Academic Science Journal

Phylogenetic Tree Construction of Gamma Coronavirus Genera and SARS-CoV-2

Alaa Khudair Abbas Al-Khafaji* and Bashar Talib Al-Nuaimi

Computer Science Department – College of Science – University of Diyala

*programmeralaa90@gmail.com

Received: 28 September 2022 **Accepted: 28 February 2023**

DOI: https://doi.org/10.24237/ASJ.02.02.695B

Abstract

The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes coronavirus disease 2019 (COVID-19), has spread worldwide. Therefore, this study aimed to build a phylogenetic tree of complete genomes of SARS-CoV-2 and other species of Gamma coronavirus to explore the possibility of finding the evolutionary relationships between them and wished to analyze them in order to forecast the best trees illustrating the sequences' evolutionary relationships and obtain a well-supported phylogenetic tree by using the Neighbor-Joining (NJ) and Maximum Likelihood (ML) methods after performing multiple sequence alignment (MSA). This study utilized 16 isolates of Gamma coronavirus species and SARS-CoV-2 retrieved from the NCBI (National Center for Biotechnology Information) database for this investigation. The experimental outcomes when applying the two methods to the same dataset show that a well-supported and trustworthy phylogenetic tree was obtained with a bootstrapping value of 100% for all branches of the tree when applying the ML method. Additionally, a well-supported and fast-constructing phylogenetic tree was obtained through the NJ method for all branches except one, where the bootstrapping value appeared to be 56%. The research was conducted in 2022 at the College of Science, Diyala University.

Keywords: SARS-CoV-2, phylogenetic tree construction, Maximum-Likelihood, COVID-19, phylogenetic inference, Neighbor-Joining.

Academic Science Journal

بناء شجرة النشوء والتطور لجينات فيروس جاما التاجي وسارس-كوفيد-2

علاء خضير عباس الخفاجي وبشار طالب النعيمي

قسم علوم الحاسبات – كلية العلوم – جامعة ديالى

الخلاصة

تفشي فيروس كورونا المتلازمة التنفسية الحادة الوخيمة 2 (SARS-CoV-2)، الذي كان سبب حدوث مرض فيروس كورونا 2019 (COVID-19) انتشر في جميع أنحاء العالم. لذلك، هدفت هذه الدراسة إلى بناء شجرة النشوء والتطور من الجينوم الكمال لـ SARS-CoV-2 وأنواع أخرى من فيروس غاما التاجي لاستكشاف إمكانية العثور على العلاقات التطورية بينهما وتحليلها من أجل التنبؤ بأفضل الأشجار التي توضح العلاقات التطورية للتسلسلات والحصول على شجرة نشوء وتطور مدعومة جيدًا باستخدام طرق الانضمام للجوار والاحتمالية القصوى بعد إجراء محاذاة تسلسل متعددة. استخدمت هذه الدراسة 16 عزلة من أنواع فيروس غاما التاجي وSARS-CoV-2 المسترجعة من قاعدة بيانات NCBI (المركز الوطني لمعلومات التكنولوجيا الحيوية) لإجراء هذا التحقيق. أظهرت النتائج التجريبية عند تطبيق الطريقتين على نفس مجموعة البيانات أنه تم الحصول على شجرة نسالة مدعومة جيدًا وجديرة بالثقة بقيمة تمهيدية بنسبة 100 ٪ لجميع فروع الشجرة عند تطبيق طريقة الاحتمالية القصوى. بالإضافة إلى ذلك، تم الحصول على شجرة نسالة مدعومة جيدًا وسريعة البناء من خلال طريقة الانضمام للجوار لجميع الفروع باستثناء فرع واحد، حيث بدت قيمة التمهيد له 56٪. تم اجراء البحث في عام 2022 في كلية العلم جامعة ديالى.

كلمات مفتاحية: سارس كوفيد 2، بناء شجرة النشوء والتطور، الاحتمال الأقصى، الاستدلال الوراثي، ربط الجوار، كوفيد 19.

Introduction

Coronaviruses belong to the Coronaviridae family and have a single strand of the positive-sense RNA genome of lengths 26–32 kb [1]. They have been found in a variety of avian hosts and mammals, including bats, mice, and dogs, among others [2]. In December 2019, a novel coronavirus, SARS-CoV-2, was discovered in Wuhan, China. It was responsible for the coronavirus disease 2019 (COVID-19) (WHO situation report May 30, 2020) [3]. SARS-CoV-2 caused a worldwide pandemic and rapidly spread throughout the world [4]. The SARS-CoV-2 virus's distinguishing characteristics is that it spreads very quickly. The virus's

evolutionary analysis revealed this. SARS-CoV-2 genetic variety was observed in a short period because of genetic variance in new viruses from their common ancestry. Even if SARS-CoV-2 is genetically similar to other coronavirus species, they demonstrate significant variances in epidemiology, pathogenicity, and host spectrum [5]. We propose phylogenetic tree construction for Gamma coronaviruses with an outgroup genome (SARS-CoV-2) to estimate and visualize the evolutionary relationships between Gamma coronaviruses genomes and the SARS-CoV-2 genome. In the field of bioinformatics, viruses' evolutionary analysis is an invaluable resource.

A phylogenetic tree is used to make estimates about how these species are related. These trees link the species together and show how they evolved [6]. Using the genes these coronavirus species have in common, a well-supported phylogenetic tree has been constructed for the Gamma coronavirus and SARS-CoV-2 genomes [7].

The NJ method is the most popular and quickest distance-based approach for reconstructing a phylogenetic tree from data sequences. An approach dependent on the distance matrix is independent of the molecular clock. It uses genetic distances as a clustering metric [8]. Because the ML method is more complex, it requires many computational steps, and the number of steps increases rapidly with the number of sequences, so it is limited to a small number of sequences. They can be performed on a supercomputer to examine many sequences simultaneously [9]. The current evolutionary methods need to be rethought to be more accurate reflections of reality [10].

