OMICS RESEARCH SYMPOSIUM

SEQUENTIAL ANALYSIS OF *SARS-CoV-2* VARIANTS

Department of Biotechnology,
SRM Institute of Science and Technology, Chennai

Anushka Jain, Priya Swaminathan

Introduction

Coronaviruses are large, enveloped, positive-stranded RNA viruses responsible for infecting a wide variety of mammalian and avian species. These viruses contain spike-like projections of glycoproteins on their surface, which appear like a crown under the electron microscope; hence, they are referred to as coronaviruses. The coronavirus genome encodes several structural and nonstructural proteins. The structural proteins are responsible for host infection, membrane fusion, viral assembly, morphogenesis, and release of virus particles, among other functions, and the nonstructural proteins (nsps) facilitate viral replication and transcription. The membrane (M), the envelope (E), and the spike protein (S) make up the structural proteins and are associated with the envelope. Among these structural proteins, the trimeric S proteins protrude from the virus envelope and are the key machinery that facilitates virus entry into the host cell.

Spike proteins
The S proteins are clove-shaped, type-I transmembrane proteins and have 3 segments: a large ectodomain, a single-pass transmembrane, and an intracellular tail. The ectodomain of S proteins consist of the S1 subunit, containing a receptor-binding domain (RBD), and the membrane-fusion subunit (S2). The host-cell receptor recognition by the RBDs on S proteins is the initial step of viral infection, and the binding interactions between the coronavirus spike and its receptor is one of the most critical factors for host range and cross-species transmission. Human coronaviruses recognize a variety of host receptors; specifically, HCoV-229E recognizes human aminopeptidase N (hAPN), MERS-CoV binds dipeptidyl peptidase-4 (DPP4), HCoV-OC43 and HCoV-HKU1 bind certain types of O-acetylated sialic acid, and HCoV-NL63 and SARS-CoV recognize angiotensin-converting enzyme 2 (ACE2). Recent structures, along with functional studies, have suggested that the SARS-CoV-2 S proteins utilize ACE2 and Transmembrane Serine Protease 2 (TMPRSS2) for host-cell entry, which are very similar to the mechanisms exploited by SARS-CoV. See the "Structure, function, antigenicity, and hACE2 receptor recognition by the SARS-CoV-2 S glycoprotein" section of this review for detailed information on the mechanism of coronavirus cell entry mediated by the viral S glycoproteins. The S proteins, common among all coronaviruses, are a major target for eliciting antibodies; therefore, structural and molecular details of S protein and its interactions with cognate receptors would be vital in developing vaccines and antiviral drugs against SARS-CoV-2.

In December 2019, patients with severe pneumonia cases of unknown cause were reported in Wuhan, China, and a novel coronavirus strain was detected from the lower respiratory tract of 4 patients. Viruses were isolated from these clinical samples, and their genomes were analyzed by deep sequencing. Phylogenetic analysis of 2019-novel coronavirus (2019-nCoV) genomes and other coronaviruses were used to establish the evolutionary history and infection sources. Interestingly, this indicated that 2019-nCoV (GenBank: MN908947.3) shares about 96% nucleotide sequence identity to bat coronavirus RaTG13 (GenBank: MN996532.1), with 79.5% and 55% identity to SARS-CoV B01 (GenBank: AY278488.2) and MERS-CoV HCoV-EMC (GenBank: MH454272.1), respectively, and belongs to the same family of viruses that caused SARS and MERS. This suggests that bats are possibly the hosts of 2019-nCoV origin, and it might have been transmitted either directly from bats or through an unknown intermediate host to infect humans. Despite high sequence similarities, a few notable and conserved variations arose in 2019-nCoV genomes that were not previously seen in betacoronaviruses. These notable features, which establish this virus as different from SARS-CoV and SARS-like coronaviruses, are (i) multiple mutations in the RBDs of S protein that may interact with ACE2 receptor, (ii) a polybasic furin-like protease site (RRAR/S) at the boundary of S1/S2 subunits rather than the single arginine observed in SARS-CoV, and (iii) the addition of 3 predicted O-linked glycans flanking the protease site. Of note, a furin-like protease site is a signature of several highly pathogenic avian influenza viruses and pathogenic Newcastle disease virus..

