

Progress Report

CSC499 Honours

Zak White

V00899901

February 28, 2022

Motivation

Within the field of biology, machine learning has been able to outperform humans in certain analytical tasks [1], but the use of black box models has been criticized for its absence of explainable output [2]. This project is an exploration of the caveats of machine learning-based approaches that don't offer a logical interpretation of their findings.

This analysis will focus on the results of an existing study (White, 2021), an investigation into mutations in ACE2 proteins that affect a host's susceptibility to SARS-CoV-2. The approach of the study was naive, comprising an artificial intelligence-based algorithm that compared protein sequences without factoring in the biological significance of the findings.

Objective

Existing at the intersection of bioinformatics and machine learning, this analysis aims to answer questions from the perspective of each field:

- Bioinformatics: what do the findings of the original project mean, ie, why might the identified residues be influential?
- Machine learning: was the original approach too naive without considering the explainability of the results? How effectively can the findings be used as an indicator of a sequence's susceptibility to the virus?

Method

1. Design a binary classification model based on the findings of the original project, and use it to classify some new sequences as susceptible or immune to the virus.
2. Investigate the model's misclassified sequences (especially those that are known to be susceptible but are classified as immune) to recognize the naivety of the original method.
3. Explore the original findings to better understand why some sites may be influential, and isolate those that are explainably influential.
4. Construct a new classification model based on the reduced findings and compare with the original.

Outline

The table below offers an initial project outline with estimated time frames.

Research & planning: <ul style="list-style-type: none">• Construct project outline and task breakdown• Research related work (similar studies or other works that focus on the application of machine learning in biology)	Jan 31 - Feb 13
Baseline model: <ul style="list-style-type: none">• Collect ACE2 protein sequences from various host organisms based on research indicating the host's susceptibility to SARS-CoV-2• Create a binary classification model based on the findings of the original study• Use the model to classify the sequences and evaluate the success of the model	Feb 14 - 27
Progress Report	Feb 28
Explaining the results: <ul style="list-style-type: none">• Use a structural analysis and results from other studies to identify key residues that affect susceptibility to the virus, drawing connections to the results of the original study• Create a subset of the key residues of the baseline study including only residues with an explainable influence over susceptibility	Feb 28 - Mar 13
Optimized model: <ul style="list-style-type: none">• Use the reduced findings to create a new binary classification model, and use the same ACE2 sequences to evaluate its correctness	Mar 14 - 27
Comparison and report	Mar 28 - Apr 22
Final Report and Presentation	Apr 22

Preliminary Work

Data

To date, twenty-four ACE2 sequences have been collected from a range of mammal hosts from the NCBI protein database [4]. Sixteen the sequences belong to hosts identified to be susceptible to the virus; the other eight are considered immune. More sequences may be retrieved as the project proceeds.

Baseline model

The approach used for the baseline classification model was a discrete binary decision tree, which takes as input a vector of features corresponding to the acids in each influential site of the original study. The model checks if any of the features match the specific mutation identified in the study, and if so, classifies the sequence as insusceptible. If no matches are found, the model classifies the sequences as susceptible.

Pos	Mutation		
31	Lysine (K)	→	Aspartate (D)
41	Tyrosine (Y)	→	Alanine (A)
66	Glycine (G) Arginine (R)	⇒	Alanine (A)
83	Tyrosine (Y)	→	Phenylalanine (F)
113	Serine (S) Arginine (R)	⇒	Asparagine (N)
353	Lysine (K)	→	Histidine (H)
426	Proline (P)	→	Serine (S)
679	Isoleucine (I)	→	Valine (V)

Table 1: Mutations that may abolish susceptibility [3]

In short, if any of the mutations from the original study (summarized in the above table) are identified in a sequence, the model classifies the sequence as insusceptible.

Results

Below is the confusion matrix for the model based on the results of classifying the twenty-four sequences. The positive class corresponds to sequences that are susceptible to the virus.

		Classification	
		Positive	Negative
Positive	Positive	10	6
	Negative	0	8

By looking at the confusion matrix, we can see that each of the negative sequences were correctly classified. However, six of the positive sequences were misclassified as negative (false negatives), meaning the model has low sensitivity. By reducing the mutations the model checks for, the optimized model should produce fewer misclassifications.

References

- [1] Kakadiaris, Vrigkas, Yen, Kuznetsova, Budoff, and Naghavi, “Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA,” *Journal of the American Heart Association*, vol. 27, no. 22, 2018. DOI: 10.1161/JAHA.118.009476.
- [2] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.
- [3] Z. White, “Identifying Mutations in ACE2 That Influence Susceptibility to SARS-CoV-2,” Nov. 2021.
- [4] “Protein - The National Center for Biotechnology Information.” (2021), [Online]. Available: <https://www.ncbi.nlm.nih.gov/protein> (visited on 02/20/2022).