Construction of Highly Accurate SARS-COV2 Variant Classifier through Machine Learning

Cynthia (Xiao Lei) Du

Department of Bioinformatics, University of Guelph

BINF 6210: Software Tools for Biological Data Analysis and Organization

Dr. Sarah (Sally) Adamowicz

October 27, 2023

Introduction

SARS-COV2 (COVID-19) impacted modern society in an unprecedented way, changing government policies, altering societal norms, driving vaccine advancements, and fostering drastic economic changes (Hosseinzadeh et al., 2022). One major factor driving COVID-19 virulence and persistence is their high viral mutation rate, giving rise to a multitude of variants, including major strains like Alpha, Beta, Delta, and Omicron (Markov et al., 2023). Although each strain differs in the range of mutations present, a major indicator differentiating between strains is mutations to the surface glycoproteins (spike proteins) responsible for entry and invasion of the host cell (Markov et al., 2023). This project attempts to create a high-accuracy identifier of COVID-19 major variants based on spike protein sequence differences to aid the classification and investigation of COVID-19 variants. We derive COVID-19 sequence data from the NCBI nucleotide database, isolate spike protein sequences, and explore different distance measures, parameters, and machine learning algorithms and their impact on the accuracy of the identifier. Finally, we identify the best-performing algorithm and present a visual representation of a highly representative decision tree with the best-predicting features. Accurately classifying different strains of COVID-19 based on spike sequence may help scientists better investigate and understand how the gene has changed over time, aid in future classification of viral sequences, analyze key differences between strains, and perhaps better predict future changes. Organizations may also use this classifier to determine the prevalence of currently circulating strains in the population. Overall, the product of this project is versatile and has potential in many biological and bioinformatic applications.

Conclusion and Discussion

After building and evaluating our machine learning models, we found that our Random Forest Model of k-mer 3 was the most accurate (98.9%) at classifying COVID-19 variants based on spike protein gene sequence, further reaching 100% accuracy on Alpha and Omicron variants. As expected, different measures of accuracy gave different results, and accuracy increased as we increased k-mer lengths up to a certain point. This finding is supported by Zhao et al. (2020), who showed in their paper (fig.3) that increasing k-mer size causes an S-shaped accuracy curve, plateauing at its upper limits. Similarly, we also saw that accuracy increases with k-mer size that plateaued after a certain point, especially for Random Forest Classification (Table 1; no increase between k-mer 3 and k-mer 4). Our data exploration step also indicated that our data had a class imbalance, wherein our Beta variant had 80%-90% less data than all the other classes. As expected, this affected the accuracy of our Beta variant classification, resulting in much lower accuracy for Beta variant classification compared to other variants (Table 2). Interestingly, our K-NN model of k-mer 3 was able to overcome the data limitations and provide an accuracy of 92.9% for the Beta variant classification, much higher compared to the highest achieved accuracy for the Beta variant by Random Forest (85.7%). This suggests that different models have different advantages, and it may be interesting to explore which models have the highest tolerance for class imbalance. In this project, we saw the effect of class imbalance when not rigorously addressed, but also found that some models are able to overcome class imbalance. Given more time, it may also be interesting to compare different methods to address class imbalance, such as by undersampling and decreasing the overall sample size or oversampling by artificially generating more data through randomized removal/modification of a single nucleotide from existing Beta variant sequences. Lastly, an interesting phenomenon was observed where Random Forest models based on k-mer 3 and k-mer 4 produced different training outcomes (estimated error of 1.92% vs 1.48% respectively), but performed identically on the validation data (Table 1) and produced the same confusion matrix with 6 incorrectly identified variants. We theorize that the 6 incorrectly identified sequences in this particular division of validation data may have differences that make them difficult to accurately group. Perhaps this could mean that they could represent intermediate states between different strains, since variants were being sequenced in real time as new variants diverged. Another next step would be isolating and examining (eg. via alignment) these particular sequences to determine SNP differences between their observed classification and their incorrectly predicted classification. All in all, we generated a highly accurate model for the classification of COVID-19 variants based on spike protein variation. This also confirms that the spike protein is one of the major targets of gene mutation between COVID-19 variants (Markov et al., 2023).

Acknowledgements

I would like to thank Dr. Sarah Adamowicz for providing fantastic example scripts in class. I am also grateful to contributors on stack overflow for describing ways to visualize Random Forest trees and the publicly available regex cheat sheets which I consulted thoroughly to properly isolate spike protein CDS.

Works Cited

Hosseinzadeh, P., Zareipour, M., Baljani, E., & Moradali, M. R. (2022). Social consequences of the COVID-19 pandemic. A systematic review. *Investigación y Educación En Enfermería*, *40*(1). https://doi.org/10.17533/udea.iee.v40n1e10

Markov, P. V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N. I., & Katzourakis, A. (2023). The evolution of SARS-COV-2. *Nature Reviews Microbiology*, *21*(6), 361–379. https://doi.org/10.1038/s41579-023-00878-2

Zhao, Z., Cristian, A., & Rosen, G. (2020). Keeping up with the genomes: Efficient learning of our increasing knowledge of the Tree of Life. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-03744-7

Figures

Table 1. Prediction accuracies on validation data for Random Forest (RF) models on various distance measures and features. Overall accuracy is calculated as correct/total identified variants. Weighted accuracy is an average of per variant accuracy. Finally, per variant accuracy is calculated based on correctly identified variants/all variants of that class.

