Deciphering the Evolutionary Relationship of SARS-CoV-2: A Graph Theory Approach

Pranjal Kumar Bora[1], Sanchaita Rajkhowa[2,*], Arun Kumar Baruah[3], Papori Bora[4]

[1]Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh, Assam, INDIA.
[2]Centre for Biotechnology and Bioinformatics, Dibrugarh University, Dibrugarh, Assam, INDIA.
[3]Department of Mathematics, Dibrugarh University, Dibrugarh, Assam, INDIA.
[4]Department of Community Science, Assam Agriculture University, Jorhat, Assam, INDIA.

Submission Date: 12-10-2022; Revision Date: 01-11-2022; Accepted Date: 30-11-2022.

ABSTRACT

A novel coronavirus called SARS-CoV-2 was discovered in Wuhan, China, in December 2019, putting an end to everyone's daily activities. SARS-ancestry Cov-19's evolutionary position are yet unknown. In this study, we used a unique graph theory approach to determine the evolutionary relationships between two bat coronaviruses believed to be the relatives of 2019-nCoV and SARS-CoV and seven known human coronaviruses. The maximum likelihood (ML) method has been used to construct the phylogenetic tree of seven coronaviruses (SARS-CoV, 2019-nCoV, MERS-CoV, HCoV-NL63, HCoV-OC43, HCoV-229E, HCoV-HKU1) and other two coronaviruses (RsSHC014 and RaTG13). Network Merge application embedded in Cytoscape 3.3 is used to merge the entire module graph for a single sequence graph and validate the results of the phylogenetic tree using centrality measures with their correlation. RaTG13 is highly correlated with 2019-nCoV (SARS-CoV-2).The novelty of the work lies in the fact that it is one of the research work that shows the evolutionary relationship of protein sequences using rapidly changing regions, instead of the conserved regions. This new graph theory approach gives 100% accuracy of the evolutionary relationship of nine protein sequences with the biologically established one. This work can be used as a pipeline to accomplish evolutionary studies of protein sequences having adjacent residues with at least one common property.

Keywords: SARS-CoV-2, Phylogenetic relationship, Graph Theory, Centrality, Correlation

Correspondence:
Dr. Sanchaita Rajkhowa,
Centre for Biotechnology and Bioinformatics,
Dibrugarh University,
Dibrugarh, Assam-786004,
INDIA.

Email id: s_rajkhowa@dibru.ac.in

INTRODUCTION

The 2019 novel coronavirus SARS-CoV-2 (2019-nCoV) is the primary pathogen responsible for a new type of pneumonia, COVID-19. The SARS-CoV-2 is a beta-coronavirus of the family Coronaviridae, including the past epidemic-causing species SARS-CoV and MERS-CoV.[1] Both of these two viruses were thought to be emerged from bats and in time transferred to humans.[2] The genome of SARS-CoV-2 shows about 80% similarity with SARS-CoV and 96% with that of BatCoV RaTG1,[3] and the primary host is often recognized as a bat (*Rhinolophus affinis*).[3] The S$_2$ subunit of the spike surface glycoprotein of SARS-CoV-2 helps regulate the fusion of the viral and cellular membranes.[4] Thus, making the spike protein an essential component for determining the specificity and level of infection of the host.

With researchers demonstrating the evolutionary patterns of SARS-CoV-2 and speculating its origin, the investigation of the phylogenetic relationships of SARS-CoV-2 is a rapidly expanding field. The primary methods used in these investigations are genome sequencing, followed by sequence alignment and similarity analysis. Numerous hypotheses have been put out regarding the origin of SARS-CoV-2. Of these, some theories propose bats to be the origin of SARS-CoV-2,[5] while others have proposed pangolin to be the origin.[6]

SCAN QR CODE TO VIEW ONLINE

www.ajbls.com

DOI: 10.5530/ajbls.2022.11.104

Ji *et al.* also reported a study in which snakes served as an intermediary in the transmission of SARS-CoV-2.[7]

From Euler's formal beginnings in the 18th century, graph theory has significantly transformed. The size or range of these graphs has grown dramatically in response to the growing interaction of nodes or objects over time. Data extraction is becoming significantly more difficult and time-consuming due to this. To overcome this constraint, graph mining is utilized, in which useful information is retrieved from a number of graph networks utilizing a concept known as graph matching or graph similarities. In a sequence of graphs G_1, G_2, G_3...G_N, graph similarities are a function to compute the similarities among graphs which is defined as sim(G_1, G_2) ∈[0, 1] has value one if G_1 and G_2 have the same similarities; otherwise, zero.[8] Several mathematical studies have been modeled to capture the dynamics of COVID-19 phylogeny. In the current study, we have adopted a novel graph comparison approach to validate the evolutionary relationship of COVID-19.

