Predicting Linear Epitopes in the Spike Protein of SARS-CoV-2

Natalie Rshaidat

Brown University

Predicting Linear Epitopes in the Spike Protein of SARS-CoV-2

Introduction

With SARS-CoV-2 infecting millions of people around the world, the study of possible vaccines is of grave interest for tackling the spread of the virus. The immune system has two methods of protecting it's cells against infection from the virus. The first method is to use T-cells that can recognize an infected cell and invoke cell death in the infected cell to prevent the infected cell from spreading the virus to other healthy cells. The other method is using B-cell's to produce antibodies that can neutralize the virus by binding to the virus' epitope and essentially blocking the virus from binding to host cells. Studying potential epitopes in the SARS-CoV-2 virus is crucial for creating effective vaccines that can neutralize the virus.

The SARS-CoV-2 virus is able to infect host cells by binding to the ACE2 receptor to gain entry into the cell. The protein that is responsible for binding to the ACE2 receptor is the spike glycoprotein. The spike protein of SARS-CoV-2 contains the S1 subunit and the S2 subunit. The s1 subunit consists of the N-terminal domain (NTD) and the receptor binding domain (RBD). The RBD contains a receptor binding motif that binds to the ACE2 receptor in the host cell, thereby allowing SARS-CoV-2 to enter into the host cell and infect it. Possible vaccines should encode for B-cell antibody proteins that can bind to the S1 subunit, thereby neutralizing the antigen by blocking it from binding to the host cells. Such a vaccine would need to look into the specific epitopes of the spike protein that the antibody would need to bind to.

The epitope of a virus is the region where an antibody can bind to the antigen. There are two types of B-cell epitopes: continuous residues in linear epitopes and discontinuous residues in conformational epitopes (Sanchez-Trincado, Gomez-Perosanz, & Reche, 2017). Machine learning models have provided optimistic results for predicting B-cell linear epitopes. However due to the challenge of effectively representing protein sequences in a computationally efficient manner, thereby motivating the goal of this study (Sanchez-Trincado, Gomez-Perosanz, & Reche, 2017). Using feature extraction methods such as n-grams on the protein sequence, in combination with the physicochemical properties of the protein sequences of differing k-mer decompositions of the spike protein can provide the model with enough information to predict if the protein is an epitope or not.

This study aims to predict epitope locations within the spike protein sequence of the SARS-CoV-2 virus using a random forest classifier trained on known linear B-cell epitopes. Due to the receptor binding domain's role in binding to a host cell and specifically the receptor binding motif's role in binding to the ACE2 receptor, this study hypothesizes that receptor binding motif will contain possible epitopes. Figure 1 shows the annotated locations of possible epitopes.

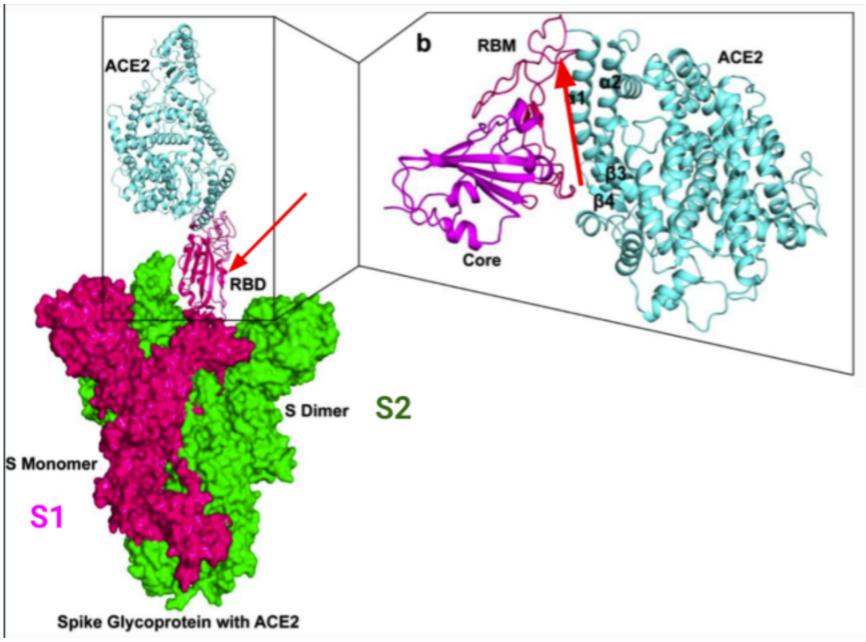


Figure 1. Annotated possible epitope locations on the spike protein image from Fig. 1 by Rathod et al., 2020, (https://link.springer.com/article/10.1007/s40203-020-00055-w/figures/1).

