Jonathan Lyda

Dr. Paten

BME-230

March 16, 2023

Predicting Virus Evolution of COVID-19 Through Historical Tracing of Variants

I. Abstract

The present study discusses the use of phylogenetic analysis to predict the emergence of disease variants in a population, using SARS-CoV-2 as an example. This involves analyzing historical data to find the emergence time for each variant. A more accurate method averaged the difference in emergence times of each predecessor variant. This method resulted in an emergence rate of roughly 142 days for Covid-19. However, the study has limitations, such as not accounting for every variant's exact phylogenetic common ancestor and not considering other factors influencing disease spread. While useful for tracking the emergence of new variants, there is a need for additional data and modeling techniques to create a reliable prediction. It is important to recognize and apply analytical methods such as phylogenetic analysis to the tracing of disease evolution patterns due to the valuable insights into pandemic preparation and management it can provide.

II. Introduction

An approach to tackle the challenges posed by disease evolution is to predict how the disease develops. This is possible by studying how a disease develops, emerges, and dies out in a population. There are several analytical methods available to anticipate disease evolution and how a virus can mutate within a community. In the wake of the Covid-19 pandemic, the need for

understanding how disease spreads throughout populations has hit a high. To help solve this challenge, implementing analytical techniques such as phylogenetic analysis, genetic analysis, and epidemiological data analysis seem to be the front-running options. Of these techniques, phylogenetic analysis is the most common and is the most promising in terms of prediction quality [3]. By understanding and using these techniques, it is feasible to trace patterns of disease evolution and gain insights into ways to prepare for, prevent, and manage pandemics. Among the prediction capabilities of phylogenetic analysis, there is the possibility to predict when disease variants will emerge in the population.

III. Methods

The overall method of phylogenetic analysis is based on using the information contained in historical data. For the purposes of this experiment, an unrestricted public SARS-CoV-2 metadata file was used. Using a different metadata file for a different disease will yield different results in relation to that specific disease. The Covid-19 metadata file included a name for each positive Covid-19 sample up until March 6, 2023, the date it was sampled, and the variant the sample is believed to belong to [1]. With this information, it is possible to find out the first recorded sample of each Covid-19 variant and ultimately chart the time it took each variant to emerge in the population. The answers to how Covid-19 evolves and spreads are ingrained in the data, making this historical data approach a

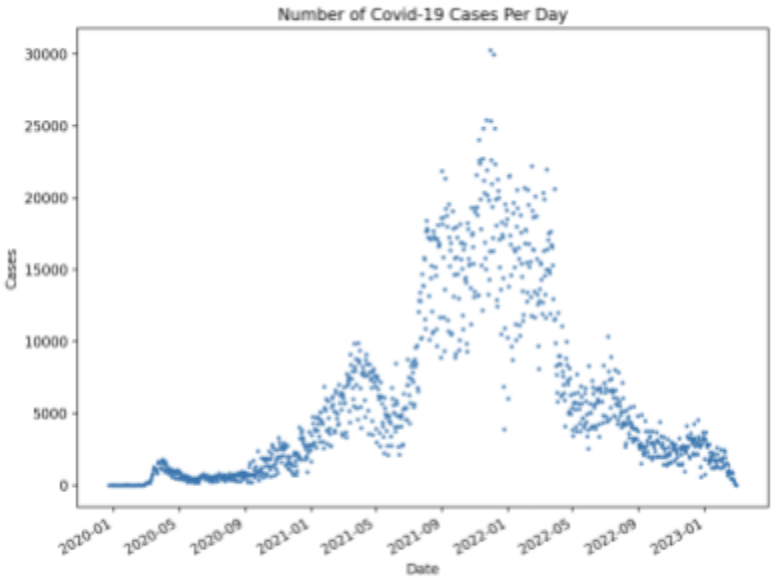

***Fig. 1**: Plot of the number of positive Covid-19 cases per day up until March 6, 2023. This plot illustrates the spike in cases and the growing and diminishing transmission rates of Covid-19 throughout the years from 2020 to 2023.*

strong method of prediction. The Wuhan/Hu-1 sample is the first recorded sample of Covid-19 and also serves as the root of the phylogenetic tree present in the metadata, as it is the common ancestor of all SARS-CoV-2 samples.

Once a time of emergence for each variant has been found, a need for a slightly better estimation of emergence time must be used. A big portion of the reasoning behind this is because there is an unknown variable of when to start counting how many days a variant takes to emerge, in addition to the idea that not all variants are a direct descendant of the first Covid-19 sample (Wuhan/Hu-1). Some variants come from other variants and executing this more accurate emergence rate method roughly takes that into account. It will simply look at all variants before the current variant is calculated and find the difference in emergence times for each predecessor variant. It then averages the results and updates the emergence time of the specific variant to a slightly more accurate number of days. After each variant is calculated with this better estimation, all variant emergence times are averaged to create a rough amount of time it takes a SARS-CoV-2 variant to emerge in the population.

