J Appl Biotechnol Rep. 2021 June;8(2):116-126

doi 10.30491/JABR.2021.260175.1319

Journal of
Applied Biotechnology
Reports

Original Article

An Insight Into SARS-CoV-2 Phylogenetics and Genomics for Sixty Isolates Occurring in India

Koteswara Reddy Gujjula[1*], Nikhil Reddy Varakala[1], Divyanshu Dhakate[1], Harikishan R. Ellamla[2], Shadrack Jabes B[3]

[1]Department of Biotechnology, Koneru Lakshmaiah Education Foundation (Deemed to be University), Green Fields, Vaddeswaram-522502, Guntur, Andhra Pradesh, India
[2]Eduard-Zintl-Institut für Anorganische und Physikalische Chemie, Centre of Smart Interfaces, Technische Universität Darmstadt, Alarich-Weiss-Straße 10, 64287, Darmstadt, Germany
[3]Department of Chemical Engineering, Federal University of São Carlos, Rod. Washington Luiz, São Carlos-13565905, Brazil

Corresponding Author: Koteswara Reddy Gujjula, PhD, Assistant Professor, Department of Biotechnology, Koneru Lakshmaiah Education Foundation (Deemed to be University), Green Fields, Vaddeswaram-522502, Guntur, Andhra Pradesh, India. Tel: +91-8555913247, Email: koteswarareddy@kluniversity.in

Received December 2, 2020; **Accepted** April 4, 2021; **Online Published** May 27, 2021

Abstract

Introduction: Analysis of genome sequences to search for encoded proteins and motifs is a most widely used technique for the prediction of new drug and vaccine targets. It can effectively leverage computational techniques to deliver effective and pragmatic advantages in the search of new drug and vaccines.

Materials and Methods: The diversity and evolution of the SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2) isolates have been examined from different geographical parts of India using phylogenetic tree analysis. A dataset of 172 Indian SARS-CoV-2 genome sequences were collected from a database and a phylogenetic tree was constructed.

Results: From the phylogenetic analysis, six different clusters were identified and from each cluster 10 genome sequences were chosen to find open reading frames (ORFs) and common encoded proteins. Four encoded proteins that are common among the 60 genome sequences and correspond to ORF7a protein, *Membrane glycoprotein, Surface glycoprotein* and *Nucleocapsid phosphoproteins* were found. Our results highlight that there are six conserved motifs with high frequency of occurrence suggesting a potential of being useful in further studies.

Conclusions: The encoded proteins and their detected sequential motifs might be useful for screening potential drugs and vaccine candidates of the SARS-CoV-2 Indian isolates in the current epidemic situation.

Keywords: SARS-CoV-2, COVID-19, Phylogenetics, Genomics, Motif, Vaccine Target

Citation: Gujjula KR, Varakala NR, Dhakate D, Ellamla HR, Jabes B S. An insight into SARS-CoV-2 phylogenetics and genomics for sixty isolates occurring in India. J Appl Biotechnol Rep. 2021;8(2):116-126. doi:10.30491/JABR.2021.260175.1319.

Introduction

The first outbreak of human coronavirus was recorded in 1965-HCoV-229E, and thereafter two outbreaks of the severe acute respiratory syndrome coronavirus (SARS-CoV) and it was out broken first in the Middle East as a respiratory syndrome coronavirus (MERS-CoV) in 2003 and in the year of 2012 were accountable for severe infection and high mortality rates.[1,2] In the past 17 years, this is one of the emergence of a novel coronavirus which was originated from China, after a severe acute respiratory syndrome (SARS) which happened in the year of 2003.[3-6] The outbreaks of the coronaviruses (CoVs) in the history of humankind have been presented in Table 1. The outbreak of a novel coronavirus was recently reported in Wuhan city, republic of China in December 2019.[7] The CoVs belong to the family of Coronaviridae and are enclosed in single-stranded and also positive-sense RNA (ribonucleic acid) viruses.[1] The CoVs are seen to be spread mostly in animals as well as in humans causing mild to severe infections. There are a couple of research papers on the analysis of these cases that were carried out by Lu et al to identify the contributing agent of pneumonia.[8]