Comparison of two graphs can be analyzed by different macro-level properties of networks like degree distributions,[9] maximum common sub-graph,[10] average path length number of spanning trees,[11] and local neighborhood of the nodes,[12] etc. The graph theory approach has also been applied to different biological networks where they have successfully established the relationship of the physicochemical properties of the amino acid residues as well as studied the evolutionary relationship of proteins using different centrality measures.[12] The evolutionary relationship among species was studied by Heymans and Singh, where each of the graphs represented enzyme interactions within a particular species. They computed the distances between some species and calculated the pair-wise similarity matrices, which were further validated by comparing them with the phylogenetic trees of those species.[13] Singh *et al.* have applied a noble graph theoretic approach to study the evolutionary relationship among fourteen different species of animals having a common neuropeptide sequence of Galanin. They applied different centrality measures and finally gave a conclusion about the relations among the amino acids, and the physicochemical properties of amino acids.[14] Das *et al.* suggested a new bipartite graph-based alignment-free sequence comparison approach based on graph theory in 2020.[15] They used this hypothesis to build evolutionary links between diverse types of organisms, such as viruses, mammals, and coronaviruses and established remarkably good and accurate correlations. Another study by Khan and Atangana looked at the mathematical modeling and dynamics of novel coronaviruses.[16]

According to the reviewed literature, there is a great demand for efficient methods for analyzing large quantities of genomic sequences in a short time, with the bulk of proposed models focusing primarily on conserved regions. When attempting to resolve deep phylogenies, more conserved sequences are likely to be more useful. Rapidly changing regions, on the other hand, have been demonstrated to be highly instructive in studying recent divergences, provided orthology remains certain. Also, no such evolutionary analysis for long, unequal lengths has been done in detail in the papers listed thus far. There is no comparison of graph theory results with those of a biologically determined phylogenetic tree in any of these investigations. One of the studies by Singh and co-workers showed a method whereas separate graph for each sequence eliminates the need for merging or union operations.[13] Previous scholars have also reached conclusions based solely on centrality scores. But, in this case, we have reached conclusions directly not just on centrality values, but also on correlation values. Furthermore, despite the availability of a significant number of proposed models, most of them do not consider non-conserved sequences in phylogenetic analysis.

Therefore, this study is an attempt to fill in the gaps indicated above, wherein we have used a novel graph theory approach to investigate the evolutionary lineage of SARS-CoV-2 utilizing the spike (S) protein in comparison to closely related viruses using non-conserved regions. According to our findings, this method can be used as part of a pipeline to perform phylogenetic analysis on protein sequences that share a common physico-chemical feature in adjacent amino acid residues.

MATERIALS AND METHODS

The graph theory model used to forecast the phylogeny considered species was created using a two-step technique. First, to construct the phylogenetic tree, the sequences of the seven coronaviruses known to mankind and two bat coronaviruses were collected and aligned and exploratory data analyses were done. And second, a novel graph theory technique was used to validate the results of the phylogenetic tree.

The proposed methodology consists of three main processes, as shown in Figure 1, starting with the data processing, where we describe how data were collected, organized and analyzed. Secondly, the phylogenetic tree modeling using the Mega X software. Moreover,

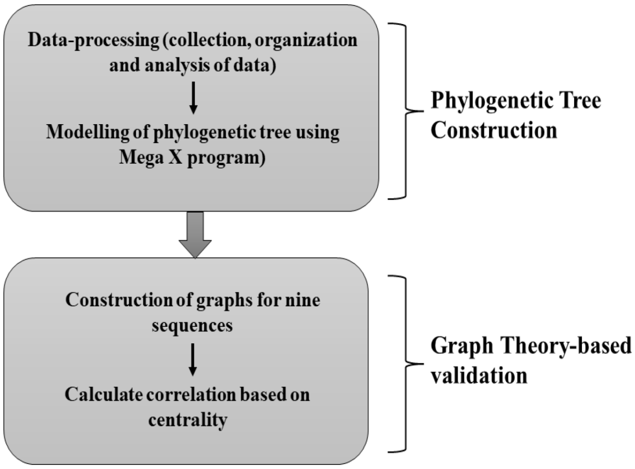
Data-processing (collection, organization and analysis of data)
Modelling of phylogenetic tree using Mega X program)
Phylogenetic Tree Construction
Construction of graphs for nine sequences
Calculate correlation based on centrality
Graph Theory-based validation

Figure 1: Graphical representation of the methodology followed.

finally, validate the results of the phylogenetic tree using various centrality measures.

Protein Sequence Alignment and Phylogenetic Tree Construction

Presently, there are seven coronaviruses (SARS-CoV, 2019-nCoV, MERS-CoV, HCoV-OC43, HCoV-229E, HCoV-NL63, and HCoV-HKU1) known to mankind, to say the least, of which many are known to be transmitted from bats.[17] Apart from the 7 human coronaviruses, 2 coronaviruses (RsSHC014 and RaTG13) originating from bat, have also been examined because it is believed that they are, respectively, the ancestors of SARS-CoV and SARS-CoV-2. All the sequences have been downloaded from NCBI(https://www.ncbi.nlm.nih.gov/) having accessions: P59594 (SARS-CoV), YP_009047204.1(MERS-CoV), QHD43416.1 (2019-nCoV), YP_009555241.1 (HCoV-OC43), YP_003767.1 (HCoV-NL63), NP_073551.1 (HCoV-229E), YP_173238.1 (HCoV-HKU1), AGZ48806.1 (RsSHC014), and QHR63300.1 (RaTG13). Therefore, in order to scrutinize the evolutionary relationship between SARS-CoV-2 and the bat CoV as compared to the other human CoVs, we performed a multiple sequence alignment (MSA) of the S-protein using the MUSCLE tool present in the Mega version X.[18]

For the phylogenetic tree construction, the maximum likelihood (ML) method has been used as the preferred statistical method. ML analysis has also been performed using the Mega version X. The most appropriate substitution model, WAG with Freqs. (+F) model has been used with rates = gamma distributed with invariant sites (G+I) unable to accept substitutions = 5 for ML. We used heuristic searches followed by fifty bootstrap iterations for the robustness of the ML tree.