The rest sections in this paper are arranged as follows, Section 2 will go through some of the most related works, while Section 3 illustrate the materials and methods used in the system, Section 4 shows the experimental results and provides a discussion upon these results also, Finally, the conclusion is presented in Section 5.

Related works

Constructing a phylogenetic tree is an important challenge in the field of bioinformatics. There has been significant progress in computer hardware and software resources in recent years,

prompting many researchers to investigate phylogenetic trees to understand the evolutionary relationships between different species.

1. T. Li, D., *et al.* [11], proposed to use the Matrix Representation with Parsimony (MRP) pseudo-sequence super-tree method and ML method to investigate the evolutionary history of SARS-CoV-2 and gain a deeper understanding of its origin and evolutionary relationships. Protein-coding sequences (CDS) of 120 SARS-CoV-2 and other coronavirus sequences were obtained from the NCBI and GenBank databases. The sequences were grouped by orthologous proteins and aligned using Fast Fourier Transform (MAFFT) with the L-INS-i method. ML phylogenetic trees were constructed and estimated using PhyML V3 with 100 bootstrap resampling based on each CDS to provide more information on the evolution of SARS-CoV-2 compared to MRP phylogenetic trees.

2. Turista, A., *et al.* [12], in this work, the phylogenetic model and tree visualization were constructed using the MEGA X program with the ML algorithm. In this study, Indonesian SARS-CoV-2 nucleocapsid gene sequences and various types of coronavirus sequences from multiple countries were obtained from the Global Initiative on Sharing All Influenza Data (GISAID) and NCBI databases to build phylogenetic trees. The validation of the phylogenetic tree was conducted by applying the Tamura-Nei (TN93) substitution model and running 1000 bootstrapped on the input datasets. The aim of this work was to determine the spread of COVID-19 in Indonesia and construct a phylogenetic tree of the SARS-CoV-2 virus from Indonesian samples and samples from multiple countries, including other coronaviruses, to understand their relationship.

3. Awoyelu, E., *et al.* [13], proposed to investigate the evolution of SARS-CoV-2 in Nigeria and determine the common ancestor of each strain by analyzing 39 SARS-CoV-2 nucleotide sequences obtained from the GISAID database. These sequences chosen for this study based on the individual's travel history and the date the sample was collected were taken from individuals in Congo, South Africa, Nigeria, France, Italy, and China. The evolutionary history was inferred using the ML algorithm based on the General Time Reversible (GTR) substitution model by MEGA 5.2 software. Multiple Sequence Alignment (MSA) was utilized with

ClustalW on the MEGA X tree-building program to align the sequences and identify the conserved regions. The bootstrap consensus tree inferred from 1000 resampling was taken to represent the evolutionary history of the species analyzed.

4. Adebali, A., *et al.* [14], this work proposed building a ML phylogenetic tree using the IQ-TREE software tool to reveal the evolutionary relationships between sequences. Sequences of over 15,000 genomes were obtained from the GISAID database. Following MSA alignment with the MAFFT algorithm for those sequences, a phylogenetic tree was built using the ML algorithm, the GTR substitution model, and 1000 bootstrap replicates. The genomes were then clustered in the phylogenetic tree based on their clade distribution, genomic characteristics, and links to previous studies. Further analysis was conducted on the clusters, mutations, and transmission patterns of the genomes from Turkey. This work aims to track the spread of SARS-CoV-2 sequences related to Turkey by analyzing representative samples using the phylogenetic tree.

5. Hussen, D., *et al.* [15], proposed phylogenetic analysis based on a nearly full gene sequence of SARS-CoV-2 samples collected from various geographical locations was performed. The sequences were obtained from the GISAID and GenBank databases and aligned using the Clustal W algorithm. The NJ algorithm, bootstrapped with 1,000 replicates, was used for sequence analysis of the coronaviruses by the MEGA 7 software to examine similarities and differences among the Erbil genome sequence and the 27 SARS-CoV-2 genome sequences downloaded. The main aim of this study was to carry out genome sequencing of the local SARS-CoV-2 variant to gain insights into the mutational variation of the virus compared to the first reported genome sequence of the SARS-CoV-2 genome isolated from Wuhan, China. In addition, this sequencing enabled a precise understanding of the novel variations, which are key indicators for viral development and vaccine response.

Although these studies produced good results when applied to SARS-CoV-2 and other coronavirus sequences obtained from NCBI datasets, they did not use the complete genome of SARS-CoV-2 and Gamma coronavirus sequences and did not obtain a well-supported tree with a bootstrapping value of 100% for all tree branches in any of the methods used in their study.

It simply means that there is some uncertainty or variation in the relationships among the different species being analyzed.

Material and Methods

We have followed a general diagram consisting of several steps to construct a phylogenetic tree of the gamma coronavirus genus and SARS-CoV-2 using NJ and ML methods, as shown in Figure (1).

1. Selection Coronavirus Molecular.

The first step is Gamma coronavirus molecular data selection. We used the nucleotide sequence of the complete genome while studying. The complete genome of (15) Gamma coronaviruses sequences and SARS-CoV-2 (NC_045512) RefSeq were summoned directly by the NCBI (National Center for Biotechnology Information) database (https://www.ncbi.nlm.nih.gov/) by accession number, which will be available on September 1, 2022.

Gamma coronavirus is one of the four coronavirus genera (Alpha, Beta, Gamma, and Delta). It belongs to the Coronaviridae subfamily, Orthocoronavirinae. They are zoonotic, enveloped, positive-sense, single-stranded RNA viruses. Coronaviruses are capable of infecting both people and animals. Table1 contains information on these Gamma coronavirus genomes.

One of the SARS-CoV-2 sequences is NC_045512 Reference Sequence (RefSeq), the first SARS-CoV-2 isolate sequencing discovered in Wuhan, China, and dated 2019-12-20 based on the appearance of the illness in the patient. Which belongs to the Beta coronaviruses subfamily and is classified in the Sarbecovirus subgenus [11, 15].

The sequences' integrity was verified. Furthermore, databases were created by identifying genomic sequences with the country, host, date of collection, and other information.