Accuracy Type	Prediction Accuracy (%)			
	RF ATC	RF k-mer 2	RF k-mer 3	RF k-mer 4 *
Overall	90.2	97.9	98.9	98.9
Weighted	81.7	92.7	95.3	95.3
Per Variant				
Alpha	90.0	99.6	100.0	100.0
Beta	57.1	78.6	85.7	85.7
Delta	84.6	93.4	95.6	95.6
Omicron	95.1	99.0	100.0	100.0

*Note: k-mer 4 no longer used for other functions due to large memory use.

Table 2. Prediction accuracies on validation data for K-Nearest Neighbours (K-NN) models on various distance measures and features. Overall accuracy is calculated as correct/total identified variants. Weighted accuracy is an average of per variant accuracy. Finally, per variant accuracy is calculated based on correctly identified variants/all variants of that class.

Accuracy Type	Prediction Accuracy (%)		
	K-NN ATC	K-NN k-mer 2	K-NN k-mer 3
Overall	80.5	97.4	98.8
Weighted	68.4	92.2	96.9
Per Variant			
Alpha	88.5	99.2	99.6
Beta	57.1	78.6	92.9
Delta	36.3	92.3	95.6
Omicron	91.7	98.5	99.5

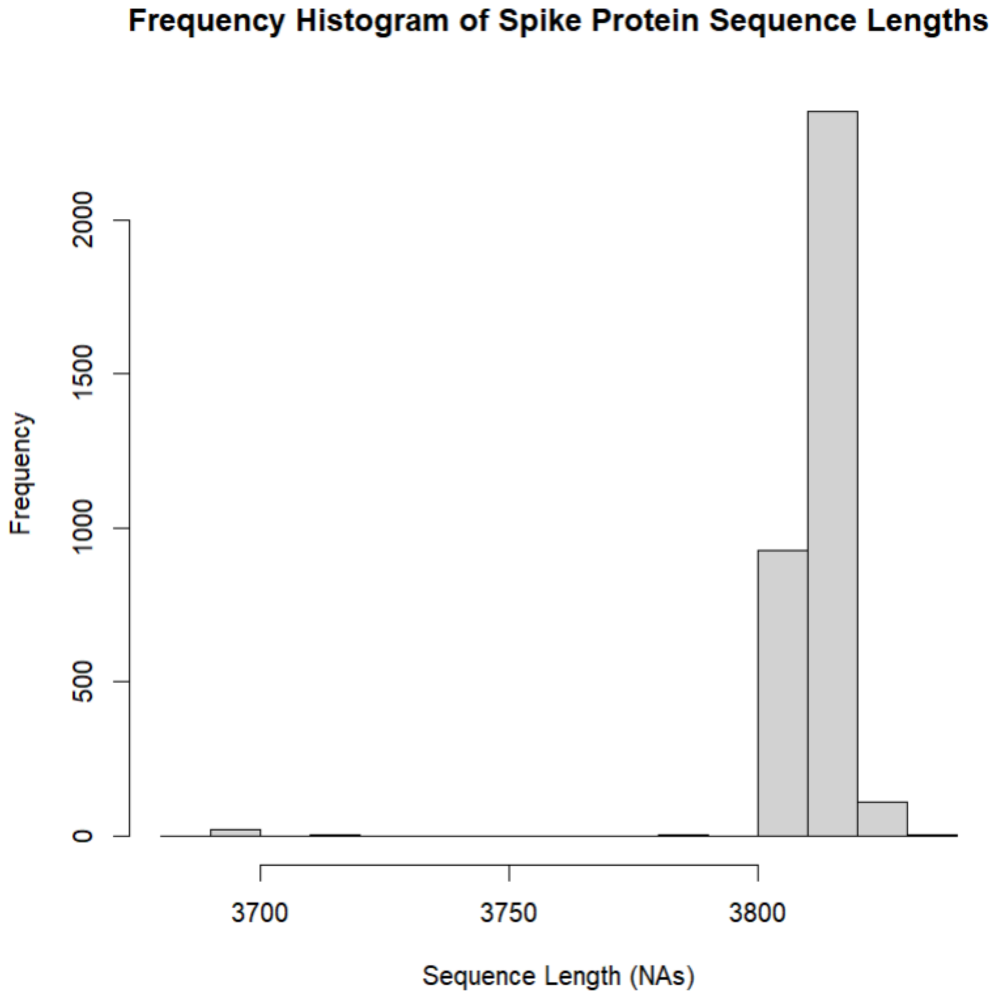
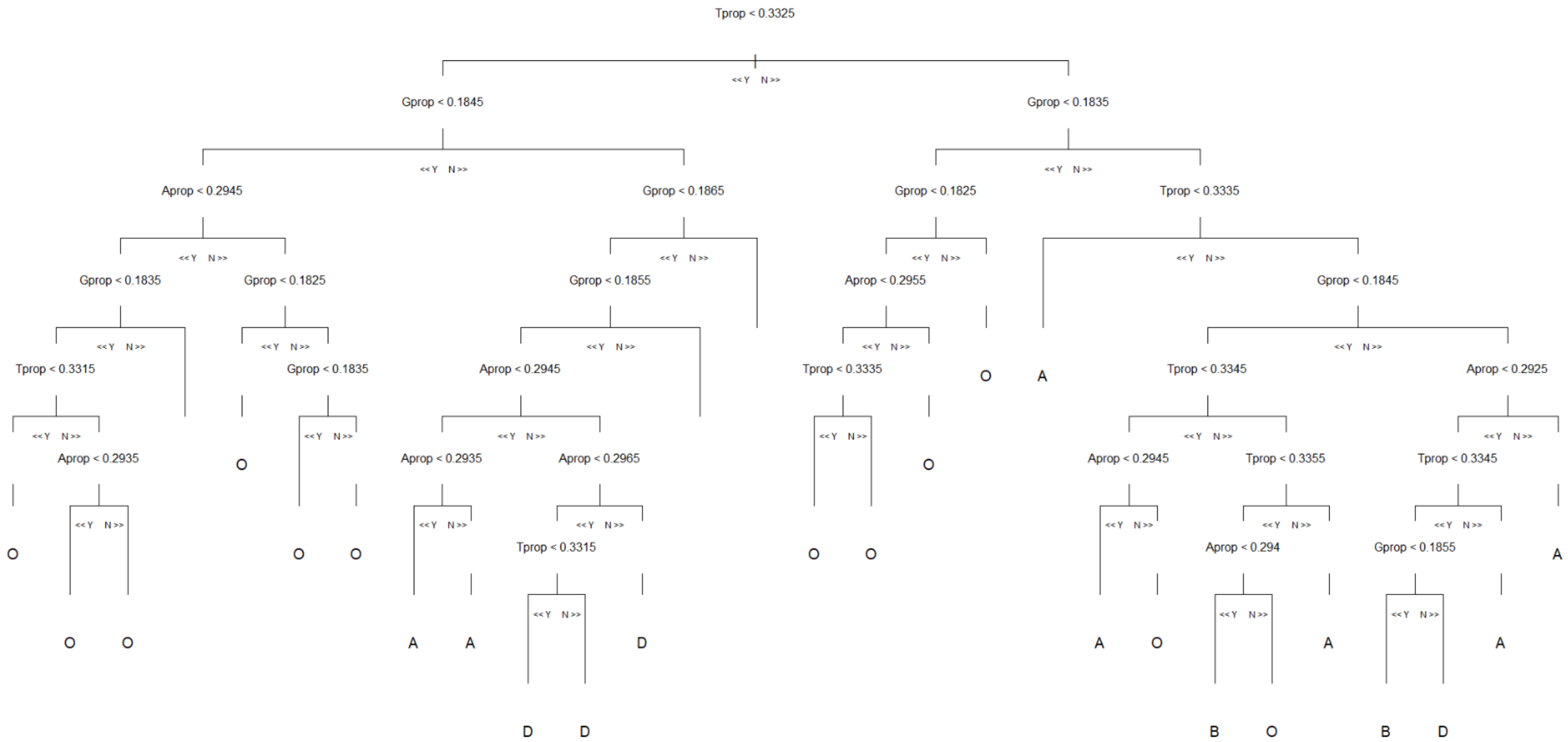