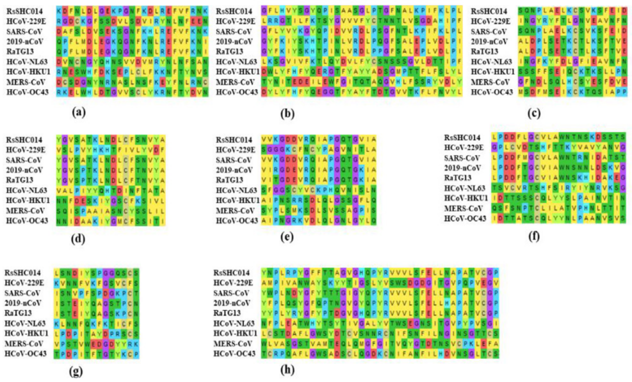

Figure 2: Aligned sequences for the spike proteins of nine viral sequences, of which 2019-nCoV (QHD43416.1), SARS-CoV (P59594), MERS-CoV (YP_009047204.1), HCoV-NL63 (YP_003767.1), HCoV-229E (NP_073551.1), HCoV-OC43 (YP_009555241.1), and HCoV-HKU1 (YP_173238.1) are the human coronaviruses and RaTG13 (QHR63300.1) and RsSHC014 (AGZ48806.1) are the two bat coronaviruses that are thought to be the ancestors of 2019-nCoV and SARS-CoV. (a), (b), (c), (d), (e), (f), (g), and (h) are the module graphs having sequence numbers (taking 2019-nCoV as reference) 301 – 325, 330 – 362, 428 – 445, 516 – 533, 537 – 555, 562 – 583, 677 – 690, and 724 – 762 respectively.

CoV Sequence-Fragment Selections for Graph Theory

All the nine viral sequences have been manually aligned, and a total of eight fragments with varying lengths have been considered for the graph theory study. Care has been taken to consider the fragments of the S-protein lying in and around the receptor-binding domain. Further, only those parts of the sequences have been considered which have non-identical amino acids in the pair-wise alignment (Figure 2).

Network Properties used in Graph Theory

A graph $G = (V, E)$ is a set of objects called node $V = \{v_1, v_2, v_3 \ldots\ldots v_n\}$ and edges $E = \{e_1, e_2, e_3 \ldots e_n\}$ such that each edge e_k is recognized with an unordered pair of vertices (v_i, v_j). $V(G)$ and $E(G)$ are denoted as vertex set and edge set of G.[28]

In a large network finding the most influential node among node numbers is a challenging task. A number of centrality measures like closeness centrality, degree centrality, eigenvector centrality etc. are used to find the most influential node. Here, we have applied degree centrality as well as degree distribution centrality in protein sequence networks. Some of the centrality measures are discussed below.

Degree centrality

Degree centrality $C_d(n_i)$ is the most straightforward centrality measure where the centrality value of a particular node is calculated as the number of vertices to which the vertex '*u*' is directly connected. This centrality measure reflects that more the connections, more the important vertex are which may be different in real-world networks. It is defined as,

$$C_d(n_i) = d(n_i) \qquad \text{[Eq.1]}$$

Degree distribution

The degree distribution is an important feature of complex networks and can be used for feature extraction techniques. Although degree distribution gives a little bit of formation about the network, one can still easily get information about the network's structure. In most cases, similar types of networks have the same degree of values for nodes. The degree distribution $P(k)$ is defined by,

$$P(k) = \frac{n_k}{n} \qquad \text{[Eq.2]}$$

Where n is the total number of nodes in a network and of them having degree.

In our study, we have considered nine S-protein sequences as shown in section 2.2. From the concept of amino acids networks studies,[19] we have defined two networks viz., module graph and sequence graph, respectively. In amino acid networks, the vertex is the twenty amino acids, and the edge is established between two amino acids if they are adjacent and have at least one common property. Here, we have considered eight properties.

The module graph is a graph of a single S-protein sequence where the vertex set is the number of amino acids present in that specific sequence and two nodes are said to be adjacent in G if they are consecutive elements in the sequence and also have at least one common property.[12] For our study, we have considered eight module graphs for a single protein sequence and got eight module graphs for the remaining sequences. To calculate the adjacency matrix, we have used R programming and used aMat Reader, as well as CytoNCA package of Cytoscape to analyze the network.

The sequence graph is a graph in which all the module graphs are merged for a particular sequence. We have used the 'Network Merge' application embedded in Cytoscape 3.3 to merge the entire module for a single sequence. Networks will be merged based on the attribute values of nodes/edges. Finally, we can get nine sequence graphs for nine S-protein sequences.

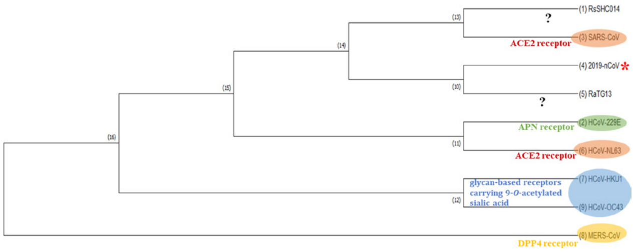

Figure 3: The phylogenetic tree for the 2019-nCoV along with the known entry receptors. The numbers present along the branches represent the bootstrap values percentage out of 50 bootstraps resembling. The star shows the 2019-nCoV, and the question marks show that the entry receptors for the viruses remain unknown.