The Alpha and Beta genera are derived from the gene pool of bats, while the Gamma and Delta genera are derived from the gene pools of avian and pigs, respectively. The avian

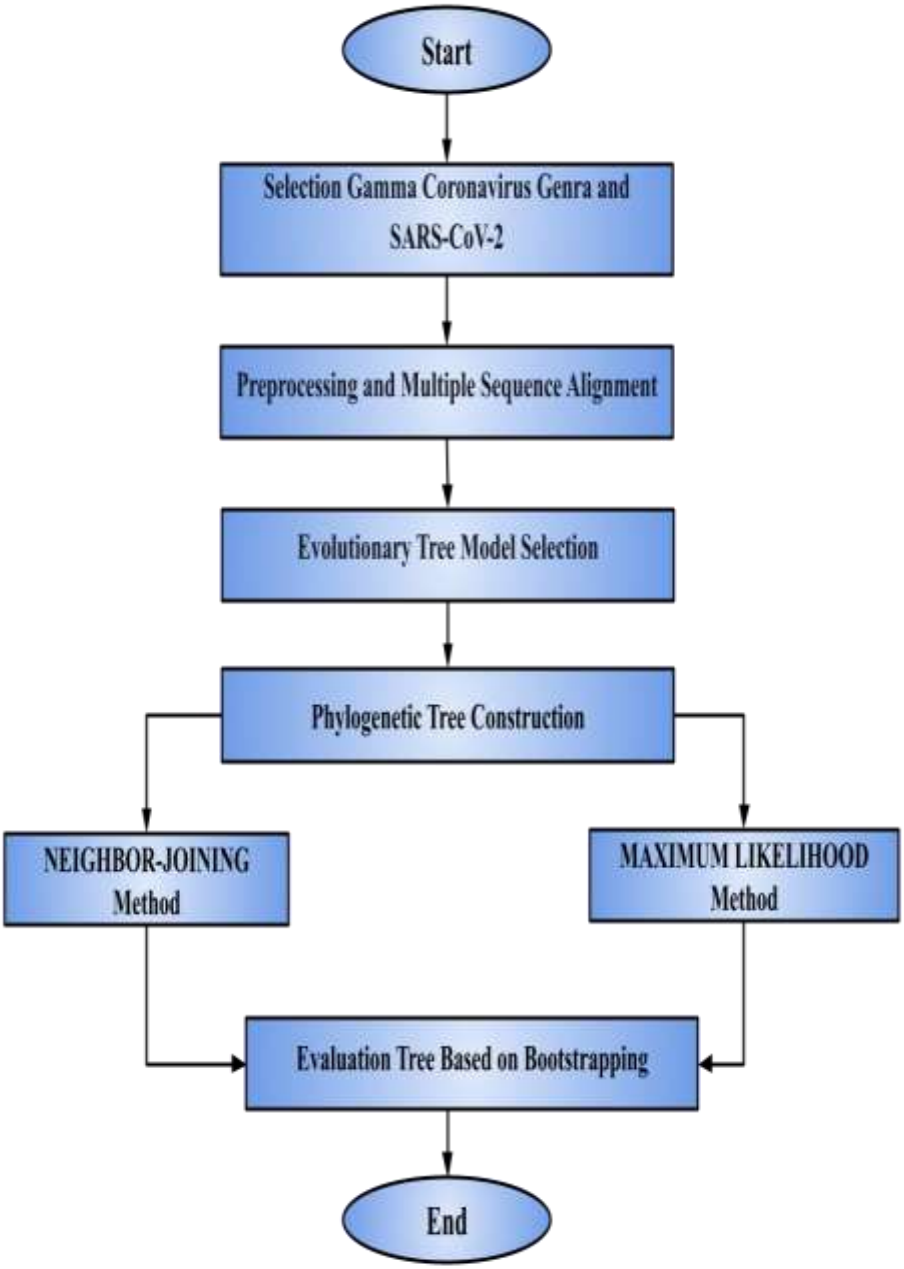

Figure 1: General diagram of the proposed system.

coronaviruses are gamma coronavirus, commonly known as coronavirus group 3 [16].

Table 1: Information about Gamma Coronavirus genomes

ACCESSION	DEFINITION	GENOME SIZE(BP)	HOST	COUNTRY	COLLECTION DATE
MW792514	Infectious bronchitis virus	27693	Gallus	China	2020
MW024789	Avian coronavirus	27486	Gallus	USA	06-Jan-2020
LN610099	Guinea fowl coronavirus	27471	guinea fowl	France	2011
KF793826	Bottlenose dolphin coronavirus HKU22	31775	bottlenose dolphin	Hong Kong	Oct-2013
MT367412	Turkey coronavirus	27614	Meleagris gallopavo	Poland	21-Jun-2016
KM454473	Duck coronavirus	27754	Duck	China	2014
MK359255	Canada goose coronavirus	28539	Branta canadensis	Canada	Aug-2017
MZ327724	Guinea fowl coronavirus	27773	Guinea fowl	France	May-2017
MT591566	Infectious bronchitis virus	27227	Chicken	Germany	Feb-2016
KR822424	European turkey coronavirus	27739	Turkey	France	2008
EU111742	Coronavirus SW1	31686	Delphinapterus leucas	China	Aug-2007
MW436465	Anser fabalis coronavirus NCN2	28466	Anatidae sp.	China	2019-03
MN690608	Bottlenose dolphin coronavirus	31728	Bottlenose dolphin	USA	2020
JQ088078	Infectious bronchitis virus	27664	Chicken	Sweden	2010
JF274479	Infectious bronchitis virus	27678	Chicken	China	07-May-2007