Functions of the S protein
The S protein on the surface of the virus is a key factor involved in infection. It is a trimeric class I TM glycoprotein responsible for viral entry, and it is present in all kinds of HCoVs, as well as in other viruses such as HIV (HIV glycoprotein 160, Env), influenza virus (influenza hemagglutinin, HA), paramyxovirus (paramyxovirus F), and Ebola (Ebola virus glycoprotein). Similar to other coronaviruses, the S protein of SARS-CoV-2 mediates receptor recognition, cell attachment, and fusion during viral infection. The trimer of the S protein located on the surface of the viral envelope is the basic unit by which the S protein binds to the receptor. The S1 domain contains the RBD, which is mainly responsible for binding of the virus to the receptor, while the S2 domain mainly contains the HR domain, including HR1 and HR2, which is closely related to virus fusion.

Membrane proteins
Membrane protein (M) is one of the important functional components that plays a significant role in maintaining virion size and shape. It assists to assemble all other structural proteins including spike (S), envelope (E), and nucleocapsid (N) and participates in the budding process. Coronaviruses form virus-like particle (VLP) via the interaction of M and E or M and N proteins, and the collective manifestation of M, N and E is mandatory for well-organized VLP production as well as its trafficking and release. In addition, M-S proteins' interaction assist incorporation of S protein into virion. The M protein also collaborates with the S protein during the cell attachment and entry and it seems that these crucial interaction may facilitate viral transmission. Moreover, viral M protein, like other viral proteins, exhibits self-association as well as interaction with other accessory and non-structural proteins. These protein-protein interactions may play a significant role in viral structural protein processing, modification, and trafficking for viral particle assembly and egress. Thus, the critical network of SARS-CoV-2 M protein with its intra-viral proteins shapes the basis of targeting M protein as a target for structure-based drug design.

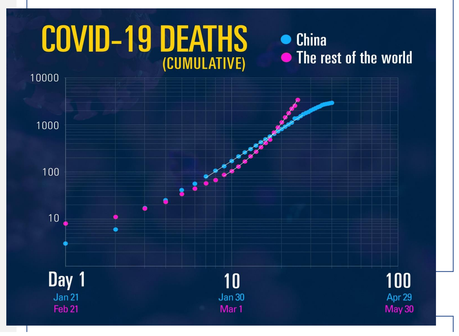
Nucleocapsid proteins
The N protein mediates ribonucleoprotein (RNP) complex formation via two key steps: packaging of the viral RNA genome and self-assembly of oligomerizations. Studies on coronavirus N-CTD suggest that the multiple packing modes of N-CTD dimers probably lead to the formation of rigid helically symmetric nucleocapsids, an unusual feature that is supported by various biochemical assays, including the disulfide trapping technique. Currently, the SARS-CoV N-CTD domain self-association has been widely studied for viral RNP assembly. However, the role of N-CTD in the self-association of SARS-CoV-2 remains unclear. Our structural data suggest that SARS-CoV-2 N-CTD possesses conserved dimerization mechanisms via multiple hydrophilic and hydrophobic interactions, similar to the CTD of other coronavirus nucleocapsid proteins. Intriguingly, the higher-order self-association of SARS-CoV-2 N-CTD seems different from that of SARS-CoV N-CTD in our studies. Previous studies showed that SARS-CoV N-CTD packs into octamers and forms a twin helix in the crystal packing, however, SARS-CoV-2 N-CTD packs into a cylindrical shape in the crystal packing. To further verify these observations, in vitro disulfide trapping assays combined with size-exclusion chromatography were performed to illustrate the status of SARS-CoV-2 N-CTD in solution. Our data suggest that the observed potential self-interactions via the β5–β6 loop and α1-helix regions in the crystal actually exist in solution, which may serve as the first step of the RNP assembly process.

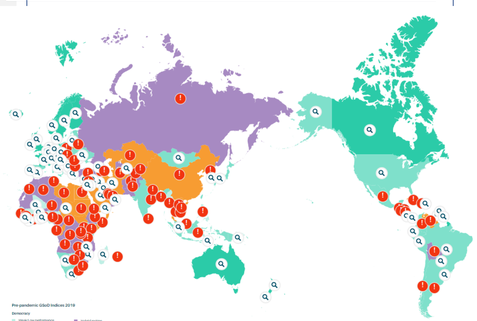
Previous studies suggest that the coronavirus nucleocapsid contains multiple RNA binding sites, including the NTD, CTD, and C-terminal IDR regions. Our previous work demonstrated that the N-terminal domain of the nucleocapsid is capable of binding to viral single-stranded 32-mer RNA. Our structural data suggest that SARS-CoV-2 N-CTD contains a positively charged channel similar to MERS-CoV N-CTD and SARS-CoV N-CTD. These surface electrostatic potential characteristics are conserved among the highly pathogenic viral nucleocapsid proteins. These positively charged channels in the α-helix-rich side are considered as potential RNA binding sites in SARS-CoV-2 N-CTD.