RESULTS

Phylogenetic Tree

Using graph theory, an ML tree was used for the phylogenetic analysis of seven human coronaviruses and two bat coronaviruses in order to discover the evolutionary connection of SARS-CoV-2. SARS-CoV-2 (2019-nCoV) and SARS-CoV were found in divergent locations, according to the analysis, which is consistent with the findings of the phylogenetic ML tree created by Feng and colleagues (Figure 3).[20] Some other relevant works also reported the divergent location of 2019-nCoV as compared to SARS-CoV and MERS-CoV.[21]

The distinct phylogenetic distances observed on clades of 2019-nCoV and SARS-CoV in the phylogenetic tree explicitly displayed evolutionary links between coronaviruses. It is remarkable that in the phylogenetic ML tree, the bat coronavirus RaTG13 showed a closer genetic distance to SARS-CoV-2. This makes sense and follows logically from evolutionary development, and it is well supported by earlier research.[20] In addition to showing SARS-CoV-2 and RaTG13 as being close, the phylogenetic ML tree also shows MERS-CoV as being the most remote relative of SARS-CoV-2.This result can be justified by the fact that 2019-nCoV and MERS-CoV have completely two different entry receptors, namely ACE2 and DPP4 receptors, respectively. The two human coronaviruses, HCoV-OC43 and HCoV-HKU1, are grouped together in the same clade because they share the same entry receptor, the glycan-based receptors containing 9-O-acetylated sialic acid, which supports the simulated phylogenetic tree.[22] Through phylogenetic analysis concerning the non-conserved regions present in the binding region, we found that both 2019-nCov and SARS-CoV are closer to bat coronaviruses RaTG13 and RsSHC014 respectively, which has also been reported by Lu *et al*.[23] Where, they state that at whole-genome level, 2019-nCoV was closer to bat-coronavirus, however, when modeling the tree using the receptor-binding domain concerning the conserved regions, they found that 2019-nCoV was closer to that of SARS-CoV.

Establishing Evolutionary Relationship using Graph Theory

To illustrate our method, we have taken sequence number 4, i.e., SARS-CoV-2 S-protein sequence having NCBI accession QHD43416.1 with sequences starting from 301 to 762 randomly (region lying in and around the receptor-binding domain of the S-protein), for each sequence. As discussed in section 2.2, we first divided the sequences into numbers of modules based on manual alignment. A total of eight modules with varying lengths and having non-identical amino acids in the pair-wise alignment have been considered. Previously, Heymans and Singh used graph node similarity to discover structural relationships between graphs. On the other hand, Singh *et al.* investigated the evolutionary relationships of fourteen species using the centrality values of a peptide called Galanin, which is composed of only 29 amino acids. The structural influence on determining the link between sequences is not adequately reflected in both investigations. Furthermore, because the evolutionary distance is based on network structure divergence, we attempted to investigate the evolution connection using the individual network structure acquired from each sequence.

Now, when the first module (M1), i.e., from position 301 to 325, is considered, a protein sequence of twenty-five amino acids is obtained where some of the residues are in repetition. For our study, we have considered the amino acids and their corresponding neighbours along with the M1, i.e., from position 301 to 325 and have checked the properties matching criteria. If these two criteria are fulfilled, then an edge is assigned to the residues. In the following Figure 4, we have drawn the M1of the fourth sequence, i.e., for SARS-CoV-2.

Similarly, eight module graphs for nine sequences have been constructed. Now, to combine this entire module graph, we have used the "merge" extension package of Cytoscape where networks will be merged based on attribute values of nodes/edges. The final sequence graph of SARS-CoV-2 is shown in Figure 5 below.

Accordingly, nine sequence graphs have been obtained having NCBI accession AGZ48806.1, NP_073551.1, sp|P59594.1, QHD43416.1, QHR63300.2, YP_173238.1, YP_009047204.1, YP_003767.1, YP_009555241.1 forRsSHC014, HCoV-229E, 2019-nCoV, SARS-CoV, RaTG13, HCoV-NL63, HCoV-HKU1, MERS-CoV, and HCoV-OC43, respectively.

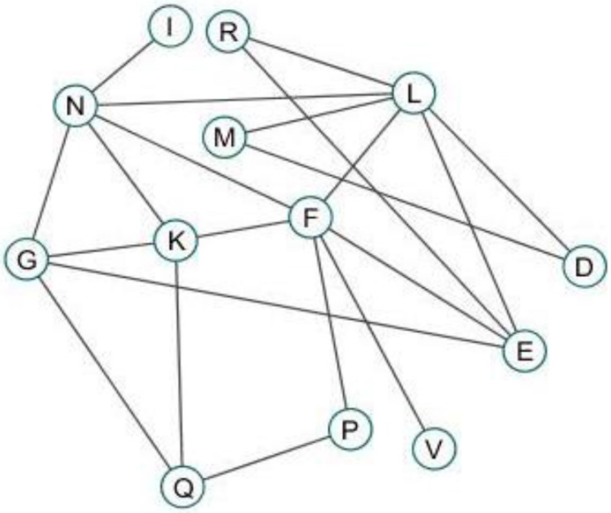

Figure 4: 1st Module graph of the sequence of SARS-CoV-2.