2. Preprocessing and Multiple Sequence Alignment.

During the preprocessing stage, data on the coronavirus is initialized and prepared for processing in the following stage. After retrieving SARS-CoV-2 and coronavirus sequences from NCBI in data frame format (which contains information such as sequence accession number, sequence character, sequence name, sequence length, country, host, collection date, etc.), they must be converted to FASTA file format. Only the sequence accession number and sequence character string are needed from the data frame during this conversion, and the resulting FASTA file is saved on the computer to be processed and used in the proposed

system. With synteny, a comparison of SARS-CoV-2 and Gamma coronavirus genomes will be accomplished. The synteny aims to find the best match or difference between the input database genomes. When regions in a genome are the same, they are called syntenic matches. This is often accomplished through matching runs of reciprocal best blast hits using k-mer matching. The results of the first representation of adjacent pairs synteny blocks representation, shown in Figure 2.a, and the dot plot or synteny map visualization, shown in Figure 2.b, for each coronavirus genera, from which these genomes were selected based on their genetic affinity to SARS-CoV-2. Sequence alignment is the process of rearranging nucleotide or protein sequences to identify areas of similarity that may be used to align the complete sequence. An evolutionary relationship may be determined by the similarity between the sequences [17]. Multiple Sequence Alignment (MSA) is a computational technique used to align more than two nucleotide or protein sequences to identify regions of similarity and dissimilarity between them. MSA involves aligning multiple sequences together by inserting gaps (or dashes) to represent the insertion or deletion of nucleotides or amino acids in each sequence. The output of an MSA is a consensus sequence that represents the most common nucleotide or amino acid at each position in the aligned sequences. MSA is simple when all the sequences are similar but challenging if they are not. If so, many gaps will be required [18].