Figure 1. Frequency Distribution of Spike Protein Sequence Lengths. Note that 6 entries from below 3000NA have been removed as outliers.

Figure 2: Example tree featuring feature discrimination for ATG proportion model. Please view other trees in R with the zoom function, they are too large to display here.

Representative Tree for AGT Proportion


Tprop < 0.3325
<<Y N>>
Gprop < 0.1845
Gprop < 0.1835
<<Y N>>
<<Y N>>
Aprop < 0.2945
Gprop < 0.1865
Gprop < 0.1825
Tprop < 0.3335
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
Gprop < 0.1835
Gprop < 0.1825
Gprop < 0.1855
Aprop < 0.2955
Gprop < 0.1845
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
Tprop < 0.3315
Gprop < 0.1835
Aprop < 0.2945
Tprop < 0.3335
O
A
Tprop < 0.3345
Aprop < 0.2925
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
Aprop < 0.2935
O
Aprop < 0.2935
Aprop < 0.2965
O
Aprop < 0.2945
Tprop < 0.3355
Tprop < 0.3345
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
<<Y N>>
O
O
O
Tprop < 0.3315
O
O
Aprop < 0.294
Gprop < 0.1855
A
<<Y N>>
<<Y N>>
<<Y N>>
O
O
A
A
D
A
O
A
A
D
D
B
O
B
D