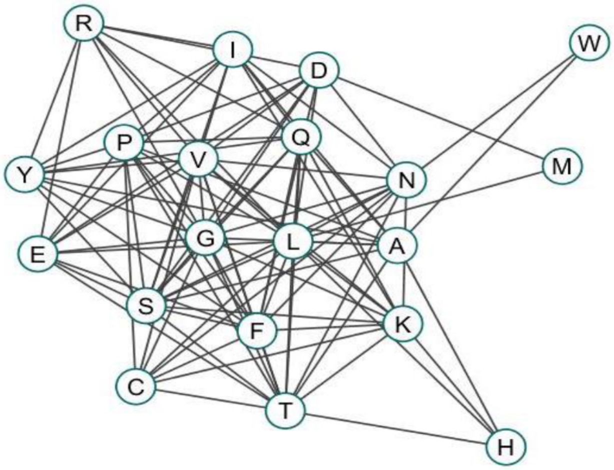

Figure 5: Sequence graph of SARS-CoV-2.

Evolutionary study through Centrality measure

Networks built using the same evolutionary method are structurally more similar to one another.[24] As a result, we can infer that networks with comparable evolutionary paths should have greater similarities in their structural makeup than networks with different evolutionary paths. Centrality measures are used to calculate the node importance in a graph. As there are many centrality measures, we have considered only that centrality measure that gives results from global perspectives. Our main objective is to find out the similarities among nine sequence graphs to study the evolutionary importance of the sequences and match the results with the phylogenetic ML tree created using the Mega version X software package, as shown in section 3.1. There are numerous methods for detecting similarities among numbers of graphs.[25] In this case, we've employed feature extraction techniques that assume that similar graphs will likely share particular characteristics, such diameter, degree distribution and eigenvalues.[26] To calculate the degree distribution among our eight modules graph, we have applied equation number (2). Degree centrality and degree distribution values of nine sequence graphs are shown in Table 1 and Table 2 respectively. All these centrality measures are calculated for undirected networks in a reasonable time. To calculate these measures, we have used the centiserve, and igraph package.

Next, we applied Pearson's correlation to find the similarities among the nine sequence graphs. Pearson's correlation values range from -1 to +1. The lower the values, the weaker the relationship and the greater the positive value more vital the relationship between the two variables. Here, in Table 3, when the correlation coefficients were analyzed, it was found that the results suggested a strong relationship among the sequences and were in agreement with those obtained in the phylogenetic ML tree performed above in section 3.

Table 3: Correlation coefficients table of the From Table 3, it has been observed that RaTG13 is highly correlated with 2019-nCoV (SARS-CoV-2). This is consistent with the findings of Li *et al*., 2020[20] and Xu *et al*., 2020[27] and also coincides with the results shown in section 3.1. This result showed the maximum correlation between RsSHC014 and SARS-CoV, HCoV-NL63 and HCoV-229E and HCoV-HKU1 and HCoV-OC43 which is similar to the results found in the evolutionary ML tree showed in section 3.1. Results also revealed that SARS-CoV,2019-nCoV, and MERS-CoV have a minimum degree of correlation. Literature showed that 2019-nCoV, MERS-CoV and SARS-CoV belong to a different lineage (Xu *et al*., 2020),[27] but they share the same family, Coronaviridae. Therefore, it validates the predictive power of the graph theory approach of the study.

CONCLUSION

Here, in this manuscript, we have attempted a graph theoretic approach to analyze the evolutionary journey of SARS-CoV-2. Our investigation is limited to two bat viruses presumed to be the ancestors of SARS-CoV-2 and seven known human coronaviruses. A phylogenetic tree using the ML method was built for the nine different coronaviruses. It was found that the closest phylogeny to 2019-nCoV was shown by the bat coronavirus RaTG13, whereas SARS-CoV showed the closest phylogeny to RsSHC014. MERS-CoV was found to be most distantly placed. After the tree was built, nine S-protein sequences were considered to create nine sequence graphs in the next step. Further analysis consisting of the different

Table 1: Sequence-wise Degree centrality values of 20 amino acids.

AA	RsSHC014	HCoV-229E	SARS-CoV	2019-nCoV	RaTG13	HCoV-NL63	HCoV-HKU1	MERS-CoV	HCoV-OC43
K	13	13	13	13	12	16	6	13	7
D	8	4	8	8	6	6	6	5	5
F	12	9	14	12	13	7	9	12	9
N	7	5	7	10	11	5	7	11	6
L	16	14	13	14	14	10	10	12	12
G	13	12	14	15	14	13	12	17	13
E	14	11	14	11	12	11	12	13	11
P	13	14	14	13	12	14	12	10	12
R	9	7	10	10	12	6	11	10	12
V	11	8	12	11	13	9	7	7	10
H	9	11	11	13	15	13	11	14	13
Y	10	11	11	10	12	7	6	11	10
S	9	9	10	10	12	12	14	9	11
Q	9	1	8	11	9	6	5	10	9
I	14	10	13	14	12	14	14	17	13
A	9	13	12	9	10	15	10	13	11
T	4	7	2	4	5	8	2	5	4
C	6	8	8	8	8	7	9	9	10
M	2	3	4	2	2	3	5	4	4
W	2	2	2	2	2	2	2	6	4

Table 2: Degree distribution of nine sequence graphs.