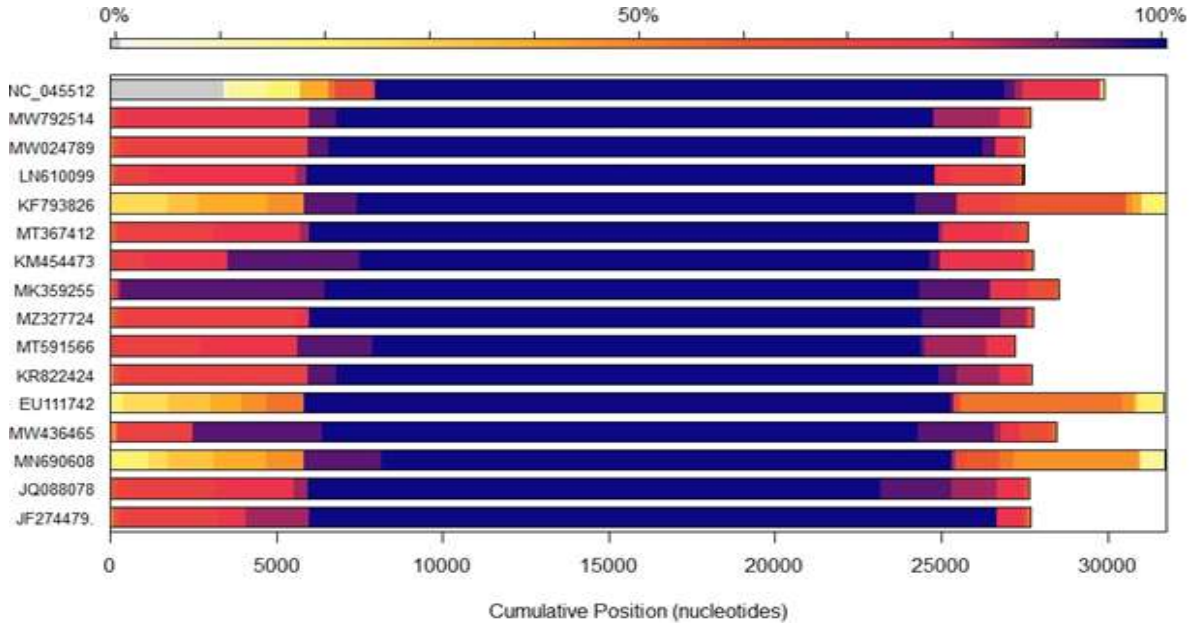

Figure 2-a

Academic Science Journal

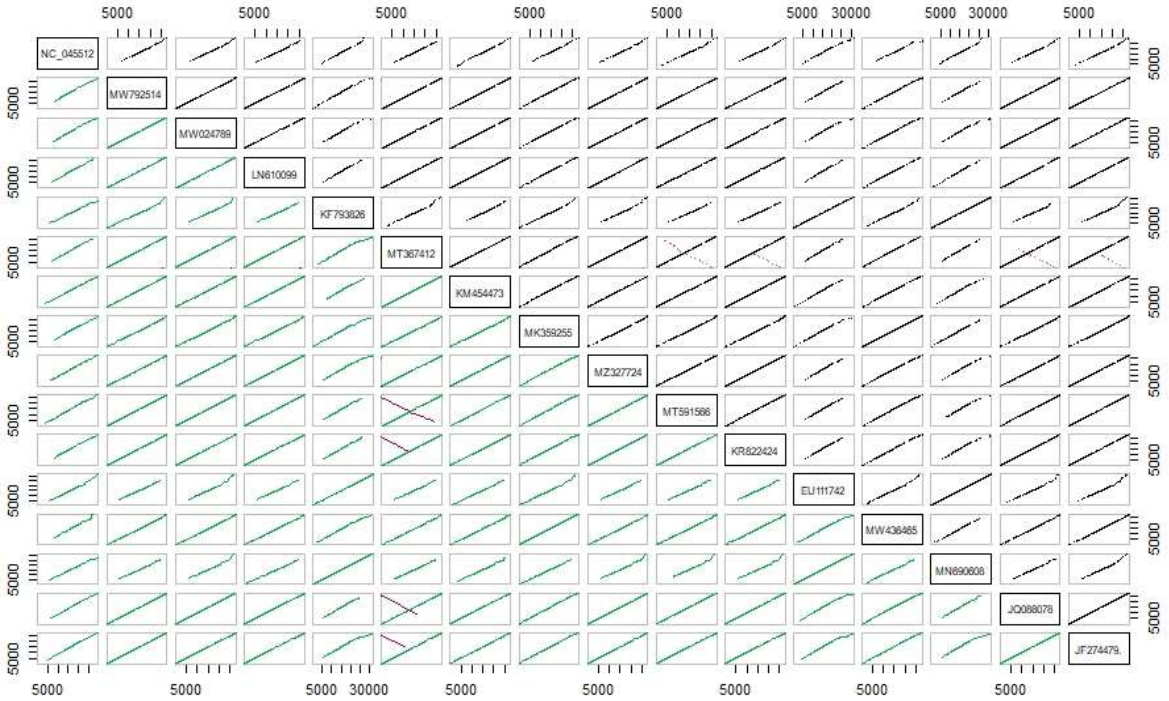

Figure 2-b

Figure 2-a Adjacent pairs synteny blocks representation of Gamma coronavirus sequences with outgroup NC_045512.2 SARS-CoV-2. When the genome shares the same region with the first genome, their coloration is the same hue, or grey if they don't. the dot plot supplies an alternate synteny map visualizing of Gamma coronavirus. Black-colored diagonal stripes represent syntenic areas having identical directions. All these species have significance.

MSA is an important stage because it demonstrates positional sym metry in sequence evolution. Only a successful and accurate sequence alignment leads to a genealogically interrelated tree [19]. Furthermore, it may assist in the detection of mutations or recombination events in pairs of closely related genomes [10].

We performed multiple sequence alignment using the ClustalW algorithm [20] for the complete genome of Gamma coronavirus, which yielded good accuracy. Figure (3) shows a partial multiple sequence alignment of coronaviruses and SARS-CoV-2.

Clustal W is a widely used for performing MSA, which uses a progressive alignment strategy. It starts by comparing pairs of sequences and generating a guide tree based on the similarity between them. The sequences are then aligned according to the guide tree, with the most similar sequences being aligned first, followed by progressively less similar sequences.

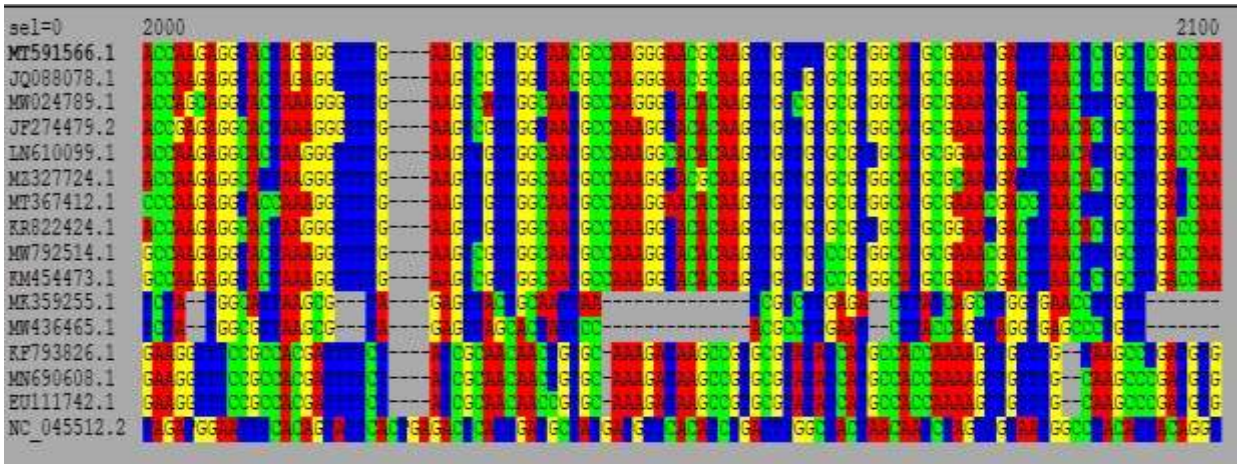

Figure 3: MSA of Gamma coronavirus sequences with outgroup NC_045512.2 SARS-CoV-2 RefSeq.

3. Evolutionary Tree Model Selection.

To construct phylogenies, several nucleotide substitution models are available for nucleotide and amino acid evolution. The complexity of the models varies with the parameters they use. These models differ in how multiple substitutions of each nucleotide are treated.

We chose Felsenstein 1981 (F81) substitution model for the Neighbor-Joining method. Felsenstein 1981 (F81) extended the JC69 model by relaxing the assumption of equal frequencies. Moreover, we optimized the different model parameters and branch lengths for the selected model of nucleotide evolution for the Maximum Likelihood method, which used (GTR+I+G) substitution models as recommended by JModelTest 2.1.10 to get optimized.

4. Phylogenetic Tree Construction.

To construct phylogenetic trees, statistical approaches are utilized to find the best explain the evolutionary relationships of the aligned sequences in a dataset [21]. We have used two

methods to construct the phylogenetic tree in this study, the first is Neighbor-Joining method, and the second is the Maximum Likelihood method.

The NJ and ML methods were used to determine and show the evolutionary relationships between Gamma coronavirus genomes and SARS-CoV-2. One of the most widely used methods of statistical estimation is that of NJ and ML [20, 21].

The NJ method [24] is based on the distance matrix to reconstruct a phylogenetic tree from aligned data sequences. It uses genetic distances as a clustering metric. The composition of the tree begins with a star, and then each pair is computed and evaluated based on whether it is linked or not, as well as the length of the branches, in order to determine the composition of the tree and the pair that yields the least amount in the collection. The number of pairs of neighbors in a tree depends on the tree 's topology. We used the NJ method to construct a phylogenetic tree. Then, we optimize the different model parameters and branch lengths for the selected model of nucleotide evolution, which uses F81 substitution models. Table 2 shows the distance matrix results of Gamma coronavirus genera.

The ML method is utilized to find the topologies and branches length that is the highest likelihood of producing the aligned data, hence giving the substitution models and tree. After the alignment stage, the likelihood value is calculated by evaluating several nucleotide substitution models [23, 24]. For the purpose of tree reconstruction, all potential sequence combinations are determined using this method, which is used to fulfill the searching area [7]. The ML is reported to be best under all circumstances [22].

ML is the most often used method for statistical inference because of its high consistency and asymptotic normal distribution [27]. The fundamental problem of this method is that they are computationally costly. However, with modern computers, this is less of a major issue [7].

