Unraveling the Genomic Mysteries of COVID-19: Using the Power of Phylogeny to Predict the Variants Before They Come into Existence

Sana Abbas

St. Andrews Episcopal Upper School, USA

ABSTRACT

There is a notable gap in the taxonomic research of COVID-19. This investigation aimed to address and fill this gap by evaluating the accuracy of different phylogenetic tree methods in displaying the evolutionary history of the virus SARS - CoV 2. This research was motivated by the potential to prevent the recurrent spread of the disease world wide and making the vaccine creation process easier by gaining the ability to predict future virus mutations through the use of phylogenetic analysis. This investigation determined which phylogenetic mechanism (Parsimony, Maximum Likelihood, Neighbor-Joining, and UPGMA) was most accurate in portraying the evolutionary history of COVID-19. The hypothesis that was inferred before beginning this investigation was that the likelihood method of phylogeny would display the evolutionary history of COVID-19 most accurately due its reputation as a phylogenetic method that conducts more thorough studies than other statistical approaches. This investigation was carried out by equally sampling COVID-19 DNA from around the globe and inputting them into software that produces different types of phylogenetic trees. The bootstrap values of the clades of each of the four types of trees produced were then observed to reach a conclusion as to which of them displayed COVID-19 most accurately. The hypothesis ended up being unsupported as it turned out that of the 4 methods tested, the neighbor-joining method proved to be the most accurate. This discovery had significant implications for future research and could be used by scientists to possibly predict new COVID-19 variants before they come into existence.

The Rationale Behind This Study

There is an extreme lack of research on COVID-19. This is especially the case when it comes to the virus taxonomy (the science of naming, classifying, and describing organisms) field. But what if this problem could be fixed? And along with that, breakthrough discoveries could be made such as uncovering undocumented COVID-19 infection sources and most importantly, predicting future mutations before they actually come into being. Such breakthroughs may enable the potential prevention of the recurrent spread of the disease worldwide. All of this can be achieved through phylogenetic analysis. This project aims to answer the question: which type of phylogenetic mechanism is most accurate in displaying the evolutionary history of COVID-19?

Independent Variables

The independent variables are the methods of phylogenetic analysis that were chosen for this study: Parsimony, maximum likelihood, neighbor-joining, and UPGMA. Phylogenetic trees are a fundamental tool in organizing our knowledge of the evolutionary history of a biological species, a gene, or any organism through one common ancestor. They are based on differences either physically or genetically. Phylogenetic trees, however, are not definitive facts, they are merely hypotheses that are made in an attempt to trace down evolutionary history since the beginning of time. Now the question arises if the concept of phylogeny can be applied to viruses since viruses do not possess one common gene that is shared by all viruses. They also do not share any characteristics with cells. The answer is yes. The concept of phylogeny can be applied to viruses. Phylogenetic analysis of viruses is used in fundamental virus research including epidemiology, virus taxonomy, diagnostics, and much more. It provides an evolutionary perspective on variations of any trait for a group of viruses that can be measured. Each of the methods that were selected for this investigation was chosen for a specific reason.

How to Read a Phylogenetic Tree

There are many parts to a phylogenetic tree. Each part helps to demonstrate the evolutionary history of the organism it is trying to portray. The first part of a phylogenetic tree towards the bottom is called the root. The root node represents the most recent common ancestor. The next part is the vertical lines otherwise known as branches. Branches represent a lineage. Other parts of the tree include nodes, which are where the lineages (branches) diverge. The lineages diverge to represent speciation events that occurred from one common ancestor. Trees sometimes also want to show synapomorphies. Synapomorphies are common characteristics that are shared by a group of taxa. The characteristics don't necessarily have to be external features, it can also include a specific sequence of DNA. Synapomorphies are usually always labeled on phylogenetic trees. Time is also represented on a phylogenetic tree. It goes from the oldest which is located at the bottom, to the most recent which is located at the top.

Interpreting phylogenetic trees can be highly complex and challenging. It is difficult to understand the formatting of a phylogenetic tree as there are many different things to consider. There are three ways a phylogenetic tree can be depicted: cladogram, phylogram and radial. A cladogram represents the relationships between different species or groups. The lengths of the branches do not represent the amount of evolutionary change or time. A phylogram depicts both the relationships between species and the amount of change or time that has occurred. In phylograms, branch lengths are proportional to evolutionary distance or time. A radial phylogenetic tree represents the exact same information as a phylogram but arranged in a circular layout. This depiction is often used for visual clarity, especially when dealing with many species. The most important thing to remember is that all three depictions are saying exactly the same thing (McLennan, 2010). Another common mistake that people make is assuming that the order of the tips of trees has meaning. The thing that matters most when reading and interpreting a phylogenetic tree is the order of branching along the time axis. One can rotate branches and not affect the structure of the tree (Taylor, 2015).