AA	RsSHC014	HCoV-229E	SARS-CoV	2019-nCoV	RaTG13	HCoV-NL63	HCoV-HKU1	MERS-CoV	HCoV-OC43
V	0.65	0.65	0.65	0.65	0.6	0.8	0.3	0.65	0.35
R	0.4	0.2	0.4	0.4	0.3	0.3	0.3	0.25	0.25
P	0.6	0.45	0.7	0.6	0.65	0.35	0.45	0.6	0.45
E	0.35	0.25	0.35	0.5	0.55	0.25	0.35	0.55	0.3
G	0.75	0.7	0.65	0.7	0.7	0.5	0.5	0.6	0.6
L	0.65	0.6	0.7	0.75	0.7	0.65	0.6	0.85	0.65
N	0.7	0.55	0.7	0.55	0.6	0.55	0.6	0.65	0.55
F	0.6	0.7	0.7	0.65	0.6	0.7	0.6	0.5	0.6
D	0.45	0.35	0.5	0.5	0.6	0.3	0.55	0.5	0.6
K	0.55	0.4	0.6	0.55	0.65	0.45	0.35	0.35	0.5
T	0.45	0.55	0.55	0.65	0.75	0.65	0.55	0.7	0.65
A	0.5	0.55	0.55	0.5	0.6	0.35	0.3	0.55	0.5
I	0.45	0.45	0.5	0.5	0.6	0.6	0.7	0.45	0.55
Q	0.45	0.05	0.4	0.55	0.45	0.3	0.25	0.5	0.45
S	0.7	0.5	0.65	0.7	0.6	0.7	0.7	0.85	0.65
Y	0.45	0.65	0.6	0.45	0.5	0.75	0.5	0.65	0.55
H	0.2	0.35	0.1	0.2	0.25	0.4	0.1	0.25	0.2
C	0.3	0.4	0.4	0.4	0.4	0.35	0.45	0.45	0.5
W	0.1	0.15	0.2	0.1	0.1	0.15	0.25	0.2	0.2
M	0	0.1	0.1	0.1	0.1	0.1	0.1	0.3	0.2

Table 3: Correlation coefficients table of the Degree distribution of nine sequence graphs.

	RsSHC014	HCoV-229E	SARS-CoV	2019-nCoV	RaTG13	HCoV-NL63	HCoV-HKU1	MERS-CoV	HCoV-OC43
RsSHC014	1	0.7239	0.9341	0.9087	0.8461	0.6809	0.6565	0.7481	0.7282
HCoV-229E	0.7239	1	0.7773	0.6707	0.7126	0.8320	0.6090	0.6434	0.6713
SARS-CoV	0.9341	0.7773	1	0.8953	0.8682	0.7160	0.7450	0.7543	0.7976
2019-nCoV	0.9087	0.6707	0.8953	1	0.9295	0.6902	0.6931	0.8194	0.7991
RaTG13	0.8461	0.7126	0.8682	0.9295	1	0.6406	0.7045	0.7623	0.8298
HCoV-NL63	0.6809	0.8320	0.7160	0.6902	0.6406	1	0.6320	0.6820	0.6351
HCoV-HKU1	0.6565	0.6090	0.7450	0.6931	0.7045	0.6320	1	0.6737	0.8628
MERS-CoV	0.7481	0.6434	0.7543	0.8194	0.7623	0.6820	0.6737	1	0.7531
HCoV-OC43	0.7282	0.6713	0.7976	0.7991	0.8298	0.6351	0.8628	0.7531	1

centrality values revealed anaccurate match with the results of the phylogenetic ML tree.

The originality of the work lies in the fact that this method uses rapidly changing regions, instead of the conserved regions, for building the phylogenetic tree using a novel graph theory method. This work can be used as a pipeline to accomplish evolutionary studies of protein sequences having adjacent residues with at least one common property. This method can bring new possibilities for studying protein sequences. Nevertheless, there is still scope for a vast amount of work to improve the current modified method, which will improve the theoretical graph methods and find its use in different protein identification problems. The graph-centric approach presented here can be further extended for other network parameters.

ACKNOWLEDGEMENT

We thank to chairperson, Centre for Computer Scienceand Applications for providing constant guidance. SR thanks Centre for Biotechnology and Bioinformatics, Dibrugarh University for providing the lab facilities to perform the phylogenetic analysis.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

SARS: Severe acute respiratory syndrome; **ML:** Maximum Likelihood; **NCBI:** National Center for Biotechnology Information; **MSA:** Multiple sequence alignment.

SUMMARY

In this manuscript we formulate a unique graph theory approach to determine the evolutionary relationships between two bat coronaviruses believed to be the relatives of 2019-nCoV and SARS-CoV and seven known human coronaviruses. It was found that the closest phylogeny to 2019-nCoV was shown by the bat coronavirus RaTG13, whereas SARS-CoV showed the closest phylogeny to RsSHC014.