The ML approach was used in this work to construct a phylogenetic tree, which can be incorporated into a statistical framework for estimating model parameters with all multiple sequence alignment data. The likelihood of a phylogenetic tree is computed first. Then, we optimized the different model parameters and branch lengths for the selected model of nucleotide evolution, which used (GTR+I+G) substitution models as recommended by

JModelTest 2.1.10 to get optimized. The output tree was visualized utilizing the Interactive Tree of Life (iTOL) online tool [28].

Table 2: Distance matrix of Gamma coronavirus species.

	MT591566	JQ088078	MW024789	JF274479	LN610099	MZ327724	MT367412	KR822424	MW792514	KM454473	MK359255	MW436465	KF793826	MN690608	EU111742	NC045512
MT591566	0.000 00000															
JQ088078	0.080 41281	0.000 00000														
MW024789	0.127 96173	0.113 87294	0.000 0000													
JF274479	0.139 88313	0.122 00772	0.107 5628	0.000 0000												
LN610099	0.154 00748	0.154 83593	0.160 1205	0.166 0297	0.000 00000											
MZ327724	0.143 26817	0.147 44605	0.152 7662	0.162 5333	0.078 46988	0.000 00000										
MT367412	0.160 62886	0.165 68558	0.165 7053	0.171 1423	0.111 29039	0.112 16405	0.000 0000									
KR822424	0.146 66308	0.147 87079	0.152 6488	0.158 1580	0.102 23158	0.100 30987	0.133 3424	0.000 000)								
MW792514	0.169 31142	0.162 83640	0.169 3869	0.171 6068	0.211 80128	0.208 05042	0.208 1232	0.206 0654	0.000 0000							
KM454473	0.310 28110	0.305 78069	0.306 0479	0.306 5421	0.281 19684	0.278 85414	0.283 8073	0.272 6023	0.268 2899	0.000 0000						
MK359255	0.525 17853	0.524 15224	05191 236	0.520 3107	0.550 06764	0.546 68160	0.546 7355	0.545 9300	0.527 4487	0.555 3747	0.000 0000					
MW436465	0.528 03961	0.521 65179	0.514 0857	0.517 1011	0.547 12059	0.548 96719	0.544 2057	0.546 8259	0.523 7685	0.548 3084	0.152 7495	0.000 0000				
KF793826	0.774 48060	0.783 81402	07779 996	0.783 4946	0.783 57714	0.781 45483	0.780 7792	0.785 9543	0.787 9781	0.820 8757	0.850 7198	0.850 6609	0.000 00000			
MN690608	0.774 79161	0.782 38343	0.776 0545	0.782 6294	0.783 26701	0.782 16789	0.781 2278	0.785 5087	0.787 0357	0.818 6413	0.849 7220	0.850 1419	0.025 08744	0.0000 0000		
EU111742	0.781 85487	0.787 02346	0.784 8548	0.785 4136	0.782 85065	0.785 44567	0.782 9529	0.787 2671	0.790 4803	0.823 5379	0.858 9070	0.857 7107	0.056 50119	0.0547 7874	00000 0000	
NC045512	0.878 70078	0.884 74286	0.886 6870	0.885 9456	0.881 76307	0.886 40331	0.888 1547	0.885 1953	0.886 1646	0.908 0109	0.944 5328	0.956 7097	1.024 21402	1.0222 5142	1.0266 9739	0.0000 0000

5. Phylogenetic Tree Bootstrapping.

Almost every statistical estimate may be evaluated for accuracy using the bootstrapping method, which is computer-based. Nonparametric estimate issues where analytic approaches are impracticable and commonly utilized can benefit from this technique [29]. The total amount

of characters present in the dataset can potentially have an artifactual effect on the values obtained via Bootstrap [30].

The bootstrapping process is performed to measure the NJ and ML tree's accuracy and prove that these methods can find the supported tree. Assign the NJ tree, distance matrices, number of bootstrap replicates to 1000 and substitution model F81 to perform a bootstrap analysis for the NJ tree. Figure 4 shows a representation of the phylogenetic trees using the NJ method with bootstrap values.

Furthermore, to do (non-parametric) bootstrap analysis for the ML tree, assign the ML tree, set the number of bootstrap replicates to 1000, and use the substitution model GTR+I+G for computing the bootstrapping values of the ML tree (See Figure 5).