To test this hypothesis, the study will set a high threshold on the classification probabilities of the epitope class when classifying a protein sequence.

Materials and Methods

Training and Testing Data

The training and testing data was collected from BediPred's linear epitope dataset (Jespersen, Peters, Nielsen, & Marcatili, 2017). The dataset was created using the Immune Epitope Database of known linear peptides. It had a collection of positive epitopes and non epitopes (negative). Only linear peptides of lengths between 5 and 25 were included in the dataset, since studies have shown that the distribution of B-Cell linear epitopes lengths lies mostly between 6 residues and 25 residues in length (Kringelum, Nielsen, Padkjær, & Lund, 2013). Peptides that were shown to be positive or negative epitopes in less than two experiments were filtered out of their respective datasets. In total there were 11,834 peptides in the positive epitope data set and 18,722 peptides in the negative epitope data set. In addition, the amino acid sequence of each peptide was in all lowercase letters with a segment in all caps, which indicated the specific epitope sequence within the peptide.

Preprocessing the Data

The Bepipred data set was processed by only selecting the epitope sequence within each peptide for training and testing. Each peptide sequence was labeled as positive (1) or negative (0), to indicate if it was an epitope or not. Additional physicochemical features were added to each peptide sequence using BioPython's protein analysis tools, since a peptide's physicochemical properties affect its ability to bind to other proteins (Cock et al., 2009). The peptide's isoelectric charge in the average pH of human blood (7.5), instability index, and it's gravy score (grand average of hydropathy) were calculated using BioPython and added as features for each peptide sequence. The peptide's gravy score provides the model with information regarding it's hydrophobic or hydrophilic properties of its side-chain. In addition, the protein sequence had to be encoded in a format that a machine learning model can understand.

Encoding Protein Sequences

Encoding the protein sequences required a computationally efficient feature selection method, which directly impacted the model's accuracy in classifying peptide sequences as epitopes or not. This study followed the approach discussed by Iqbal, Faye, Samir, & Md Said in their study of effective feature extraction methods for classification models in bioinformatics (Iqbal, Faye, Samir, & Md Said, 2014). The method uses all combinations of n-gram descriptors frequency as features. The feature set included all combinations of the 20 most common amino acids in pairs or 2 (n = 2). The size of n was chosen to be 2 because it was less computationally expensive as when n was set to 3. Having n set to 2 resulted in 400 features, compared to the 8000 features when n was set to 3. With these extra features our feature set in total included 403 features. Each tuple of combinations was hashed to an index between 0 and 400 to efficiently store counts in a vector. The steps of converting each protein sequence into their feature frequencies included:

1. Decomposing each sequence into their 2-mer subsequences.

2. Creating a vector of length 400 and storing the counts of each 2-mer subsequence in their respective indices
3. Once all 2-mer subsequences are counted, each value of counts in the feature vector gets divided by the total number of amino acids in the sequence.
4. The physicochemical features were calculated and then the feature vector was extended to include these features.

With each peptide's features extracted the data was then split using an 80:20 split. 80% of the data was used for training to avoid overfitting the model, and 20% of the data was left aside for testing the model's accuracy at predicting protein's it hasn't seen before.

Model Selection

A classification model was used for classifying peptides as epitopes. A random forest classifier (RFC) was chosen, because of its promise of improved accuracy over other classification models (Jespersen, Peters, Nielsen, & Marcatili, 2017). After the RFC model was trained on the training data, the model's hyperparameter's were optimized using a cross-validated search over various parameter settings. The number of estimators, which is the number of different decision trees in the forest that the model uses to classify peptides as epitopes or not. Each decision tree in the model processes the features and works its way down the tree until it reaches a leaf node and if it lands in a leaf that classifies the protein as an epitope then the tree will classify the protein as an epitope. The RFC model then counts each tree's vote and depending on which class has the higher number of votes, the model will classify it as that class with the higher vote count. The optimal number of estimators (trees) that produced the highest accuracy, was 800. The other hyperparameter that was optimized was the max depth of each decision tree. The max depth relates to the how many splits of the data the tree should have before classifying a protein sequence. The optimized max depth that produced the highest accuracy score was 250. The last hyperparameter that was optimized was the maximum number of features that a tree uses to classify a sequence. The optimized maximum number of features that produced the highest accuracy was the square root of the number of features, which is about 20 features. The criterion that produced the highest accuracy was the gini impurity criterion.