Maximum Parsimony

Parsimony was chosen as a method because it is known as the most common and simplest phylogenetic tree-based method. The method of parsimony involves grouping taxa together in ways that minimize the amount of evolutionary change. First, the MP algorithm starts considering a tree with a particular topology. Then, it infers the minimum number of character sequences required to explain all the nodes of the tree at every sequence position. After all of the tree topologies are analyzed, the tree with the fewest number of changes is chosen. “Parsimony analysis is very fast and has been demonstrated to be quite effective in many situations” (Lemey, 2009). Although parsimony does well in recovering the true evolutionary tree, there have been instances where it has failed very badly. It will however always remain part of phylogenetic analysis.

Maximum Likelihood

Maximum likelihood was chosen because it uses probabilities to infer the likelihood of the tree structure. Although it is a lengthy process, it is known to be very accurate and precise. This method is similar to maximum parsimony in a way because it also examines every reasonable tree topology along with finding the support for each by examining every DNA sequence position. But the difference is that it calculates the probability of expecting each possible nucleotide in the ancestral nodes. The likelihood of the tree structure is then inferred using these probabilities. The likelihood of all tree topologies is searched in this same manner and the most likely tree is then chosen as the best tree. The process of constructing a likelihood tree is more lengthy than some other methods. This is due to the fact that likelihood trees have to be tested prior to being constructed. “The actual process is complex, especially because different tree topologies require different mathematical treatments, so it is computationally demanding” (Lemey, 2009).

Neighbor-Joining

The neighbor-joining method constructs a phylogenetic tree by sequentially finding pairs of “neighbors,” or the pairs of operational taxonomic units (OTU’s) that are connected by a single interior node (Lemey, 2009). Neighbor-joining was chosen as a tree-based method because it is a distance matrix method. Neighbor-joining is one of two of the distance-matrix methods. What this means is that it takes a matrix of pairwise evolutionary distances between the DNA sequences and uses that data to build the tree. The pairwise distances are obtained from DNA sequence alignment algorithms. The neighbor-joining method is the most widely used distance-based method in phylogenetic analysis.

UPGMA

UPGMA (Unweighted Pair-Group Method with Arithmetic Mean) is a type of clustering method and is the second most commonly used distance matrix method. It is used commonly in microbial epidemiological studies. First, the program locates the pair of taxa with the smallest amount of distance between them. Then it takes this pair and defines the branching that’s present between them as half of that distance, which in turn places a node at the midpoint of the branch. Then it combines the two taxa into a cluster and rewrites the matrix with the distance from the cluster to each of the remaining taxa. This process is then repeated and reiterated until the matrix consists of one single entry. A UPGMA tree is then built using that set of matrices. The main disadvantage of using the UPGMA method is that it assumes that the evolutionary rate is the same in all of the branches. This means that all of the mutations of the organism happened at the exact same rate, which is entirely unrealistic. As a result of this assumption, the UPGMA method tends to produce inaccurate trees when evolutionary rates vary.

Bayesian Analysis

The last and most complex method of phylogenetic analysis that should be considered in this study is Bayesian analysis. Bayesian analysis is a method of statistical inference. It was named after the English mathematician Thoman Bayes. Bayes theorem is a simple mathematical formula used for calculating conditional probabilities. Bayesian inference of phylogeny combines the information in the prior with the data likelihood to create the posterior probability of trees. The posterior probability of trees is the probability that the phylogenetic tree is correct given the data, the prior, and the likelihood model.

Dependent Variable Acquired Through Bootstrapping

The dependent variable is the accuracy of the phylogenetic mechanism to display COVID-19. This can be determined by making connections between which countries are linked by air travel and if the COVID-19 distribution reflects that. It also can be done by checking for group similarities. However, there is a more accurate way of determining the accuracy of a phylogenetic method. The Bootstrap Test is a common method of analysis that assesses the quality of tree branches by evaluating the statistical support that they have. It does this by creating bootstrap replicates from the genetic data. "A bootstrap replicate is a shuffled representation of the DNA sequence data" (Hall, 2011). This method is commonly used whenever the underlying sample distribution is unable to be derived analytically or in some situations is entirely unknown. This is an extremely simple process of measuring internal consistency of molecular data by determining if slightly modified alignments still support the same clades that were generated in the original tree.

The model that was used to determine the results is the Kimura 2 Parameter (K2P) model. Its goal is to measure how different two DNA sequences are from each other. The Kimura 2-parameter was also used in opposition to the one parameter model which is the simplest but unrealistic as it makes the assumption that the rates at which transitions and transversions occur are simultaneous. Instead, the Kimura 2-parameter takes into account the possibility of the rate's occurrence being different by calculating the branch lengths based on two different rates; the rate of transition mutations that occur between two similar base pairs and the rate of transversion mutations that occur between two different base pairs. This model takes into account the fact that transition mutations occur far more frequently than transversions by giving two different probabilities to these mutations and in turn, producing more accurate results.