The cases of the SARS-CoV-2 has been rapidly expanding, with an increase in the number of cases throughout the world.[9,10] A novel coronavirus renamed as SARS-CoV-2, contributory agent of the coronavirus disease 2019 (COVID-19), as on September 16, 2020, spread to over 207 countries and caused 29 908 541 confirmed cases globally with 942 346 reported mortalities.[10] The transmission of the SARS-CoV-2 will be high in largely populated country like India. In India, the first laboratory-confirmed infection by SARS-CoV-2 was reported on January 30, 2020. Since then, it has been reported from 32 states/union territories. The total reported SARS-CoV-2 cases in India as on September14, 2020 was 4 926 914 with a total mortality over 79 754[10] (Figure 1). In the present pandemic,

Table 1. Outbreaks of Coronavirus

Outbreak	Year	Coronavirus	Effect
I	1965	HCoV-229E	Infected humans
II	2003	SARS-CoV	Infected humans with mild symptoms
III	2012	MERS-CoV	High infection and mortality rate
Present	December, 2019	SARS-CoV-2	Disaster

Sources: Chen et al,[3] Ghosh et al,[1] Hu et al,[6] Zhao et al,[2] Zhong et al,[4] Zhu et al.[7]

the isolation of the SARS-CoV-2 is significant for increasing and weighing diagnostic reagents, for the screening of vaccine candidates and for the antiviral research. The present research is an important study to understand the genomic nature of the spreading SARS-CoV-2. The isolation of the virus and its whole genomic characterization of the virus were studied and it was identified as a novel CoV, named 2019-nCoV through a next-generation sequencing.[3] The virus characterization study discovered that it is an enclosed RNA virus with a gene sequence size of 30 000. Many research have reported the antimicrobial compounds[11-13] which can be used to study the encoded viral protein and ligand interactions in the design of novel anti-viral drugs.[14-16]

Currently a large dataset on SARS-CoV-2 sequences are accessible from isolates which have been sampled and tested during the pandemic situation from different parts of India. The present study is divided into two sections. The first section deals with the diversity and evolution of SARS-CoV-2 with the progression of the pandemic situation overtime across the different geographical parts of India. The second section deals with the identification of sequential motifs and open reading frames (ORFs), which encoded proteins for screened isolates from phylogenetic analysis.

Material and Methods

Genome Sequence Dataset

The whole genomic sequence data was collected from the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/).[17] This is while only 172 SARS-CoV-2 full genome were submitted from India until the 2nd of June 2020. Full-length genome sequences of 18 different parts of Indian SARS-CoV-2 were taken and used the first viral genome sequence named ncov2019-Wuhan-hu-1/2019 (NCBI accession no: NC_045512.2) as reference. From the analysis of the 172 genomic sequences in the phylogenetic tree, 60 sequences were taken to further evolutionary analysis (Table 2).

Evolutionary Analysis

A phylogenetic tree is a diagram that represents evolutionary relationships among species or organisms. The pattern of branching in a phylogenetic tree reflects how species or other groups are evolved from a series of common ancestors and estimates the evolutionary relationship among taxa or biological sequences and their hypothetical common ancestor. Most molecular phylogenetic trees estimate the statistically significant relationships among the species/sequences.[18-20]

Interned Parameters of MEGA

The Molecular Evolutionary Genetics Analysis (MEGA) software was used for the analysis of multiple sequence alignments for this study. The MEGA software is used to measure evolutionary distance among the species or sequences for the construction of neighbour joining phylogenetic tree based on customized algorithmic set.[21] Evolutionary distance was estimated by using probability parameter (p-distance) under Jukes/Cantor model for two datasets of 172 and 60 genome sequences.

Parameters Used in iTOL Server

The constructed trees were visualized through an Interactive Tree Of Life (iTOL) (https://itol.embl.de).[22] The iTOL server is used to differently colour the selected set of species or clades and it is also useful to analyse their evolutionary relationships among the species or sequences. Display mode: Circular; Parameters: 210° Rotation and 350° Arc; Invert: No; Branch