REFERENCES

1. HIT: Linking herbal active ingredients to targets | Nucleic Acids Research | Oxford Academic [cited Jun 22, 2022]. Available from: https://academic.oup.com/nar/article/39/suppl_1/D1055/2507403.
2. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. Mar 2019;17(3):Art no. 3. doi: 10.1038/s41579-018-0118-9, PMID 30531947.
3. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3. doi: 10.1038/S41586-020-2012-7, PMID 32015507.
4. Tortorici MA, Veesler D. Structural insights into coronavirus entry. Adv Virus Res. 2019;105:93-116. doi: 10.1016/bs.aivir.2019.08.002, PMID 31522710.
5. Li X, Song Y, Wong G, Cui J. Bat origin of a new human coronavirus: There and back again. Sci China Life Sci. Mar 2020;63(3):461-2. doi: 10.1007/s11427-020-1645-7, PMID 32048160.
6. Lam TTY, Jia N, Zhang YW, Shum MH, Jiang JF, Zhu HC, *et al*. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020;583(7815):282-5. doi: 10.1038/s41586-020-2169-0, PMID 32218527.
7. Cross-species transmission of the newly identified coronavirus2019-nCoV – Ji −2020 –Journal of Medical Virology – Wiley Online Library [cited Jun 22, 2022]. Available from: https://onlinelibrary.wiley.com/doi/10.1002/jmv.25682.
8. Labriji A, Charkaoui S, Abdelbaki I, Namir A, Labriji EH. Similarity measure of graphs. Int J Recent Contrib Eng Sci IT. 2017;5(2):Art no. 2. doi: 10.3991/ijes.v5i2.7251.
9. Antunes N, Bhamidi S, Guo T, Pipiras V, Wang B. Sampling based estimation of in-degree distribution for directed complex networks. J Comput Graph Stat. 2021;30(4):863-76. doi: 10.1080/10618600.2021.1873143.
10. Huang X, Lai J, Jennings SF. Maximum common subgraph: Some upper bound and lower bound results. BMC Bioinformatics. 2006;7;Suppl 4, no. SUPPL.4:1-9. doi: 10.1186/1471-2105-7-S4-S6/FIGURES/1.
11. Liu JB, Daoud SN. Number of spanning trees in the sequence of some graphs. Complexity. 2019;2019:1-22. doi: 10.1155/2019/4271783.
12. Zager LA, Verghese GC. Graph similarity scoring and matching. Appl Math Lett. Jan 2008;21(1):86-94. doi: 10.1016/J.AML.2007.01.006.
13. Heymans M, Singh AK. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics. 2003;19;Suppl 1:i138-46. doi: 10.1093/BIOINFORMATICS/BTG1018, PMID 12855450.
14. Singh SG, ACK. Analysing amino acids in Galanin graph theoretical approach. Int J Recent Innov Trends Comput Commun. 2017;5(6):342-346-342-346. doi: 10.17762/IJRITCC.V5I6.774.
15. Das S, Das A, Bhattacharya DK, Tibarewala DN. A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets. Genomics. Nov 2020;112(6):4701-14. doi: 10.1016/j.ygeno.2020.08.023, PMID 32827671.
16. Khan MA, Atangana A. Modeling the dynamics of novel coronavirus (2019-nCov) with fractional derivative. Alex Eng J. 2020;59(4):2379-89. doi: 10.1016/j.aej.2020.02.033.
17. Armijos-Jaramillo V, Yeager J, Muslin C, Perez-Castillo Y. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. Evol Appl. 2020;13(9):2168-78. doi: 10.1111/eva.12980, PMID 32837536.
18. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547-9. doi: 10.1093/MOLBEV/MSY096, PMID 29722887.
19. Yan W, Yu C, Chen J, Zhou J, Shen B. ANCA: Aweb server for amino acid networks construction and analysis. Front Mol Biosci. 2020 [online];7:582702. doi: 10.3389/fmolb.2020.582702, PMID 33330622.
20. Li T, *et al*. Phylogenetic super tree reveals detailed evolution of SARS-CoV-2 [scirep]. Sci Rep. 2020;10(1), Dec. 2020:22366. doi: 10.1038/s41598-020-79484-8, PMID 33353955.
21. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol. Apr 2020;79:104212. doi: 10.1016/j.meegid.2020.104212, PMID 32004758.
22. Liu DX, Liang JQ, Fung TS. Human Coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae). EncyclVirol. 2021:428-40. doi: 10.1016/B978-0-12-809633-8.21501-X.
23. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. Lancet. Feb 2020;395(10224):565-74. doi: 10.1016/S0140-6736(20)30251-8, PMID 32007145.
24. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. J Proteome Res. Apr 2020;19(4):1351-60. doi: 10.1021/ACS.JPROTEOME.0C00129, PMID 32200634.
25. Koutra D, Parikh A, Ramdas A, Xiang J. Algorithms for graph similarity and subgraph matching; 2011.
26. Watts DJ, Bak P. Small worlds: The dynamics of networks between order and randomness. Physics Today. 2000;53(11):54-5. doi: 10.1063/1.1333299.
27. Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, *et al*. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci. 2020;63(3):457-60. doi: 10.1007/S11427-020-1637-5, PMID 32009228.
28. Hazarika P, Bora PK, Baruah AK, Bora P. Study of Codon Degeneracy Based on Similarity Measure. Asian J Biol Life Sci. 2022;11(2):1-11.DOI: 10.5530/ajbls.2022.11.10.

Cite this article: Bora PK, Rajkhowa S, Baruah AK, Bora P. Deciphering the Evolutionary Relationship of SARS-CoV-2: A Graph Theory Approach. Asian J Biol Life Sci. 2022;11(3):776-84.