This project aims to understand a little more about these proteins across the different organisms like Bat, Human, and understand how they have evolved over a long period of time to become the dangerous viral structures that are claiming millions of lives across the world. The proteins taken into consideration include membrane proteins, nucleocapsid proteins and spike proteins. These are the most essential components of the virus. They play a direct role in the growth and spread of the virus by controlling its activities. So, the structures of these proteins are explained above and a multiple sequence alignment is conducted.

Diagram descriptions:
Figure 1: Scatterplot graph to indicate the rise in death cases of COVID-19 since January 2020. The graph shows a comparison between the situations of China and the rest of the world over the same time period by explaining in two different color codes.

Figure 2: A diagram that explains the impact of COVID across the world. The orange and purple spaces explain that the virus was growing exponentially on these lands. In countries indicated in green, the virus was growing at a regular phase and claiming many lives.





Methods

Multiple sequence alignment
Multiple sequence alignment (MSA) methods refer to a series of algorithmic solution for the alignment of evolutionarily related sequences, while taking into account evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions. These methods can be applied to DNA, RNA or protein sequences. A recent study in Nature reveals MSA to be one of the most widely used modeling methods in biology, with the publication describing ClustalW pointing at #10 among the most cited scientific papers of all time. Indeed, a large number of in silico analyses depend on MSA methods. These include domain analysis, phylogenetic reconstruction, motif finding and a whole range of other applications.

MSA is indeed an important modeling tool whose development has required addressing a complex combination of computational and biological problems. The computation of an accurate MSA has long been known to be an NP-complete problem, a situation that explains why over 100 alternative methods have been developed these past three decades. Original MSA methods (MSAMs) and their applications have been extensively covered by several reviews. To avoid redundancy, we will focus here on the main developments that have taken place over these past 10 years and put them in a broader historical context when needed. The three first sections will detail the general algorithmic framework of MSAMs and show how it relates to the newest methods and their application to all sorts of biological sequences (proteins, RNA, DNA). The fourth part will cover method validation and available benchmarks, with a special emphasis on the newest generation designed to cater for evolutionary and structural modeling. The last part of this review will deal with the quantification of local reliability within MSAs. This task had long been identified as instrumental, and possibly more important than the computation of the models—necessarily approximate. It is, however, only recently that systematic approaches have been developed with the explicit aim of quantifying local reliability, thus allowing a systematic filtering and weighting for downstream modeling. We will review these methods in the light of the latest reports.

Phylogenetic analysis and parsimony
In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor. The diagram is known as a phylogenetic tree. Phylogenetic analysis is important for gathering information on biological diversity, genetic classifications, as well as learning developmental events that occur during evolution. With advancements in genetic sequencing techniques, phylogenetic analysis now involves the sequence of a gene to understand the evolutionary relationships among species. DNA being the hereditary material can now be sequenced easily, rapidly, and cost-effectively, and the data obtained from genetic sequencing is very informative and specific. Also, morphological estimates can be used to infer evolutionary developments, especially in cases where genetic material is not available (fossils). A phylogenetic tree, also known as phylogeny, is characterized by a series of branching points expanding from the last common ancestor (root) of all operational taxonomic units up to the most recent organisms (tips). The tree is compared to leaves (tips), nodes, and branches, wherein two nearby nodes (taxonomic units) are connected by one branch (internal branch).

In a phylogenetic tree, leaves representing species, populations, individuals, or genes can be connected to nodes through branches (external branch). The branches represent the passage of genetic information between subsequent generations, and branch lengths denote genetic change or divergence. The degree of divergence is generally estimated using the average number of nucleotide substitutions per site. While analysing a phylogenetic tree from the root toward the tips, a node represents the exact position from where two or more descendant lineages are generated from an ancestral lineage. In the newly generated lineages, evolution occurs autonomously. The particular branching pattern created by lineage splitting is called topology, which represents the evolutionary development of the contemporary generation through progressive branching of lineages. Depending on the study requirement, a phylogenetic tree can be rooted or unrooted, as well as scaled or unscaled. The proper rooting of a phylogenetic tree is required to better understand the directionality of evolution and genetic divergence. In a scaled tree, a proportional relationship exists between the branch length and the amount of genetic divergence that took place on that branch. In contrast, all branches are of equal length in an unscaled tree, and there is no correlation between the branch length and genetic divergence.