Spike Protein Sequence Collection and Preprocessing

The protein sequence of the spike glycoprotein was collected from the genome sequence from Wuhan in January of 2020 (Wu et al., 2020). Since the RFC model was trained and tested on peptides of lengths between 6 and 25, the spike sequence had to be decomposed into k-mers with k values from 6 to 25. Each depcomposed k-mer subsequence had its features extracted using the same methods used for preprocessing the Bepipred data set. Once each subsequence had its feature vector populated, it was then passed to the trained RFC model and its prediction and class probabilities were stored. If it was predicted as an epitope, then its epitope class probability was checked against a threshold value. If its probability was higher than the threshold value, then it was stored as a highly likely epitope.

Epitope Classification Threshold

Due to the severity of the implications of predicting a false positive, a threshold of 0.80 was set. A threshold had to be set because it predicted over a thousand epitopes in the subsequences of the spike protein. In addition, some sequences overlapped in their sequences so having a threshold filtered sequences that only had a part of an epitope in its sequence.

Model Evaluation

To test the accuracy of the RFC model, the number of correct predictions was calculated during prediction of the testing set. The model's 95% confidence interval of its error rate was also calculated. In addition, its above the curve score (AUC) was calculated. If the model had an AUC of 0.5 or less, then the model wasn't able to discriminate between epitopes and non-epitopes. This study required the model to have a mean AUC greater than 0.7, as a mean AUC score greater than 0.7 indicated that the model was able to discriminate between epitopes and non-epitopes.

Results

Model Prediction Accuracy

The RFC model's had an accuracy rate of 78.29% on the testing set of 7,636 peptide sequences. The model's 95% confidence interval of its error rate was between 22.73% and 20.67%. In addition, the model's mean AUC score over 10 iterations was 87.7, which means the model is able to discriminate between epitopes and non-epitopes. The model's accuracy was directly related to the features selected in the model. The four most important features in the model were plotted along with their importance score in classification of protein sequences, the plot is shown in Figure 2.

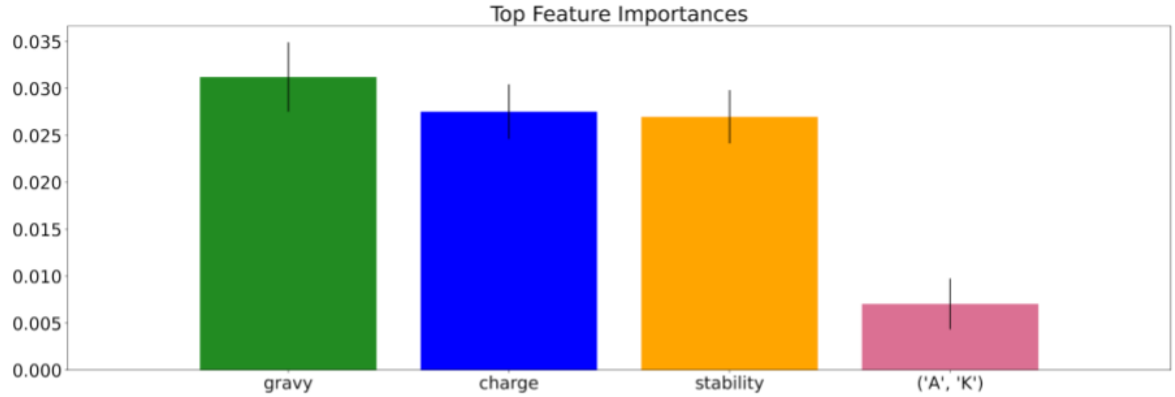
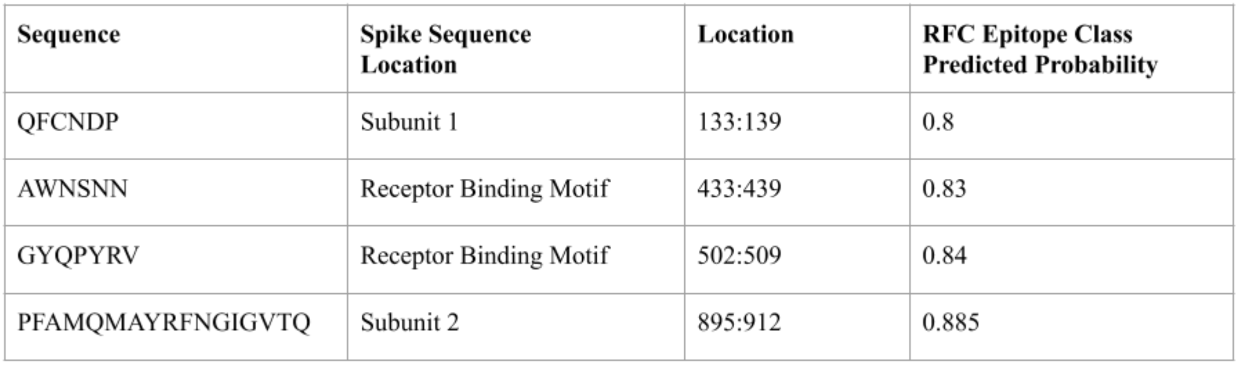


