

Using Phylogenetic trees and substitutions to Capture Relationship of SARS-Cov-2 Variants and SARS-like Genomes

Names: Liz Wyman Z1884762
Kleo Bano Z1940978
Chris Troyer Z1945059
Roberto Rivas Z1906735

Emails: Z1884762@students.niu.edu
Z1940978@students.niu.edu
Z1945059@students.niu.edu
Z1906735@students.niu.edu

Main Question: How do variants of SARS-COV-2 (Omicron, Alpha, etc.) and other SARS-like genomes compare to each other?

Background & Significance:

The COVID19 pandemic created a newfound necessity for research in SARS-COV-2 viruses and variants. Since the start of the pandemic more and more variants have been discovered and sequenced. We will use different algorithms to create multiple phylogenetic trees. This will illustrate the different evolutionary relationships between variants and SARS-like genomes. Using multiple algorithms will allow us to create a clear picture of the evolutionary relationship and compare algorithm outputs. Most people do not understand the difference between any of the SARS genomes enough to care about getting an updated shot, so this should help people understand the differences between them to get them thinking about updated covid shots.

Data:

We will use SARS-Cov-2, variants, and SARS-like genomes that can be downloaded from the NCBI website (data files are prepared by Dr. Hou). We will also use the multi-alignment and pairwise alignment data files provided. Each genome has around 30K DNA bases. We will first conduct all vs. all genome comparison and obtain the similarity measurement between every pair of genomes. The similarity is a numeric value between 0 and 1.

The study:

A comparison of phylogenetic trees will be used to show differences and similarities between SARS-COV-2, variants, and SARS-like genomes. This will help show why the different genomes and given different names and why SARS-COV-2 variants are variants instead of a different SARS-like genome. We will calculate branch lengths for each algorithm and compare trees. We will map indels, substitutions, and log gap rates to compare how those values change or modify the output on a specific phylogenetic tree. In the case of low substitution rate, we expect the branches to be close together. In the case of high substitution rates, we expect the branches to be far apart from one another. We will first start with the SARS-COV-2 and variants, then move onto SARS-like genomes. Graphs will be constructed and provided as a visual aid to view the data that is found and collected.

Programming language & system: Java, Linux, Python

Library packages: BioJava, Biopython, Matplotlib

**Additions: Python and matplotlib were used to create Phylogenetic trees to compare genomes

Work environment: GitHub will be used for collaboration on all files and research.

Contribution: Liz – report writing, graph designing 25 %

Kleo - design of the program, implementation/debugging/documentation 25 %

Chris - design of the program, implementation/debugging/documentation 25 %

Roberto - report writing, graph designing 25 %

Initial data (same from proposal):

PW FILE	SCORE	MATCHES	SUBSTITUTIONS	INDELS(GAPS)
SARS2.ALPHA.SING.MAF	2,811,044	29,808	31	10
SARS2.BETA.SING.MAF	2,842,956	29,819	26	18
SARS2.DELTA.SING.MAF	2,839,244	29,761	33	0
SARS2.GAMMA.SING.MAF	2,836,621	29,767	35	13
SARS2.OMICRON.SING.MAF	2,837,255	29,828	39	36
SARS2.SARS.SING.MAF	1,589,514	23,749	5,860	369

Distance matrix based on score

	alpha	beta	delta	gamma	omicron	sars
<i>alpha</i>	0	31,912	28,200	25,577	2,611	1,221,530
<i>beta</i>	31,912	0	3,712	6,335	5,701	1,253,442
<i>delta</i>	28,200	3,712	0	2,623	1,989	1,249,730
<i>gamma</i>	25,577	6,335	2,623	0	634	1,247,107
<i>omicron</i>	26,211	5,701	1,989	634	0	1,247,741
<i>sars</i>	1,221,530	1,253,442	1,249,730	1,247,107	1,247,741	0

Distance matrix based on matches to sars 2

	alpha	beta	delta	gamma	omicron	sars
<i>alpha</i>	0	11	47	41	20	6,059
<i>beta</i>	11	0	58	52	9	6,070
<i>delta</i>	47	58	0	6	67	6,012
<i>gamma</i>	41	52	6	0	61	6,018
<i>omicron</i>	20	9	67	61	0	6,079
<i>sars</i>	6,059	6,070	6,012	6,018	6,079	0