Phylogenetic analysis provides an in-depth understanding of how species evolve through genetic changes. Using phylogenetics, scientists can evaluate the path that connects a present-day organism with its ancestral origin, as well as can predict the genetic divergence that may occur in the future. Phylogenetics has many applications in medical and biological fields, including forensic science, conservation biology, epidemiology, drug discovery and drug design, prediction of protein structure and function, and gene function. A more accurate understanding of relationship among species is now possible in a molecular phylogenetic analysis using gene sequencing data. Also, the Linnaean classification (based on relatedness in obvious physical traits) of newly evolved species can be done using molecular phylogenetic analysis.

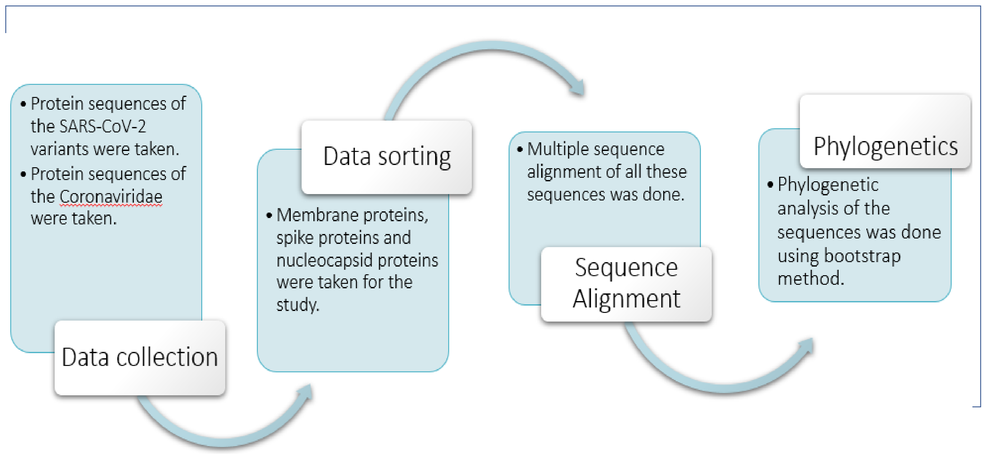
The parsimony score on networks is defined as the sum of the substitution costs along all the edges of the network; and show that certain well-known algorithms that calculate the optimum parsimony score on trees, such as Sankoff and Fitch algorithms extend naturally for networks, barring conflicting assignments at the reticulate vertices. We provide heuristics for finding the optimum parsimony scores on networks. Our algorithms can be applied for any cost matrix that may contain unequal substitution costs of transforming between different characters along different edges of the network. We analyzed this for experimental data on 10 leaves or fewer with at most 2 reticulations and found that for almost all networks, the bounds returned by the heuristics matched with the exhaustively determined optimum parsimony scores.

The parsimony principle states that the simplest explanation that explains the greatest number of observations is preferred over more complex explanations. Most phylogeneticists recognise that inferring genealogy rests on the principle of parsimony, that is, choosing evolutionary trees so as to minimize requirements for ad hoc hypotheses of similarity of observed characters.

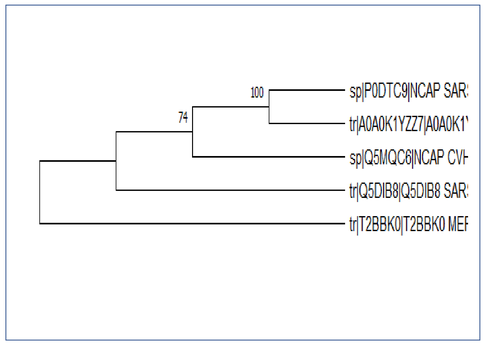
The cost of each character change event from a parent to child along an edge is weighted as a substitution cost of the parental state to the child state on the edge. The parsimony approach seeks a phylogenetic tree/network that, when we reconstruct the evolutionary events leading to the data on the leaves, minimizes the sum of the weights on the edges. We then face two important problems. First, we must be able to make a reconstruction of events on each vertex of the tree/network, such that the sum of the substitution costs on the edges is minimized (optimize the parsimony score on a network). Second, we must be able to search among all (or a subset of) possible phylogenetic networks for the one(s) that minimizes the parsimony score (find the network that has minimum parsimony score). This problem is NP-hard even for phylogenetic trees; and heuristic methods have been developed to reconstruct the phylogenetic network with the given number of reticulation vertices. In this paper, we restrict ourselves to the first of these issues, namely in establishing a parsimony criterion and to provide algorithms to achieve heuristics on finding the optimal score for any given phylogenetic network.