In summary, the Kimura 2 Parameter model assists in accurately calculating genetic distances while the bootstrap method provides a way to statistically validate the reliability of the phylogenetic trees produced. The K2P model and bootstrapping were used together to help improve the reliability of the results of the phylogenetic analysis.

Table 1. The confidence level that corresponds to each range of bootstrap values.

Bootstrap Values	Confidence
>90%	Strongly Supported
70%-90%	Well Supported
50%-70%	Weakly Supported
<50%	Not Supported

Impact on Science

Determining which type of phylogenetic tree is best suitable for accurately displaying the evolutionary history of COVID-19 will have a tremendous impact on science. It would not only assist phylogeneticists to conduct their investigations but would help make more COVID-19 studies possible (especially in the virus taxonomy field). There already is a lack of COVID research so an increase in phylogenetic evaluation of the virus would lead to even more breakthroughs. Phylogenetic trees provide an evolutionary perspective on variations of any trait for a group of viruses that can be measured. This includes looking at past mutations and predicting what may happen to the virus's genes in the future. The benefits of phylogenetic research are that it enables you to possibly detect undocumented COVID-19 infection sources. Knowing such crucial information could potentially prevent the recurrent spread of the disease worldwide. Additionally, having an accurate way of inferring future mutations could enable scientists to predict the variants of COVID-19 before they come into being.

Hypothesis

The hypothesis for this study is: If the likelihood method of phylogeny is the best way to display COVID-19, then it will display its evolutionary history most accurately. This was inferred due to the fact that likelihood trees are known to produce better, more accurate results than other statistical approaches. The null hypothesis is that these different mechanisms of phylogeny do not have any effect on the accuracy of COVID-19 being displayed.

Prior Research

An investigation was done previously concerning the same topic and with the same intentions. However, there are some key differences between that investigation and this one. First of all, that investigation utilized a smaller set of DNA sequences. While that one used only 200, this one uses 224 making it already stand a higher chance of being more accurate. Secondly, that investigation only tested 3 phylogenetic tree methods: parsimony, likelihood, and neighbor-joining. UPGMA was never taken into consideration. Out of those 3, neighbor-joining was shown to be the best. The last difference is the most important one. In the previous study, the accuracy of the phylogenetic tree methods was determined by creating a list of all similar groupings on each tree. Then, all of the similar groupings were color-coded and made into a singular tree. Using this tree, comparisons were made between the different tested methods. These comparisons included variant identification, clade and group visibility, and overall accuracy. The overall accuracy of the methods was determined by making connections between which countries are linked by air travel and if the COVID-19 distribution reflects that. However, it was primarily done by checking for group similarities. This method was overall not very accurate as it was subject to plenty of human error and did not utilize statistics or any other logical reasoning to reach a conclusion. This study, however, uses bootstrap analysis in order to determine the most accurate out of the 4 methods, therefore, making this one a far more accurate and reliable study.

Methods

Materials

1. MEGA 11.0.10 (64-bit)
2. Tracer (2018)-05-01 - v1.7.1
3. Mesquite Software Version 3.70 (2021)
4. PAUP 4.0a169
5. Genbank 41.D1 (2012): D36-D42
6. 11th Generation Intel® CoreTM i7-1165G7 Processor (12MB Cache, up to 4.7 GHz)
7. macOS Monterey 12.1 (21C52)
8. MAFFT Version 7

Procedure

The procedure that was used to conduct this investigation is as follows. First, COVID-19 samples were equally collected from all around the United States starting from 2020 to 2023. Four were collected from each state and territory (56 in total) on Genbank (Benson, 2012) (1 per year), resulting in 224 sequences in total. Those files that were previously concatenated by their appropriate state/territory, were then merged into one huge data file. This file was then run through a MAFFT (Katoh, 2006) Server to align the DNA. Each of them were aligned in FASTA format enabling the inputting of the data file into the different phylogenetic software. It was then time to create the experimental phylogenetic trees. A phylogenetic tree was first created with the likelihood method. This was done using the software MEGA (Nei, 1993). A phylogenetic tree was then created with the parsimony method using PAUP (Swofford, 2003). Next, a neighbor-joining tree was created using Mesquite (Madison, 2021). And finally, a phylogenetic tree using the UPGMA method was created using MEGA (Nei, 1993). Before running the data file to generate the trees on all of these software, their parameters were all set to utilize bootstrap analysis with 1000 replications as well as with a Kimura 2 parameter model to ensure the most accurate trees possible. These phylogenetic trees were also observed using the viewing software Tracer (Rambaut, 2018). All of the different clades, groups, and variants that the phylogenetic methods contained were identified. Then, the bootstrap values were recorded and analyzed to determine which tree did the best job, by recording the number of clades each tree had with a bootstrap value of over 70%.