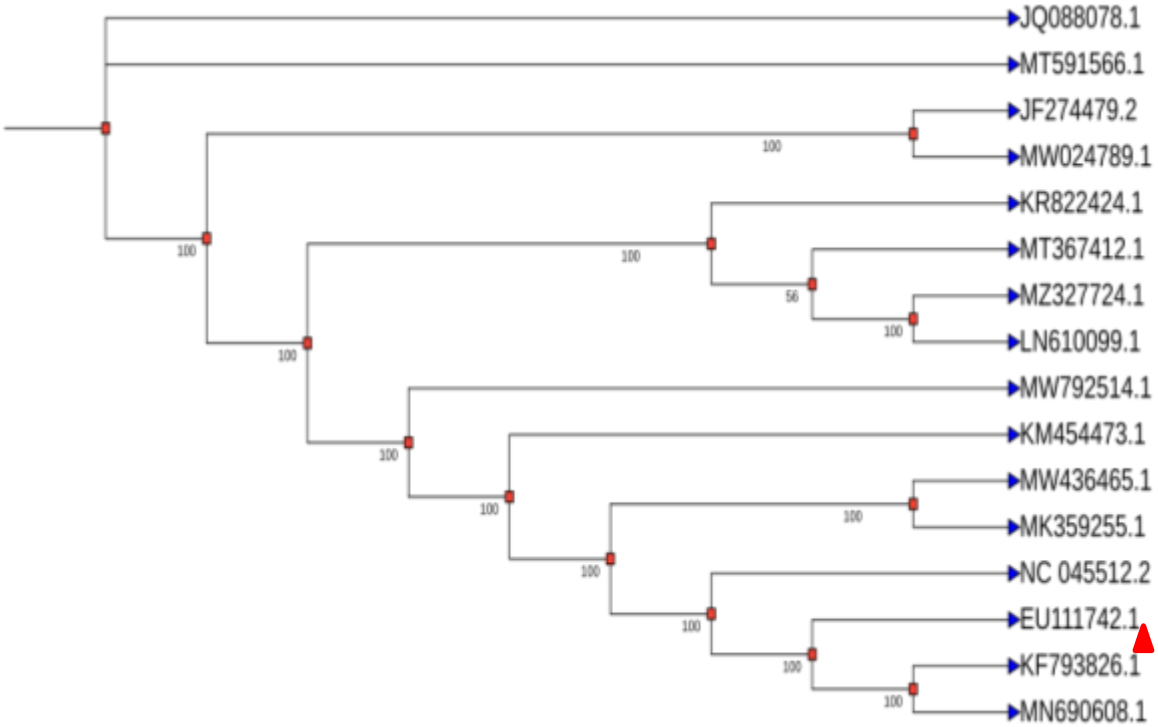

Figure 4: The phylogenetic tree of (15) Gamma Coronaviruses using the NJ method with 1000 bootstraps, the red triangle indicating SARS-CoV-2 RefSeq.

Academic Science Journal

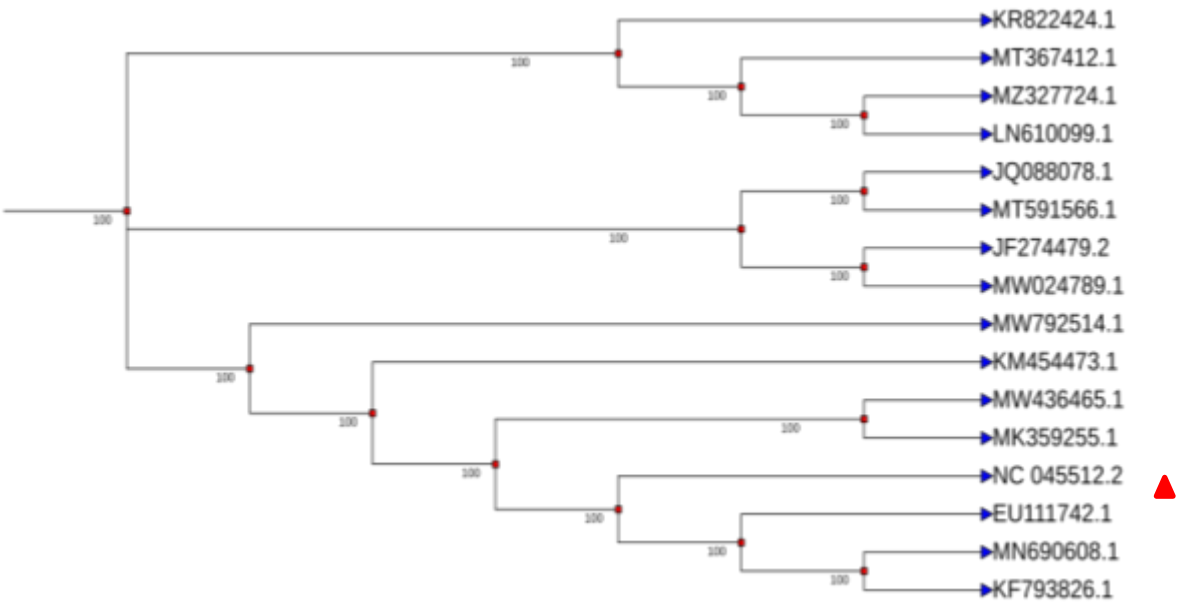

Figure 5: The phylogenetic tree of (15) Gamma coronaviruses using the ML method with 1000 bootstraps, the red triangle indicating SARS-CoV-2 RefSeq

Results And Discussion

Two phylogenetic trees based on complete genome sequences have been obtained using the NJ and ML methods to show the evolutionary relationships among SARS-COV-2 and Gamma coronaviruses. The first is for Gamma coronavirus species with outgroup (NC_045512.2) SARS-CoV-2 using the NJ method shown in Figure 4, and the second by using the ML method as shown in (Figure 5). We removed samples that were too divergent from others. Despite the close relationship among these organisms, the genome length and content are very different, as shown in Table 1.

For all the 15 Gamma coronavirus genomes with the SARS-CoV-2 Wuhan-Hu-1 (NC_045512) outgroup, the MSA process was performed to align them as previously stated (show Figure 3). Figure (2) shows the initial representation of synteny for all of them. Gamma coronavirus genomes are remarkably similar and match each other perfectly, as observed. Furthermore, just a few recombination events have taken place inside these genomes and reflect divergent

evolutionary origins. The areas are also shown in orange and yellow. We also see that some regions in the outgroup genome appeared in grey.

Figure 4 depicts the NJ phylogenetic tree obtained for Gamma coronavirus genomes. Three main branches appear in the tree obtained, the first two branches each containing one Gamma coronavirus genome; this shows how far apart they are in terms of genetics and evolutionary relationship and that they are not descended from one ancestor, and this is due to large mutations and recombination. The third branch contained the rest of the Gamma coronavirus genomes, and all branches were good, except for a value of 56 that appeared between.

The SARS-CoV-2 were grouped into a group with (EU111742) Coronavirus SW1, (KF793826) Bottlenose dolphin coronavirus HKU22, and (MN690608) Bottlenose dolphin coronavirus. This shows the extent of their genetic closeness, the close evolutionary relationship between them, and that they are from one ancestor. We have noticed here that the sequence lengths of these three genomes are more than 31,000 and that the rest of the sequence lengths are between 27,000 and 28000, as shown in Table 1. Therefore, the length of the sequences could be one of the factors that led to the convergence between it and SARS-CoV-2, in addition to the genetic similarity and the few mutations.

The ML phylogenetic tree was obtained using Gamma coronavirus's complete genome in conjunction with the GTR Gamma substitution model. Figure (5) illustrates this, all branches show a well-supported. The obtained ML tree is trustworthy and well-supported.