An often used structure to represent the evolution of sequences with reticulations is a family of trees each describing the evolution of a segment of the sequence. In previous approaches, the parsimony criterion on a network has been defined as the sum total of the substitution costs on the edges of a tree (a subgraph of the network) that minimizes the parsimony score of the site. The parsimony scores for networks given in these definitions are NP-hard to compute. Moreover, a major problem with this extension is that it favors more complex evolutionary relationships by adding larger numbers of edges to trees over simpler ones that contain fewer additional edges, thus having the potential of overestimating the amount of reticulation (horizontal events) in the data. An ad hoc solution has been provided by the authors, namely to restrict blocks of contiguous sites to optimize on the same tree, rather than choosing site-specific most parsimonious tree. However, it is not clear how these blocks are chosen.

Diagrammatic explanation:
Figure 3: It gives a chart of the steps followed to get the multiple sequence alignment and phylogenetic results. The sequences are taken from UniProt and NCBI Genbank. This is further employed to perform the MSA and parsimony of the protein sequences.



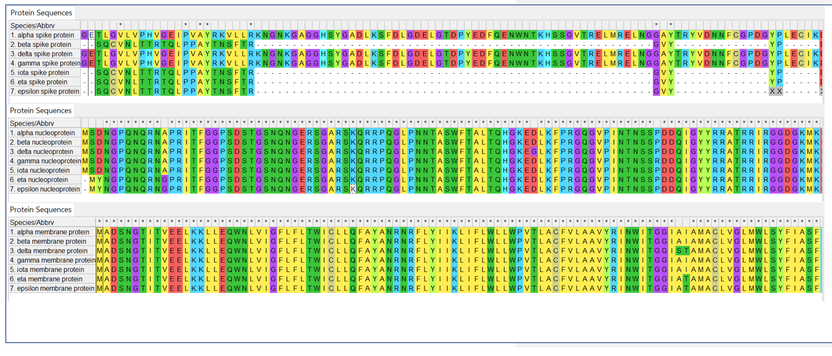
Results



Results from the *Coronaviridae* Family
The phylogenetic analysis of the family's membrane proteins showed 100% similarity between the bat isolate and the COVID-19 sequence. Further, 88% similarity was found between them and the MERS virus. The SARS and human isolate looked more different. The multiple sequence alignment showed that the MERS and the bat isolate were closer to each other, and the other three sequences were closer. The alignment of the nucleocapsid proteins talked more about the massive amount of variations that each isolate had gone through to be specialised in their own manners. They showed very little similarities. The phylogenetic analysis showed 100% similarity in the bat isolate and the COVID-19 isolate, while many differences were seen between others.
The spike proteins showed major dissimilarity between the sequences. They looked barely alike.

Results from the *SARS-CoV-2* variants
The membrane proteins showed more than 99% similarity during multiple sequence alignment. Alpha variant showed great similarity with beta, iota and gamma variants in the parsimony tree.(figs. 7, 9)
The nucleoproteins showed nearly complete similarity in multiple sequence alignment. But the parsimony tree showed many differences between the sequences(figs. 6, 10)
The spike proteins of alpha, gamma and delta variants were similar in sequence. There were many spaces in the sequences of beta, iota, eta and epsilon variants, as shown by multiple sequence alignment. The parsimony results also showed that the alpha, gamma, and delta variants were similar.(figs. 5,8)

Figure 4: the dendrogram shows the evolutionary connection between the members of the coronavirus family.