A Note on Safety and Ethics

This is an entirely risk free method that utilizes only previously-collected sequences from a database to experiment with. This project is entirely computer-based and does not require precautions before completing it. There is also nothing ethically wrong with performing this study.

Controls

In a phylogenetic study, especially when dealing with genetic or sequence data, the concept of a “control” is not typically applicable in the same way it is in experimental settings. In experimental research, a control group is often used to compare against the experimental group to assess the impact of a treatment or intervention. In a phylogenetic study, the primary focus is on understanding the evolutionary relationships and patterns among organisms or sequences. Instead of a control group, known variants were used. In this study, it seemed that the most logical course of action would be to use the 4 main variants of concern in the United States: Alpha, Beta, Delta, and Gamma. Another good positive control is the consistency with epidemiological data such as expected patterns of virus spread based on location.

Uncertainty

Systematic uncertainty in phylogenetic studies can arise from various sources. The main source of uncertainty in this investigation, however, was during the data collection stage. Besides, using a relatively small data set, the data was collected through means of systematic sampling. The results would have definitely been affected by this. Typically when professionals are performing phylogenetic studies, they have access to far more powerful computers with much more RAM than the general population is able to obtain. Therefore, the data set was reduced in order to accommodate the limits of the computers used. Unfortunately, there is no real way to be able to fix this and a far more accurate study would definitely be able to be performed without these limitations. Furthermore, a greater number of trials would also be conducted as a result of using more advanced and powerful lab computers. Creating these few trees already poses to be difficult for traditional computers making numerous trials out of the question to prevent crashing the computer.