Conclusion

In this article, a phylogenetic tree of Gamma coronaviruses was constructed to show the evolutionary relationships between the SARS-CoV-2 genome and other species of Gamma coronaviruses by using the NJ and ML methods. The ML method obtained a highly accurate and trustworthy phylogenetic tree, but it was slow, unlike the NJ method, which was fast in building the phylogenetic tree.

The phylogenetic tree and the ML method appear to be effective and beneficial for monitoring the evolution of the SARS-CoV-2 lineage and Gamma coronavirus genomes, due to the trustworthy and well-supported tree obtained. As a result, strains of SARS-CoV-2 should be continually monitored, as the development of variants in the COVID-19 pandemic is probable and may occur quickly.

A future work for research to intended to construct a phylogenetic tree of SARS-CoV-2s to show the evolutionary relationships with Delta coronavirus genomes, revealing a variety of SARS-CoV-2s mutations, insertion-deletion (INDELs) and single nucleotide polymorphisms (SNPs). It is critical that the variety and development of SARS-CoV-2 be closely monitored at all times in order to control and treat COVID-19 and avoid another epidemic.

References

1. S. Su, Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses, Trends Microbiol., 24(6), 490–502(2016)
2. D. Cavanagh, Coronavirus avian infectious bronchitis virus, Vet. Res., 38(2), 281–297(2007)
3. B. Morel, Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult, Mol. Biol. Evol., 38(5), 1777–1791(2021)
4. T. Li , Phylogenetic supertree reveals detailed evolution of SARS-CoV-2, Sci. Rep., 10(1), Dec(2020)
5. M. Sallam, A. Mahafzah, Molecular analysis of sars-cov-2 genetic lineages in Jordan: Tracking the introduction and spread of covid-19 UK variant of concern at a country level, Pathogens, 10(3), 1–12, Mar, (2021)
6. J. Rizzo, E. C. Rouchka, Review of Phylogenetic Tree Construction, Univ. Louisv. Bioinforma. Lab. Tech. Rep. Ser. 1, (2007)
7. B. Al-Nuaimi, B. Alkindy, J.-F. Couchot, M. Salomon, C. Guyeux, Ancestral Reconstruction and Investigations of Genomic Recombination on Campanulides Chloroplasts, (2017)

8. E. Rouchka, J. Rizzo, Review of Phylogenetic Tree Construction, Bioinforma. Lab. Tech. Rep. Ser., June, (2017)
9. D. W. Mount, Bioinformatics Sequence and Genome Analysis, (2001)
10. C. Guyeux, B. Al-Nuaimi, B. AlKindy, J. F. Couchot, M. Salomon, On the ability to reconstruct ancestral genomes from Mycobacterium genus, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics, 10208 LNCS, 642–658(2017)
11. T. Li, Phylogenetic supertree reveals detailed evolution of SARS-CoV-2, Nat. Sci. Reports, 10(1), (2020)
12. D. D. R. Turista, A. Islamy, V. D. Kharisma, A. N. M. Ansori, Distribution of COVID-19 and phylogenetic tree construction of sars-CoV-2 in Indonesia, J. Pure Appl. Microbiol., 14(May), 1035–1042(2020)
13. E. H. Awoyelu, E. K. Oladipo, B. O. Adetuyi, T. Y. Senbadejo, O. M. Oyawoye, J. K. Oloke, Phyloevolutionary analysis of SARS-CoV-2 in Nigeria, New Microbes New Infect., 36, 100717(2020)
14. O. Adebali, Phylogenetic analysis of sars-cov-2 genomes in Turkey, Turkish J. Biol., 44, Special issue 1, 146–156(2020)
15. B. M. Hussen, D. K. Sabir, Y. Karim, K. K. Karim, H. J. Hidayat, Genome sequence analysis of SARS-COV-2 isolated from a COVID-19 patient in Erbil, Iraq, Appl. Nanosci., 0123456789(2022)
16. P. C. Y. Woo, Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavi, J. Virol., 86(7), 3995–4008(2012)
17. S. Vijayakumar, A. Bhargavi, U. Praseeda, S. A. Ahamed, Optimizing sequence alignment in cloud using hadoop and MPP database, In: Proceedings - 2012 IEEE 5th International Conference on Cloud Computing, CLOUD 2012, 819–827(2012)
18. M. Dipl, I. C. Horejš, H. Horejš-Kainrath, U. Bodenhofer, J. Kepler, Multiple Sequence Alignment with R, (2016)

19. S. R. Amit Roy, Molecular Markers in Phylogenetic Studies-A Review, J. Phylogenetics Evol. Biol., 02(02), (2014)
20. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res., 22(22), 4673–4680(1994)
21. J. Xiong, Essential bioinformatics,(2006)
22. L. Le Cam, Maximum Likelihood: An Introduction,(1990)
23. I. J. Myung, Tutorial on maximum likelihood estimation, J. Math. Psychol., 47(1), 90–100(2003)
24. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., Mol. Biol. Evol., 4(4), 406–42591987)
25. L. Addario-Berry, B. Chor, M. Hallett, J. Lagergren, A. Panconesi, T. Wareham, Ancestral Maximum Likelihood Of Evolutionary Trees Is Hard, Journal of Bioinformatics and Computational Biology, 2(2), 257–271(2004)
26. A. K. A. Al-khafaji, B. T. Al-nuaimi, Phylogenetic Tree Construction to Reveal the Detailed Evolution of SARS-CoV-2, J. Algebr. Stat., 13(2), 538–549(2022)
27. K.-T. F. Jian-Xin Pan, Growth Curve Models and Statistical Diagnostics, Manchester 1984, (2006)
28. I. Letunic, P. Bork, Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation, Nucleic Acids Res., 49(W1), W293–W296(2021)
29. B. Efron, E. Halloran, S. Holmes, Bootstrap confidence levels for phylogenetic trees,(1996)
30. P. S. Soltis and D. E. Soltis, Applying the Bootstrap in Phylogeny Reconstruction,(2003)