Predicting Future Mutations of COVID-19 Using Machine Learning

Torpey, Keith (School: East Airport International School)

The pandemic brought about by Covid-19 has greatly disrupted the world. The virus has different mutations that can be seen in the population, but there are also many mutations that exist within a person, not seen in the population. These intra-host variations of the virus can be quantified using variant allele frequency (VAF) data. With machine learning processes, we tested the hypothesis that VAF data can be used to predict future mutations of the virus. We found the difference in allele frequencies between the first and second waves of Covid-19 for each base position. We set thresholds for the difference in frequencies that could be considered as a mutation. We then binarized our targets using these thresholds and determined whether a base position mutates or not. We then used PyTorch and Google Colab to train a machine learning system that took the VAF data as an input and matched it to the targets in an attempt to predict whether a base position mutates or not using VAFs. The system was tested against a portion of the dataset that was held out, and the results were analyzed. The results obtained showed that the model could not efficiently predict the positive cases where a base position mutates. Across different thresholds the results for the accuracy, precision and recall varied, ranging from 97.5%, 42.9% and 52.8% to 51.6%, 1.9% and 2.5% respectively. The study did not put forward conclusive evidence to the research hypothesis that VAF data could predict future mutations. Further research could be done to explore the hypothesis with changes made to the VAFs to improve the data.