Figure 2. Bar plot of top four importance features used in the random forest classifier model.

The most important features in order of most important to least important was the gravy score, the isoelectric charge in human blood, the instability index, and the AK frequency. This supports this study’s claim that the physicochemical features of a peptide are important when trying to classify if it is an epitope or not. However, the inclusion of the physicochemical features had little effect on the accuracy of the model, which was unexpected. When the physicochemical features were left out the max depth of the classification trees was optimized to 600, which is more than with the physicochemical features. It is possible that the features may not affect the prediction of the peptides, but rather the time it takes for the tree to classify a feature set.

Spike Protein Predictions

The model predicted 72 epitopes in the subsequences of the spike protein that satisfied the 0.8 threshold, but majority of these subsequences overlapped with each other. The more distinct epitopes with highest epitope class probabilities are shown in Table 1, along with their respective location in the spike protein and what protein part of the spike protein they are a part of.

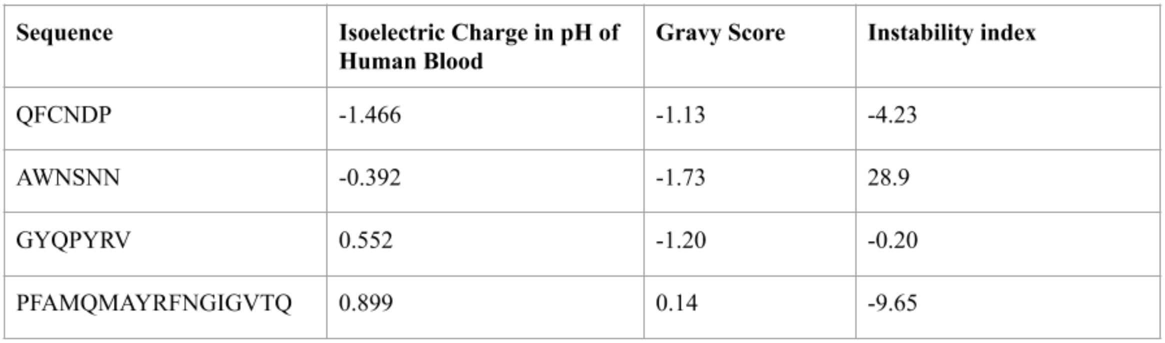


Sequence	Spike Sequence Location	Location	RFC Epitope Class Predicted Probability
QFCNDP	Subunit 1	133:139	0.8
AWNSNN	Receptor Binding Motif	433:439	0.83
GYQPYRV	Receptor Binding Motif	502:509	0.84
PFAMQMAYRFNGIGVTQ	Subunit 2	895:912	0.885

Table 1. The predicted epitopes that were above the probability threshold of 0.8.

The predicted epitopes supported the study’s hypothesis, as it found two very likely epitopes in the receptor binding motif of the spike protein. An unexpected epitope location was the epitope in the subunit 2. However, other studies have shown that that specific epitope interacts with MHCI and MHC II alleles, which is important in immuno-detection (Awadelkareem, Mohammed, & Gaafar, 2020). The “QFCNDP” and the “AWNSNN” predicted epitope sequences were the only epitope sequence that were not in the model’s training data set. The “GYQPYRV” sequence is located in the SARS-CoV spike protein sequence, without any mutations nor substitutions. The “GYQPYRV” predicted epitope sequence was found as part of two epitope sequences “**GYQPYRV**VVLSFELLNAPATV” , and “**GYQ**PYRVVVLSFELLNAPATV” in the training set, which is why it had such a high class probability. In addition, the “PFAMQMAYRFNGIGVTQ” sequence was also found in the SARS-CoV spike protein sequence without any mutations nor substitutions. The “PFAMQMAYRFNGIGVTQ” sequence was located in one of the positive epitope peptide sequences in the data set, which explains its high epitope class probability. Parts of the “PFAMQMAYRFNGIGVTQ” sequence was also located in three other positive epitope sequences, “TAGWTFGAGAALQI**PFA**”, ”**AMQMAYRF**”, and “**GIGVTQNVL**YENQKQI”, which explains for its very high class probability as an epitope.

