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

Library packages: BioJava

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

Contribution: Liz – report writing 25 %
Kleo - design of the program, implementation/debugging/documentation 25 %
Chris - design of the program, implementation/debugging/documentation 25 %
Roberto - report writing 25 %

Timeline:
1-2 weeks – Code phylogenetic trees and branch lengths
1-2 weeks after coding complete – Analyze data

Initial data:

pw file	score	matches	substitutions	indels (gaps)
sars2.alpha.sing.maf	2811044	29808	31	10
sars2.beta.sing.maf	2842956	29819	26	18
sars2.delta.sing.maf	2839244	29761	33	0
sars2.gamma.sing.maf	2836621	29767	35	13
sars2.omicron.sing.maf	2837255	29828	39	36
sars2.sars.sing.maf	1589514	23749	5860	369

distance matrix based on score

	alpha	beta	delta	gamma	omicron	sars
alpha	0.0	31912.0	28200.0	25577.0	26211.0	1221530.0
beta	31912.0	0.0	3712.0	6335.0	5701.0	1253442.0
delta	28200.0	3712.0	0.0	2623.0	1989.0	1249730.0
gamma	25577.0	6335.0	2623.0	0.0	634.0	1247107.0
omicron	26211.0	5701.0	1989.0	634.0	0.0	1247741.0
sars	1221530.0	1253442.0	1249730.0	1247107.0	1247741.0	0.0

```
distance matrix based on matches to sars 2
          alpha    beta   delta   gamma  omicron    sars
alpha       0.0    11.0    47.0    41.0     20.0  6059.0
beta       11.0     0.0    58.0    52.0      9.0  6070.0
delta      47.0    58.0     0.0     6.0     67.0  6012.0
gamma      41.0    52.0     6.0     0.0     61.0  6018.0
omicron    20.0     9.0    67.0    61.0      0.0  6079.0
sars     6059.0  6070.0  6012.0  6018.0   6079.0     0.0

distance matrix using substitution rate
            alpha      beta     delta     gamma   omicron      sars
alpha    0.000000  0.000168  0.000069  0.000136  0.000267  0.196874
beta     0.000168  0.000000  0.000236  0.000303  0.000435  0.197042
delta    0.000069  0.000236  0.000000  0.000067  0.000198  0.196805
gamma    0.000136  0.000303  0.000067  0.000000  0.000131  0.196738
omicron  0.000267  0.000435  0.000198  0.000131  0.000000  0.196607
sars     0.196874  0.197042  0.196805  0.196738  0.196607  0.000000

distance matrix using gap rates
            alpha      beta     delta     gamma   omicron      sars
alpha    0.000000  0.000268  0.000335  0.000101  0.000869  0.011974
beta     0.000268  0.000000  0.000603  0.000167  0.000601  0.011706
delta    0.000335  0.000603  0.000000  0.000436  0.001204  0.012309
gamma    0.000101  0.000167  0.000436  0.000000  0.000768  0.011873
omicron  0.000869  0.000601  0.001204  0.000768  0.000000  0.011105
sars     0.011974  0.011706  0.012309  0.011873  0.011105  0.000000
```

Initial analysis:
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.

Looking at the gap rate between SARS-COV-2 and Sars sequences we can see there is a much higher number of indels between these two sequences as compared to the SARS-COV-2 variants. In our phylogenetic trees we can predict that the SARS sequence will be the outgroup for creating a harmonious picture.

The distance matrices will be used to make the phylogenetic trees. These different matrices will give a wide picture of the types of phylogenetic tress we will create.