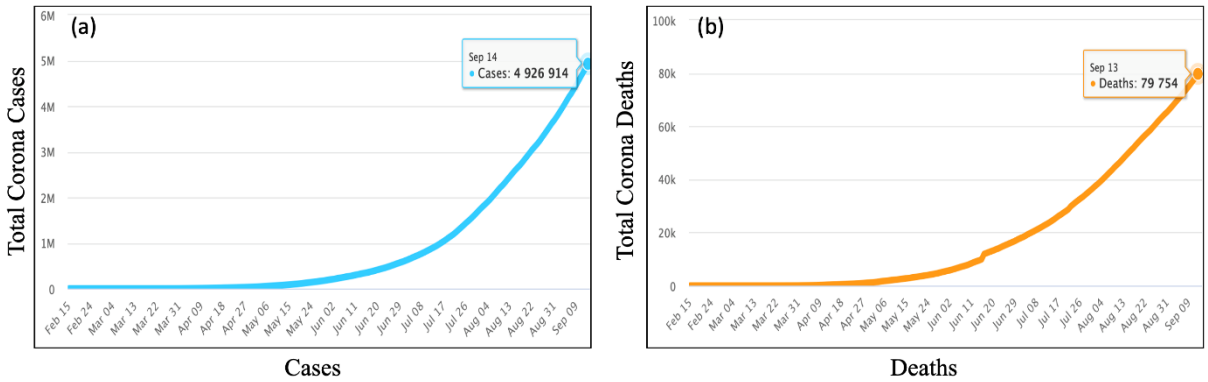

Figure 1. a) Total Coronavirus Cases, b) Total Coronavirus Deaths in India as on 14th September, 2020 (Culp, 2020, Worldmeter-2020).

Table 2. The List of SARS-Cov-2 Isolates From Different Geographical Parts of India

Isolates/Source	Accession Number*	Location/State
hCoV-19/India10/NCDC-4874/2020	EPI ISL 436459	Tamil Nadu
hCoV-19/India45/NCDC-2525/2020	EPI ISL 436421	Maharashtra
hCoV-19/India151/GUJRT/2020	EPI ISL 447045	Gujarat
hCoV-19/India35/NCDC-2519/2020	EPI ISL 436419	Madhya Pradesh
hCoV-19/India105/CCMB J014/2020	EPI ISL 447558	Telangana
hCoV-19/India53/UP/2020	EPI_ISL_435060	Uttar Pradesh
hCoV-19/India135/GUJRT/2020	EPI_ISL_447549	Gujarat
hCoV-19/India69/BNGL/2020	EPI_ISL_437539	Bengal
hCoV-19/India136/GUJRT/2020	EPI_ISL_447548	Gujarat
hCoV-19/India146/GUJRT/2020	EPI_ISL_447052	Gujarat
hCoV-19/India155/GUJRT/2020	EPI_ISL_447039	Gujarat
hCoV-19/India31/NCDC-02323/2020	EPI_ISL_435089	Andhra Pradesh
hCoV-19/India23/Ladakh/NCDC-01760/2020	EPI_ISL_435106	Ladakh
hCoV-19/India67/NCDC-4444/2020	EPI_ISL_436453	Madhya Pradesh
hCoV-19/India/Maharashtra42/NCDC-02330/2020	EPI_ISL_435077	Maharashtra
hCoV-19/India33/NCDC-3941/2020	EPI_ISL_436441	Bihar
hCoV-19/India107/CCMB_J048/2020	EPI_ISL_447559	Telangana
hCoV-19/India54/UP/2020	EPI_ISL_436413	Uttar Pradesh
hCoV-19/India48/Kerala/1-31/2020	EPI_ISL_413523	Kerala
hCoV-19/India109/CCMB_J067/2020	EPI_ISL_447561	Telangana
hCoV-19/India115/CCMB_J304/2020	EPI_ISL_447567	Telangana
hCoV-19/India113/CCMB_J278/2020	EPI_ISL_447565	Telangana
hCoV-19/India108/CCMB_J043/2020	EPI_ISL_447560	Telangana
hCoV-19/India96/NCDC-02334/2020	EPI_ISL_435080	Tamil Nadu
NC_045512.2 SARS-CoV-2	NC_045512.2	China, Wuhan
hCoV-19/India/TMLND84/2020	EPI_ISL_447584	Tamil Nadu
hCoV-19/India61/NCDC-4877/2020	EPI_ISL_436461	Madhya Pradesh
hCoV-19/India/DELHI2/2020	EPI_ISL_436454	Delhi
hCoV-19/India116/CCMB_J166/2020	EPI_ISL_447568	Telangana
hCoV-19/India63/NCDC-4874/2020	EPI_ISL_436459	Madhya Pradesh
hCoV-19/India167/GUJRT/2020|EPI_ISL_426415	EPI_ISL_426415	Gujarat
hCoV-19/India/Maharashtra39/NCDC-3965/2020	EPI_ISL_436444	Maharashtra
hCoV-19/India/DELHI16/2020	EPI_ISL_435071	Delhi
hCoV-19/India/UP55/2020	EPI_ISL_435100	Uttar Pradesh
hCoV-19/India/TMLND92/2020	EPI_ISL_435091	Tamil Nadu
hCoV-19/India/Maharashtra44/NCDC-02310/2020	EPI_ISL_435085	Maharashtra
hCoV-19/India/TMLND91/2020	EPI ISL 435091	Tamil Nadu
hCoV-19/India/TMLND89/2020	EPI ISL 435093	Tamil Nadu
hCoV-19/India118/CCMB J224/2020	EPI ISL 447571	Telangana
25hCoV-19/India58/NCDC-4879/2020	EPI ISL 436463	Uttar Pradesh
hCoV-19/India147/GUJRT/2020	EPI ISL 447051	Gujarat
24hCoV-19/India/UP57/2020	EPI ISL 435082	Uttar Pradesh
hCoV-19/India160/GUJRT/2020	EPI ISL 447041	Gujarat
hCoV-19/India119/CCMB J230/2020	EPI ISL 447572	Telangana
hCoV-19/India15/DELHI/2020	EPI ISL 447041	Delhi
13hCoV-19/India34/NCDC-3934/2020	EPI ISL 436439	Kerala
28hCoV-19/India6/DELHI-3961/2020	EPI ISL 436443	Delhi
hCoV-19/India/TMLND88/2020	EPI ISL 435094	Tamil Nadu