Data Analysis and Discussion

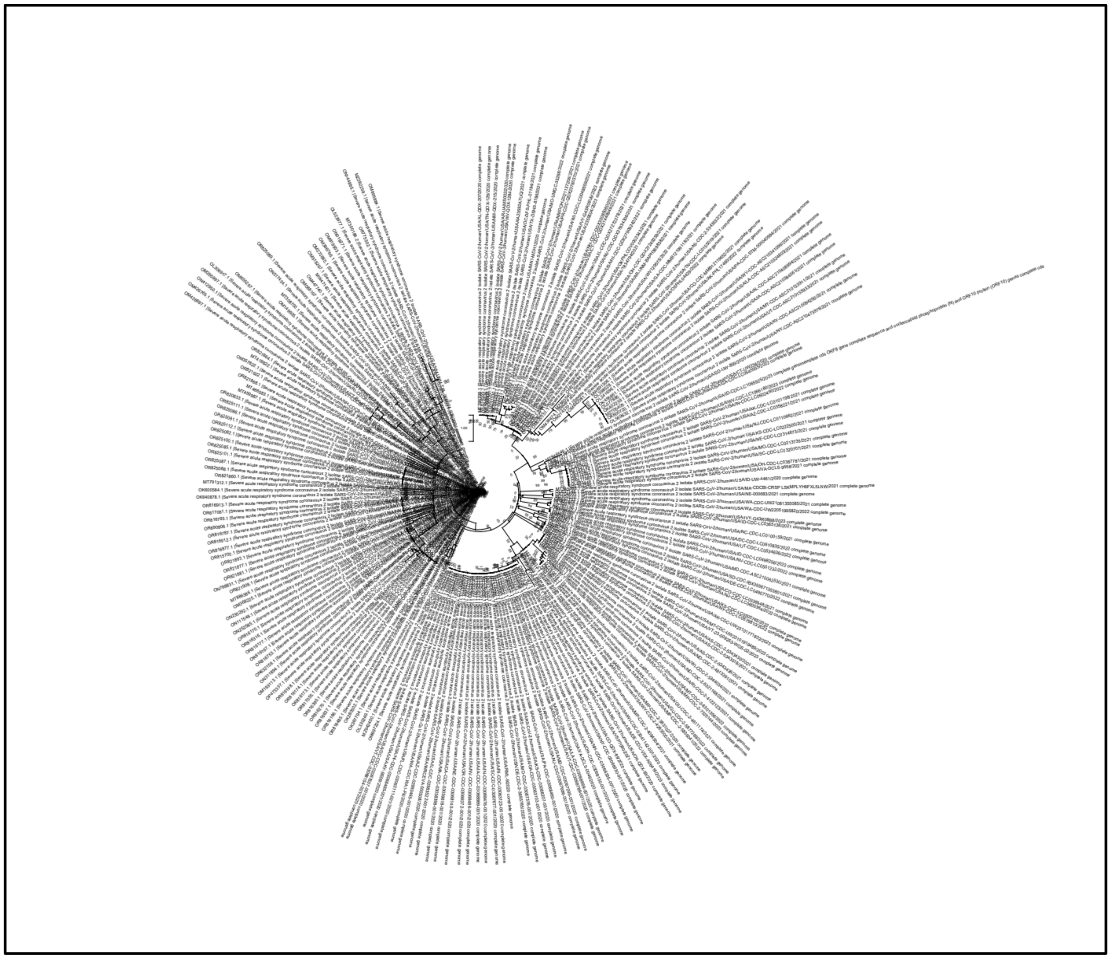

Figure 1. Maximum parsimony tree created using the software PAUP

For the maximum parsimony method, there were a total of 210 clades that were created. Out of those 210, 104 of them were strongly supported meaning that the parsimony method created 104 clades that had a bootstrap value of 90% or above and upon numerous replications, would still be supported. 22 out of the 210 clades were well supported, meaning that they had bootstrap values from 70%-89% making them almost always supported upon numerous replications. 14 of the clades out of 210 were weakly supported, meaning that they had bootstrap values from 69% to 50% making them only sometimes supported upon numerous replications. The remaining 70 clades were all not supported, meaning they all had bootstrap values that fell under 50% making them completely unsupported upon numerous replications.