IV. Results

Based on the March 6, 2023 metadata of Covid-19, the estimation came out to be around 142 days for a Covid-19 variant to emerge in the population. This is indeed a better prediction than the previous, less accurate method which had a variant emergence rate of 206 days. The emergence rate of 142 days provides a more

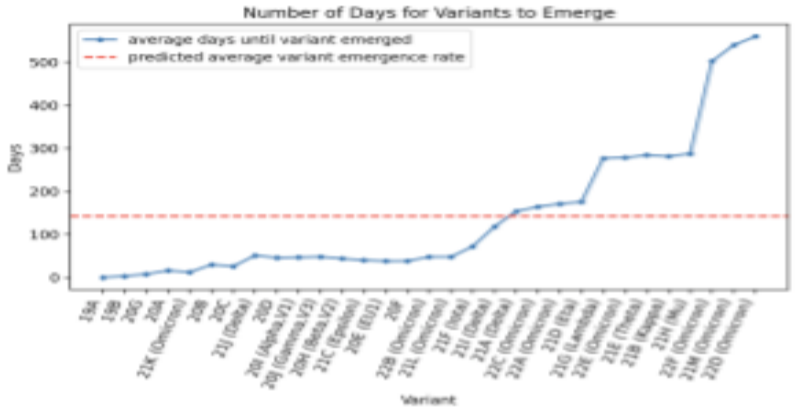

Fig 2: Plot of the number of days it took each Covid-19 variant to emerge in the population. Also has the predicted average variant emergence rate plotted for comparison.

accurate representation of the phylogenetic relationships among the Covid-19 samples, as it adjusts the emergence rate based on the timing and the ancestral variants of a specific variant. With the other estimation, the emergence rate of 206 days assumes that each variant is derived from the original Covid-19 sample (Wuhan/Hu-1) and may oversimplify the relationship between the variants.

V. Discussion

This study was motivated by the catastrophic damage Covid-19 has caused throughout the world. Despite SARS-CoV-2 being the main focus of the study, the logic of this experiment can be translated to any other virus or disease. The sought out goal of this experiment was to figure out the time it takes for an infectious disease to mutate enough to gain a variant. The idea behind this is to give researchers an approximate window of time until they know a disease will mutate. For example, if a highly transmissible virus (such as Covid-19) were to have a breakout somewhere in the world, knowing the amount of time until a variant emerges in the population would be valuable insight. Knowing this window of time could give scientists a deadline for coming up with a vaccine, shelter-in-place orders, or research goals.

A major issue and error found with the logic of this experiment was caused by not completely implementing the idea that every variant is related through the phylogeny of Covid-19. In other words, measuring the correct variant emergence time would require the knowledge of what exact common ancestor, or other variant, each variant stems from. The current implementation doesn't measure this exact number but rather finds the difference in time between all of the possible variants the specific variant could have come from, then averages these numbers to get the emergence time for the specific variant. A better and more accurate version would essentially

traverse down the phylogenetic tree and keep track of how long it took each variant to emerge based on when it split from its common ancestor. Even with this more accurate change, predicting how Covid-19 variants will spread and emerge in the population is a sophisticated process that would require the knowledge of more factors. These factors include the virus's mutation rate, transmission rate, selective pressure, and the population's immune status [2]. Therefore, while monitoring the emergence of new variants and tracking their spread is important, additional data and modeling techniques would be required to create an accurate and dependable prediction of a new variant's emergence in the population.

For the future of this study, the implementation can be worked to be a more accurate estimator of variant emergence times for any disease given to the python script. With more time, a phylogenetic tree data structure could be created based on the inputted metadata file. By using this phylogeny, the emergence time of each variant in the population can be more precisely determined as we would have information on the duration it took for each variant to evolve from its common ancestor. This is what the original study aimed to achieve in a more naive implementation.

VI. Bibliography

[1] McBroome et al. (2021). A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. https://academic.oup.com/mbe/article/38/12/5819/6361626

[2] Miller, James Kyle et al. "Forecasting emergence of COVID-19 variants of concern." PloS one vol. 17,2 e0264198. 24 Feb. 2022, doi:10.1371/journal.pone.0264198

[3] Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. Nature. 2008 May 29;453(7195):615-9. doi: 10.1038/nature06945. Epub 2008 Apr 16. PMID: 18418375; PMCID: PMC2441973.