Further analysis of the physicochemical properties are shown in table 2.



Sequence	Isoelectric Charge in pH of Human Blood	Gravy Score	Instability index
QFCNDP	-1.466	-1.13	-4.23
AWNSNN	-0.392	-1.73	28.9
GYQPYRV	0.552	-1.20	-0.20
PFAMQMAYRFNGIGVTQ	0.899	0.14	-9.65

Table 2. The predicted epitopes physicochemical properties.

The “PFAMQMAYRFNGIGVTQ” epitope was the only epitope in the table that was found to be hydrophobic by its positive gravy score. Hydrophobic proteins are more likely to be membrane bound. There does not appear to be any trend in the physicochemical properties and the model’s epitope class prediction probability. So, even though the features are important, the model still depends on its sequence frequency in its classification.

Conclusions

In conclusion, the majority of the model's predicted epitopes with the highest class probabilities for being an epitope were previously trained on them or at least parts of them. The predicted epitope sequences that it had never seen before with a high epitope probability were the "QFCNDP" sequence and the "AWNSNN" sequence. The results from the spike protein epitope predictions supported the hypothesis that the location of the epitopes would be in the receptor binding motif of the spike protein.

The limitations of this study are that the RFC model replies on past epitope data and its predictions are limited on what it was trained on. The model will predict epitopes that it has seen before much higher than ones it has never seen before. In addition the model was limited in its ability to accurately represent the protein sequences as more expansive feature sets were more computationally expensive.

Discussion

This study was unable to provide evidence in support of its predictions regarding the location of possible epitopes in the reception binding motif of the spike protein. One key finding was the identification of important features that can be used to predict possible epitopes. The study also found limitations in the predictive power of classification models trained on known epitope data sets.

Future explorations could look at the peptide's structural alignments with known conformational B-cell epitopes. Another avenue of research could be the feature representation of the peptide sequences, such as solely physicochemical features and structural features. In addition, future directions should research structural epitopes for the differing structural representations of the spike protein, i.e. the up vs the down conformational structures. Explorations in conserved epitopes that can be effective to the variations of the virus infecting different global populations could also be of value in vaccine production.

References

Awadelkareem, E. A., Mohammed, N. O., & Gaafar, B. B. M. (2020). *Epitope-based peptide vaccine design against spike protein (S) of novel coronavirus (2019-nCoV): An immunoinformatics approach.* (). doi:https://doi.org/10.21203/rs.3.rs-30076/v1

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., . . . Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics [computer software] Oxford University Press.

Iqbal, M. J., Faye, I., Samir, B. B., & Md Said, A. (2014). Efficient feature selection and classification of protein sequence data in bioinformatics. *TheScientificWorld, 2014*, 173869-12. doi:10.1155/2014/173869

Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research, 45*(W1), W24-W29. doi:10.1093/nar/gkx346

Kringelum, J. V., Nielsen, M., Padkjær, S. B., & Lund, O. (2013). Structural analysis of B-cell epitopes in antibody:Protein complexes. *Molecular Immunology, 53*(1-2), 24-34. doi:10.1016/j.molimm.2012.06.001

Rathod, S. B., Prajapati, P. B., Punjabi, L. B., Prajapati, K. N., Chauhan, N., & Mansuri, M. F. (2020). Peptide modelling and screening against human ACE2 and spike glycoprotein RBD of SARS-CoV-2. *In Silico Pharmacology, 8*(1), 3. doi:10.1007/s40203-020-00055-w

Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and methods for T- and B-cell epitope prediction. *Journal of Immunology Research,* *2017*, 1-14. doi:10.1155/2017/2680160

Wu, F., Zhao, S., Yu, B., Chen, Y., Wang, W., Song, Z., . . . Zhang, Y. (2020). A new coronavirus associated with human respiratory disease in china. *Nature (London), 579*(7798), 265-269. doi:10.1038/s41586-020-2008-3