APPENDIX A: Protein sequence database used in Chapter VII downloaded from NCBI *https://www.ncbi.nlm.nih.gov/.*

Sl. No.	Module	NCBI accession number	Non-identical amino acid residues	Position in the alignment
1	M1	AGZ48806.1	KDFNLDLGEKPGNFKDLREFVFRNK	From position 301 to 325 in alignment file
2		NP_073551.1	RGDCKGFSSDVLSDVIRYNLNFEEN	
3		sp|P59594.1	DAFSLDVSEKSGNFKHLREFVFKNK	
4		QHD43416.1	QPFLMDLEGKQGNFKNLREFVFKNI	
5		QHR63300.2	QPFLMDLEGKQGNFKNLREFVFKNI	
6		YP_003767.1	DVNCNGYQHNSVVDVMRYNLNFSAN	
7		YP_173238.1	RNESWHFDKSEPLCLFKKNFTYNVS	
8		YP_009047204.1	DCSDGNYNRNASLNSFKEYFNLRNC	
9		YP_009555241.1	RKELWHLDTGVVSCLYKRNFTYDVN	
1	M2	AGZ48806.1	GFLHVYSGYQPISAASGLPTGFNALKPIFKLPL	From position 330 to 362 in alignment file
2		NP_073551.1	LRRGTILFKTSYGVVVFYCTNNTLVSGDAHIPF	
3		sp|P59594.1	GFLYVYKGYQPIDVVRDLPSGFNTLKPIFKLPL	
4		QHD43416.1	GYFKIYSKHTPINLVRDLPQGFSALEPLVDLPI	
5		QHR63300.2	GYFKIYSKHTPINLVRDLPPGFSALEPLVDLPI	
6		YP_003767.1	LKSGVIVFKTLQYDVLFYCSNSSSGVLDTTIPF	
7		YP_173238.1	DWLYFHFYQERGTFYAYYADSGMPTTFLFSLYL	
8		YP_009047204.1	TYNITEDEILEWFGITQTAQGVHLFSSRYVDLY	
9		YP_009555241.1	DYLYFHFYQEGGTFYAYFTDTGVVTKFLFNVYL	
1	M3	AGZ48806.1	SQNPLAELKCSVKSFEID	From position 428 to 445 in alignment file
2		NP_073551.1	INGYRYFTLGNVEAVNFN	
3		sp|P59594.1	SQNPLAELKCSVKSFEID	
4		QHD43416.1	ALDPLSETKCTLKSFTVE	
5		QHR63300.2	ALDPLSETKCTLKSFTVE	
6		YP_003767.1	INGFKYFDLGFIEAVNFN	
7		YP_173238.1	SSSFFSEIQCKTKSLLPN	
8		YP_009047204.1	GFNDLSQLHCSYESFDVE	
9		YP_009555241.1	MSDFMSEIKCKTQSIAPP	
1	M4	AGZ48806.1	YGVSATKLNDLCFSNVYA	From position 516 to 533 in alignment file
2		NP_073551.1	VSLPVYHKHTFIVLYVDF	
3		sp|P59594.1	YGVSATKLNDLCFSNVYA	
4		QHD43416.1	YGVSPTKLNDLCFTNVYA	
5		QHR63300.2	YGVSPTKLNDLCFTNVYA	
6		YP_003767.1	VALPIYYQHTDINFTATA	
7		YP_173238.1	NNFDESKIYGSCFKSIVL	
8		YP_009047204.1	SQISPAAIASNCYSSLIL	
9		YP_009555241.1	NNIDAAKIYGMCFSSITI	
1	M5	AGZ48806.1	VVKGDDVRQIAPGQTGVIA	From position 537 to 555 in alignment file
2		NP_073551.1	SGGGKCFNCYPAGVNITLA	
3		sp|P59594.1	VVKGDDVRQIAPGQTGVIA	
4		QHD43416.1	VIRGDEVRQIAPGQTGKIA	
5		QHR63300.2	VITGDEVRQIAPGQTGKIA	
6		YP_003767.1	SFGGSCYVCKPHQVNISLN	
7		YP_173238.1	AIPNSRRSDLQLGSSGFLQ	
8		YP_009047204.1	SYPLSMKSDLSVSSAGPIS	
9		YP_009555241.1	AIPNGRKVDLQLGNLGYLQ	
1	M6	AGZ48806.1	LPDDFLGCVLAWNTNSKDSSTS	From position 562 to 583 in alignment file
2		NP_073551.1	GPLCVDTSHFTTKYVAVYANVG	
3		sp|P59594.1	LPDDFMGCVLAWNTRNIDATST	
4		QHD43416.1	LPDDFTGCVIAWNSNNLDSKVG	
5		QHR63300.2	LPDDFTGCVIAWNSKHIDAKEG	
6		YP_003767.1	TSVCVRTSHFSIRYIYNRVKSG	
7		YP_173238.1	IDTTSSSCQLYYSLPAINVTIN	
8		YP_009047204.1	QSFSNPTCLILATVPHNLTTIT	
9		YP_009555241.1	IDTTATSCQLYYNLPAANVSVS	
1	M8	AGZ48806.1	YNPLRPYGFFTTAGVGHQPYRVVVLSFELLNAPATVCGP	From position 724 to 762 in alignment file
2		NP_073551.1	AMPIVANWAYSKYYTIGSLYVSWSDGDGITGVPQPVEGV	
3		sp|P59594.1	YWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATVCGP	
4		QHD43416.1	YFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGP	
5		QHR63300.2	YYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPATVCGP	
6		YP_003767.1	NFPLEATWHYTSYTIVGALYVTWSEGNSITGVPYPVSGI	
7		YP_173238.1	LCSTDAFLGWSYDTCVSNNRCNIFSNFILNGINSGTTCS	
8		YP_009047204.1	WLVASGSTVAMTEQLQMGFGITVQYGTDTNSVCPKLEFA	
9		YP_009555241.1	TCRPQAFLGWSADSCLQGDKCNIFANFILHDVNSGLTCS	