Distance matrix using substitution rate

	alpha	beta	delta	gamma	omicron	sars
<i>alpha</i>	0	1.68×10^{-4}	6.9×10^{-5}	1.36×10^{-4}	2.67×10^{-4}	1.97×10^{-1}
<i>beta</i>	1.68×10^{-4}	0	2.36×10^{-4}	3.03×10^{-4}	4.35×10^{-4}	1.97×10^{-1}
<i>delta</i>	6.9×10^{-5}	2.36×10^{-4}	0	6.7×10^{-5}	1.98×10^{-4}	1.97×10^{-1}
<i>gamma</i>	1.36×10^{-4}	3.03×10^{-4}	6.7×10^{-5}	0	1.31×10^{-4}	1.97×10^{-1}
<i>omicron</i>	2.67×10^{-4}	4.35×10^{-4}	1.98×10^{-4}	1.31×10^{-4}	0	1.97×10^{-1}
<i>sars</i>	1.97×10^{-1}	1.97×10^{-1}	1.97×10^{-1}	1.97×10^{-1}	1.97×10^{-1}	0

Distance matrix using gap rates

	alpha	beta	delta	gamma	omicron	sars
<i>alpha</i>	0	2.68×10^{-4}	3.35×10^{-4}	1.01×10^{-4}	8.69×10^{-4}	1.20×10^{-2}
<i>beta</i>	2.68×10^{-4}	0	6.03×10^{-4}	1.67×10^{-4}	6.01×10^{-4}	1.17×10^{-2}
<i>delta</i>	3.35×10^{-4}	6.03×10^{-4}	0	4.36×10^{-4}	1.20×10^{-3}	1.23×10^{-2}
<i>gamma</i>	1.01×10^{-4}	1.67×10^{-4}	4.36×10^{-4}	0	7.68×10^{-4}	1.19×10^{-2}
<i>omicron</i>	8.69×10^{-4}	6.01×10^{-4}	1.20×10^{-3}	7.68×10^{-4}	0	1.11×10^{-2}
<i>sars</i>	1.20×10^{-2}	1.17×10^{-2}	1.23×10^{-2}	1.19×10^{-2}	1.11×10^{-2}	0

Distance matrix:

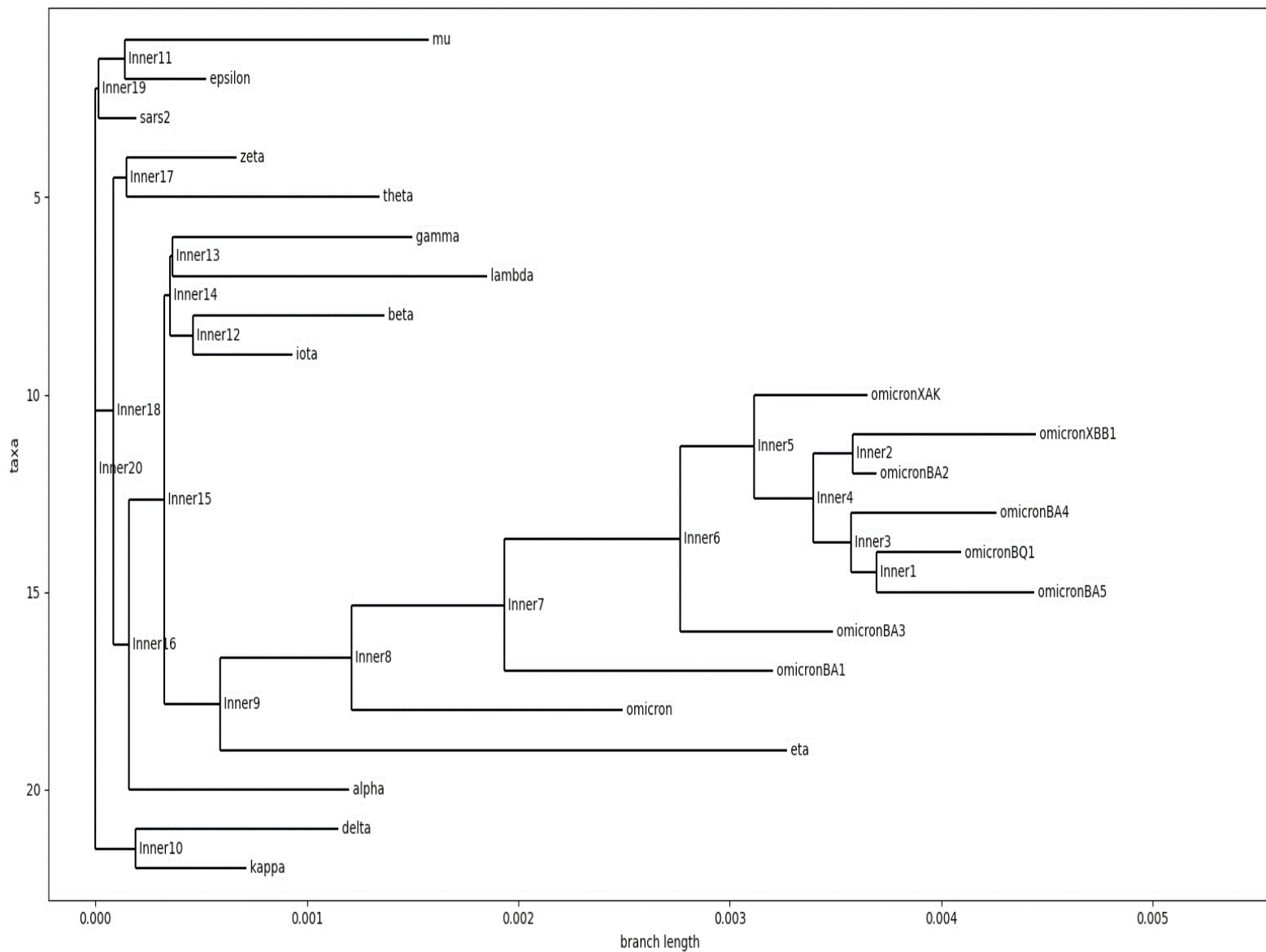
There are 4 different distance matrices above that show how closely related SARS-COV-2 variants and sars are to one another. Comparing the data from all different matrices shows the substantial differences between SARS-COV-2 variants and sars. The variants differences to each other vary within a smaller range of numbers, suggesting some variants are more closely related than others. This is useful in determining precautions to use against each variant.

Substitution rate:

Comparing the wildtype SARS-COV-2 genome to SARS-COV-2 variants and SARS-like genome sars, shows how the substitution rate is substantially greater for sars than any of the variants. Also, the substitution rates comparing sars to variants are within a ~ 0.002 margin to the substitution rate of sars to SARS-COV-2 of 0.194913. This shows how closely related the variants are compared to one another and the wildtype SARS-COV-2 genome, which is why they are not considered variants of SARS-like genome sars.