References:
Wang, J. T., Lin, Y. Y., Chang, S. Y., Yeh, S. H., Hu, B. H., Chen, P. J., & Chang, S. C. (2020). The role of phylogenetic analysis in clarifying the infection source of a COVID-19 patient. The Journal of infection, 81(1), 147–178. https://doi.org/10.1016/j.jinf.2020.03.031

Ciotti M, Angeletti S, Minieri M, Giovannetti M, Benvenuto D, Pascarella S, Sagnelli C, Bianchi M, Bernardini S, Ciccozzi M: COVID-19 Outbreak: An Overview. Chemotherapy 2019;64:215-223. doi: 10.1159/000507423

Liubov K, Anastasia P, Georgy I, Alexey S, Anna S, Anastasia K, Ilya G, Yury I, Anastasia B, et al. Isolation and phylogenetic analysis of SARS-CoV-2 variants collected in Russia during the COVID-19 outbreak, International Journal of Infectious Diseases, (2020) Vol. 99, 40-46, https://doi.org/10.1016/j.ijid.2020.07.024.

Thanat C, Evolving COVID-19 conundrum and its impact, Genomics and Evolutionary Medicine Unit, https://www.pnas.org/cgi/doi/10.1073/pnas.2007076117
Yuen KY, Lau SK, Woo PC. Wild animal surveillance for coronavirus HKU1 and potential variants of other coronaviruses. Hong Kong Medical Journal = Xianggang yi xue za zhi. 2012 Feb;18 Suppl 2:25-26. PMID: 22311357.

Mahase E. Delta variant: What is happening with transmission, hospital admissions, and restrictions? BMJ 2021; 373 :n1513 doi:10.1136/bmj.n1513

Hirofumi Y, Masato T, Hirofumi K, Satoru T, Fumiya A, Prinzmetal's variant form of angina as a manifestation of alpha-adrenergic receptor-mediated coronary artery spasm: Documentation by coronary arteriography, American Heart Journal, Vol. 91(2), 1976, 148-155, https://doi.org/10.1016/S0002-8703(76)80568-6.

Mittal, A., Manjunath, K., Ranjan, R. K., Kaushik, S., Kumar, S., & Verma, V. (2020). COVID-19 pandemic: Insights into structure, function, and hACE2 receptor recognition by SARS-CoV-2. PLoS pathogens, 16(8), e1008762. https://doi.org/10.1371/journal.ppat.1008762

Kannan, L., Wheeler, W.C. Maximum Parsimony on Phylogenetic networks. Algorithms Mol Biol 7, 9 (2012). https://doi.org/10.1186/1748-7188-7-9

Conclusions

From the two sub-parts, we can successfully conclude the following:
The human strain of COVID-19 is much stronger than the other strains of SARS, MERS and bat

The *Coronaviridae* family holds the ability to wipe out massive populations of humans but making use of its membrane, nucleocapsid and spike proteins.

The advancements in the methods of technologies has helped us determine these sequences and perform the operations we needed.

The variants of the *SARS-CoV-2* virus have the alpha, delta, gamma and beta variants being very closely related to each other.

Thus, they are categorised as variants of concern.

The variants eta, epsilon and iota are similar to each other in terms of their sequences and mode of action, and very little information was found on them.

Thus they are categorised as variants of interest.

The spike proteins change rapidly and play a very important role in infecting the host organism.

The membrane proteins and nucleocapsid proteins are conserved at several positions.

These properties of the variants can be extensively used to understand the regression curve of the virus and the possibilities of any further waves that might hit human populations across the world in the future.
The metadata from these sequences and the experiments performed using the wet lab software can be used to build vaccines and medical setups that can be useful in the combat against these viruses that swiftly take the form of a pandemic.

The following tools of bioinformatics were used:
NCBI GenBank- It was used to procure nucleotide sequences of the variants and collect their fastq data.

ExPASy- This was a translation tool to convert the data from nucleotide to protein format so that the spike, membrane and nucleotide proteins could be identified from the complete genomes of the variants as they were not found separately on the other databases.

UniProt- It was used to collect the protein sequences of the strains of the *Coronaviridae* family.

BLAST- The tool was used to correctly identify the proteins of the variants by comparing the data found with earlier known sequences.

MEGAX- This tool was used to understand the molecular and evolutionary traits of family strains and the variants by performing their multiple sequence alignments. They were further studied by understanding their phylogenetic construct using the maximum parsimony method involving bootstrapping.

Diagrammatic explanation:
Figure 8: This figure shows the evolutionary connection between the variants and their rate of similarity along the dendrogram.

Figure 9: This figure explains the curve of the virus in India and how the past two waves have been for the Indians.