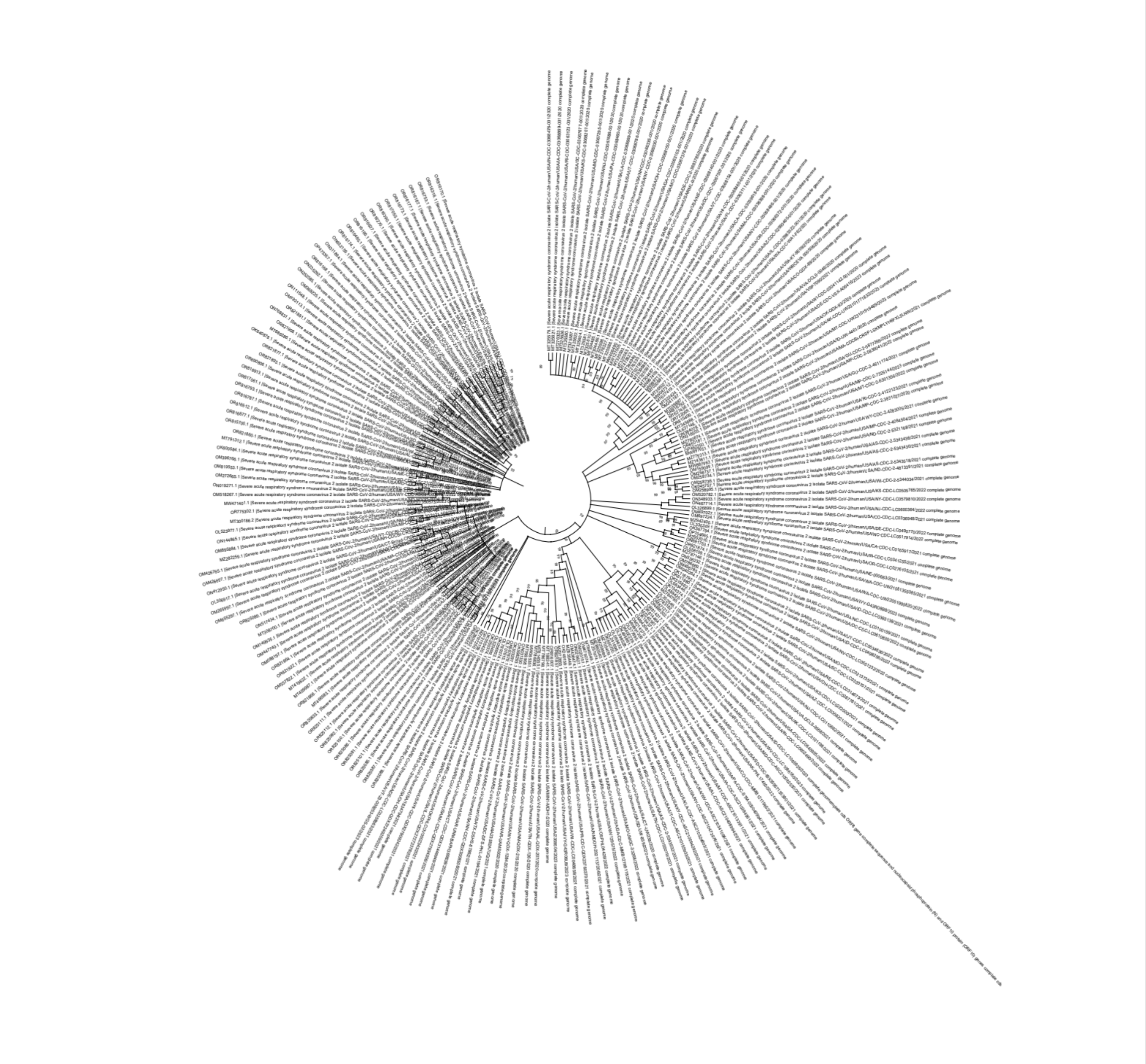

Figure 2. UPGMA tree created using the software MEGA

For the UPGMA method, there were a total 106 clades that were created. Out of those 106, 91 of them were strongly supported meaning that the UPGMA method had constructed 91 clades that all had a bootstrap value of 90% or above and would still be supported after numerous repications. 4 of the clades were well supported with bootstrap values ranging from 70% to 89%, 4 of them were weakly supported with bootstrap values from 69% to 50% making them only sometimes supported upon numerous replications. And lastly, 7 of the clades were not supported upon numerous replications as they all contained bootstrap values that were below 50%.

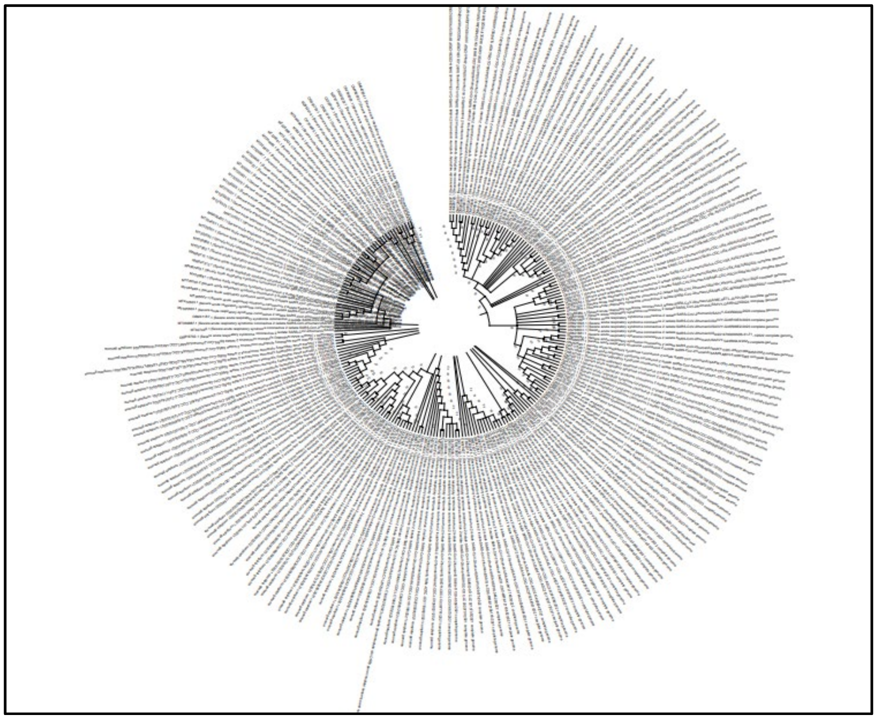

Figure 3. Maximum likelihood tree created using the software MEGA

For the maximum likelihood method, there were a total of 130 clades that were produced. Out of those 130,51 had bootstrap values over 90% and were strongly supported, 35 were well supported, and 44 with bootstrap values ranging from 69% to 50% were weakly supported upon numerous replications. The likelihood method did not produce any clades that were entirely unsupported.

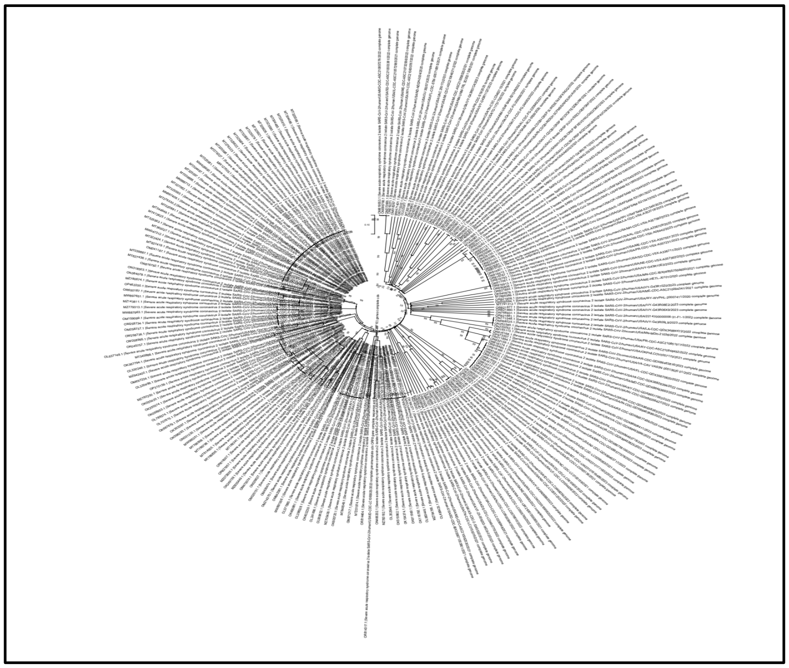

Figure 4. Neighbor-Joining tree created using the software Mesquite

For the neighbor-joining method, there were a total of 98 clades created. 97 of those clades were strongly supported upon numerous replications. And the remaining 1 clade was still well supported with a bootstrap value only falling slightly under at 89%. This method yielded no clades that fell under the bootstrap value of 89% meaning that it produced only extremely accurate clades that would stay consistent no matter what.

