive boosting methods also showed an improved performance of 55% recall and 0.63 f1-score. However, the best performing model was seen to be the XGBoost. The XGBoost model boasted a validation score of 66% recall and 0.71 f1-score. This is the best performing model on our validation set. When ran on the test set, it still performed very nicely with 64% recall and 0.68 f1-score, which is much better than any of the other models in our project. Interestingly, there is a large drop in model performance when training only on general b-cell data and testing on the SARS dataset. This implies that each organism (SARS, Influenza, etc.) may have their own specific patterns and fitting a model on the general data would overfit it to the organisms and viruses that it has already seen. The hope is that the deep learning methods can pick up on patterns within protein and peptide sequences that will be able to predict B-cell Epitopes with a higher performance.

RNN and LSTM:

Applications of LSTM RNNs with an attention layer have proven to be successful in predicting epitopes. In this project, we attempted many different recurrent neural networks, LSTMs, and GRUs on the peptide sequences to see whether neural networks would pick up on any patterns. However, none of these RNNs were able to predict unseen data with high accuracy. The models would strictly only predict one class and were not learning any patterns. One solution may be to apply a masked self-attention mechanism to the bidirectional LSTMs. This would mask the site of the peptide within the parent protein's sequence and force the attention mechanism to apply different weights towards these specific outputs of the bidirectional LSTM. Further experimentation on these models must be done to identify epitopic patterns in these sequences.

Methods:

Data for this project was taken from Kaggle. This dataset was cleaned and taken from the Immune Epitope Database (IEDB) and UniProt (UniProt) by the original publishers. The data came in three tables, two of which had final labels for the target variable, while one did not. The general table as well as peptides specific to SARS proteins had labels and were combined to create one larger dataset. Peptide features were then added using Biopython's protein parameters module. The amount of each amino acid within the peptide sequences were also added to the dataset, giving insight to amino acid abundance within each peptide. Analysis was primarily conducted with Pandas and Machine Learning was done with Sci-kit Learn. Deep Learning was attempted using Pytorch.

References

1. Ahmad, Tarek A., Amrou E. Eweida, and Salah A. Sheweita. "B-cell epitope mapping for the design of vaccines and effective diagnostics." Trials in Vaccinology 5 (2016): 71-83.
2. www.iedb.org
3. The UniProt Consortium, "UniProt: the Universal Protein Knowledgebase" in 2023, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, D523–D531
4. Chou, Peter Y., and Gerald D. Fasman. "Prediction of protein conformation." Biochemistry 13.2 (1974): 222-245.
5. Emini, Emilio A., et al. "Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide." Journal of virology 55.3 (1985): 836-839.
6. Kolaskar, Ashok S., and Prasad C. Tongaonkar. "A semi-empirical method for prediction of antigenic determinants on protein antigens." FEBS letters 276.1-2 (1990): 172-174.
7. Parker, J. M. R., D. Guo, and R. S. Hodges. "New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites." Biochemistry 25.19 (1986): 5425-5432.
8. Kini, R. Manjunatha, and Herbert J. Evans. "A hypothetical structural role for proline residues in the flanking segments of protein-protein interaction sites." Biochemical and biophysical research communications 212.3 (1995): 1115-1124.
9. Noumi, Toshiaki, et al. "Epitope prediction of antigen protein using attention-based LSTM network." Journal of Information Processing 29 (2021): 321-327.