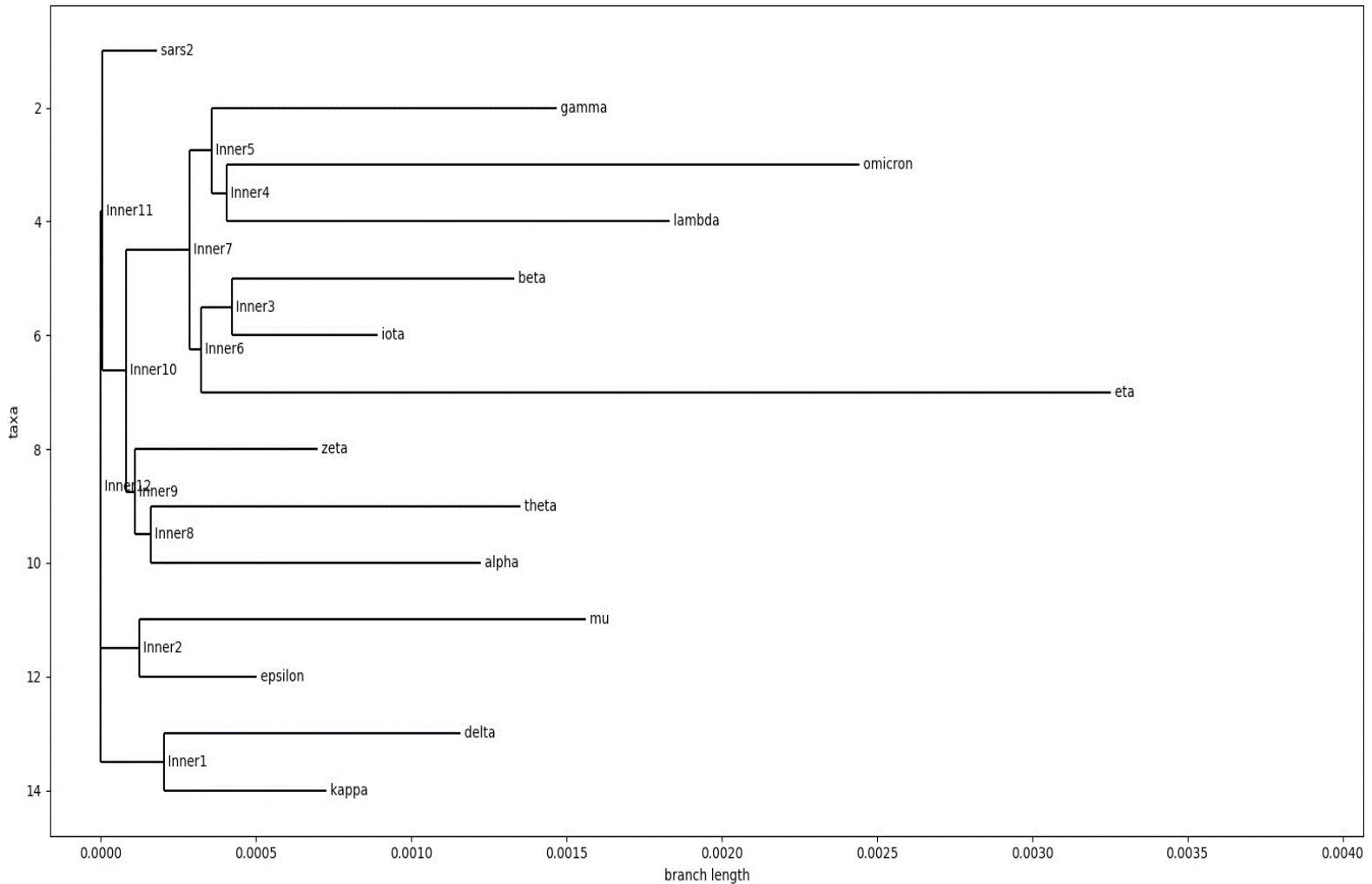
Phylogenetic tree:

As you can see in the tree below, the omicron variants have longer branch lengths and are clustered together. This shows how closely related the omicron variants are, while showing how they differ from other SARS-COV-2 variants. The only SARS-COV-2 variant that has a relatively similar branch length is eta. Although, more recent omicron variants have longer branches. A branch indicates evolutionary change from the root sars2 and as the branch length increases it shows the amount of evolutionary change also increases. This also suggests how a covid shot that targeted the alpha variant would not be as effective at targeting omicron variants because of how much they changed. More and more variants keep evolving, so it is important to keep an eye on their level of relation to one another to better understand how to take precautionary measures against them.



SARS-COV-2 variant tree:

Looking at the tree below that removes omicron variants from the equation can give a better look at how the SARS-COV-2 variants relate to one another. Epsilon shows the least amount of evolutionary change and eta shows the greatest when comparing only SARS-COV-2 variants. This is determined by considering the branch length. The longer the branch length the more distant of an evolutionary relationship.



Current plan:

The current plan is to continue to build trees that examine the evolutionary relationship between variants. We will use biophyton instead of biojava as the latter did not yield the simplest results or code. This includes cleaning up our current trees and extending the number of sequences for each tree using SARS-like genomes. Finally we will analyze the data based on these phylogenetic trees.