Table 2. A summary of all of the clades produced by each type of tree along with the confidence levels

Tree Type	Strongly Supported	Well Supported	Weakly Supported	Not Supported	Total # of Clades
Parsimony	104	22	14	70	210
UPGMA	91	4	4	7	106
Likelihood	51	35	44	0	130
Neighbor-Joining	97	1	0	0	98

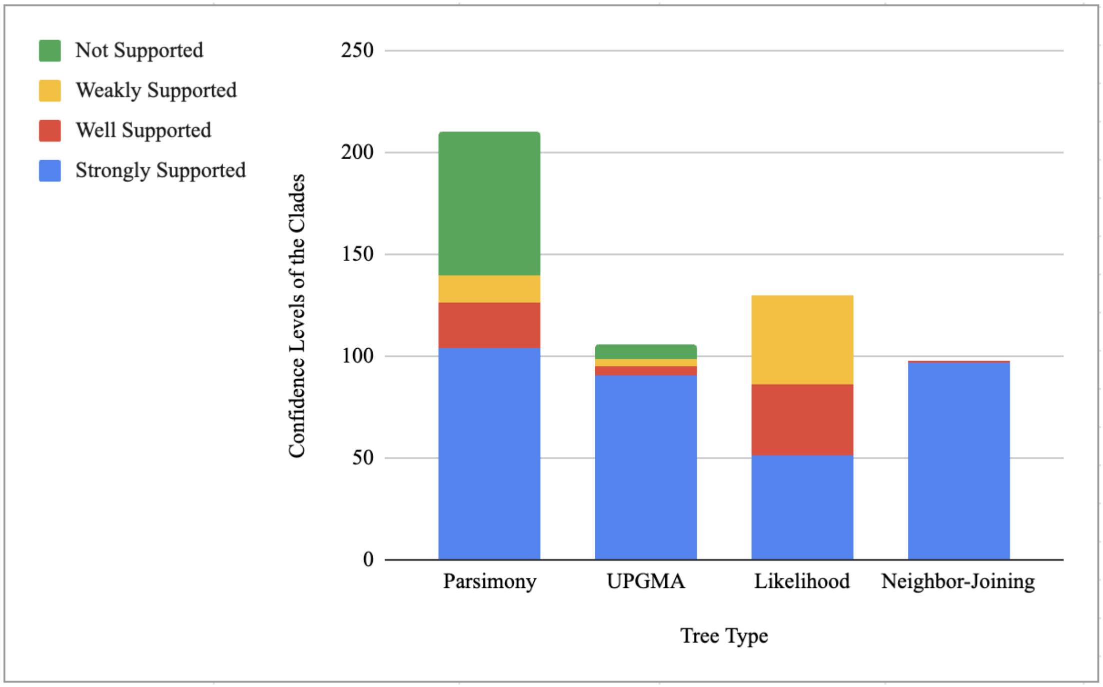

Figure 5. Graph summarizing the confidence levels of the clades produced for each tree type

Conclusion

Considering all of the above, the results have demonstrated that the neighbor-joining method of phylogeny is the best mechanism to display the evolutionary history of COVID-19. The alternative hypothesis that was inferred was incorrect. It was initially thought that the maximum likelihood method of phylogeny would display the evolutionary history of COVID-19 most accurately because it's known as one of the most accurate phylogenetic methods in taxonomy. However, the rate of evolution of COVID-19 was overlooked. In a number of cases, it had been previously proposed that the evolution of viruses within certain RNA virus genes proceeds at a constant rate. For example, this clock-like evolution had been proposed for human influenza. The constancy of evolutionary rates does not hold in many other cases. Bacterial mutations happen at random but the rate at which they occur depends on the actual virus itself. The RNA polymerase that makes copies of the genes of the virus generally lacks the ability to proofread its code. Making RNA viruses like SARS - CoV 2 prone to high mutation rates. This resulted in multiple variants arising during the pandemic. Thus, the COVID-19 DNA sequences that were collected for the experiment did not evolve at a clock-like pace making the neighbor-joining method of phylogeny the most suitable method for displaying its evolutionary history.

Although the likelihood method did do a decent job it still could not compare to the neighbor-joining method because likelihood is focused on models of evolution rather than actual phylogenetic patterns. Parsimony's failure was to be expected as it is known as the simplest form of phylogenetic analysis and its simplicity makes it not the most accurate. Even though neighbor-joining did prove to be the most accurate method in this project, further research is still necessary for declaring the neighbor-joining method of phylogeny as the best possible method to accurately display the evolutionary history of SARS - CoV 2.

Limitations

There are a number of things that would be altered if this investigation were to be conducted again in the future. Firstly, other phylogenetic methods such as Bayesian Analysis would be considered and compared to the methods that were used here. It would be interesting to see if these methods would outperform neighbor-joining. The second thing that could possibly be changed is building a computer from scratch that contains more memory with a better processor. Having more RAM would make it possible to conduct more trials, although it is not entirely certain whether a common person would be able to construct a computer with that good of a processor and if that much RAM would even be good enough to conduct over 3 trials. It is likely that this kind of in depth study would only be able to be conducted using a laboratory computer only available to scientists. That being said, this study has proven the ability of the neighbor-joining method to display COVID-19's evolutionary history and therefore, its ability to be utilized by scientists to get a deeper taxonomic understanding of the virus, and potentially gaining the ability to predict future mutations of the virus before they come into being. Being able to have such a lens into the virus's future would make creating effective vaccines easier, not only for COVID-19 but for other RNA viruses with a similar structure to SARS - CoV 2. The final outcome of this investigation demonstrates that the neighbor-joining method of phylogenetic analysis is the best method to display the evolutionary history of COVID-19 most accurately.

Acknowledgments

I would like to thank the instructors at St. Andrews Episcopal Upper School for providing me with the time and facilities to conduct this study. I would also like to thank Professor Kiel Kietlinski for kindly donating his PHD textbooks and helping me begin my investigation.

References

Bartolucci, F., & Scrucca, L. (2010). *Point Estimation Methods with Applications to Item Response Theory Models. In International Encyclopedia of Education (3rd ed., pp. 366–373). essay,* Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01376-2.

Berkeley Edu. (n.d.). *Phylogenetic systematics (evolutionary trees)*. Understanding Evolution . https://evolution.berkeley.edu/evolibrary/article/phylogenetics_08#:~:text=What%20is%20parsimony%3F,requires%20the%20fewest%20evolutionary%20changes

Cunningham, M. (2014). *How can you tell if a phylogenetic tree is accurate?* . Researchgate. https://www.researchgate.net/post/How-can-you-tell-if-a-phylogenetic-tree-is-accurate

Goloboff, P.A., Catalano, S.A., Marcos Mirande, J., Szumik, C.A., Salvador Arias, J., Källersjö, M. and Farris, J.S. (2009), Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics, 25: 211-230. https://doi.org/10.1111/j.1096-0031.2009.00255.x

Hall, B. G. (2018b). *Phylogenetic trees made easy: A how-to Manual*. Sinauer Associates, imprint of Oxford University Press.

Huelsenbeck, J. P., & Rannala, B. (2004, December 1). *Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models*. OUP Academic. https://academic.oup.com/sysbio/article/53/6/904/1651356

Katoh, K. (2006). *Multiple alignment program for amino acid or nucleotide sequences*. MAFFT 7.526. https://mafft.cbrc.jp/alignment/software/

Lemey, P., Salemi, M., & Vandamme, A.-M. (2009). *The phylogenetic handbook: A practical approach to DNA and protein phylogeny*. Cambridge University Press.

Maddison, W. P., & Maddison, D. R. (2023). *Mesquite: A modular system for evolutionary analysis. Version 3.81*. Mesquite Project. http://www.mesquiteproject.org/

McLennan, D. A. (2010, September 29). *How to read a phylogenetic tree - evolution: Education and outreach*. BioMed Central. https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-010-0273-6

Miller, R. E., McDonald, J. A., & Manos, P. S. (2004). Systematics of Ipomoea subgenus Quamoclit (Convolvulaceae) based on ITS sequence data and a Bayesian phylogenetic analysis. *American journal of botany*, *91*(8), 1208–1218. https://doi.org/10.3732/ajb.91.8.1208

Munar, M. P. (2021, May 29). *How to analyze phylogenetic trees | interpret bootstrap values and sequence divergence*. YouTube. https://m.youtube.com/watch?v=QlMwSqNbKA8

Nei, M. (1993). *Molecular Evolutionary Genetics Analysis*. MEGA. http://www.megasoftware.net/

Rambaut, D. (2018). *Tracer v1.7.2* . Tracer | BEAST Documentation. https://beast.community/tracer

Swofford, D. (2003). *PAUP* (* phylogenetic analysis using PAUP)*. PAUP Phylogenetic Analysis Using PAUP. https://paup.phylosolutions.com/

U.S. National Library of Medicine. (1982). *GenBank Overview*. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/genbank/