Table 2. Continus

Isolates/Source	Accession Number*	Location/State
hCoV-19/India153/GUJRT/2020	EPI ISL 447042	Gujarat
hCoV-19/India100/GMC-KN443/2020	EPI ISL 431103	Gujarat
15hCoV-19/India/Mstra38/NCDC-3985/2020	EPI ISL 436446	Maharashtra
hCoV-19/India159/GUJRT/2020	EPI ISL 444480	Gujarat
hCoV-19/India6/DELHI-3961/2020	EPI ISL 436456	Delhi
hCoV-19/India48/Kerala/1-31/2020|EPI_ISL_413523	EPI_ISL_413325	Kerala
hCoV-19/India/TMLND92/2020	EPI_ISL_435190	Tamil Nadu
hCoV-19/India6/DELHI-3961/2020	EPI ISL 436344	Delhi
13hCoV-19/India34/NCDC-3934/2020	EPI ISL 436934	Kerala
15hCoV-19/India/Mstra38/NCDC-3985/2020	EPI ISL 436644	Maharashtra
hCoV-19/India/TMLND92/2020	EPI_ISL_435990	Tamil Nadu
hCoV-19/India119/CCMB J230/2020	EPI ISL 447275	Telangana

*GISAID/NCBI Accession number.

Length: Ignore; Labels: Aligned; Label rotation: On; Label Alignment: Left; Label Shift:0; Label Font: Aerial; Branch lines: Normal; Branch Gradients: Off; Scaling Factors: Hor 1 Vec 1; Leaf Sorting: Default; Internal node Symbols: one child off; (https://itol.embl.de).[22]

Open Reading Frames and Encoded Proteins

In a molecular genomic study, an ORF is the part of a reading frame that has the ability to be translated into potential protein encoding segments. One common use of ORF is as a piece of proof to assist in gene identification. Long ORFs sequences are often used, along with other information, to initially identify potential protein-coding sections or functional RNA-encoding regions in a genome sequence.[23] The ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/) is a tool which finds all ORFs of a selectable minimum size in a given sequence or in a sequence already available in the database and it gives graphical output. This software detects all possible ORFs using a standard or alternative genetic code. The resulting amino acid sequence can be stored in various file formats and also possible search sequences in the database using the smart BLAST server.

Identification of Sequential Motifs

A motif is a sequence pattern which occurs repeatedly in a group of similar sequences. Motifs are represented as position-dependent scoring (or probability) matrices, that define the probability of each possible alphabet at each position in the pattern.[24] First viral genome sequence named ncov2019-Wuhan-hu-1/2019 is used to find the conserved motifs for assigning the function for the representative sequence using Multiple EM for Motif Elicitation (MEME) server.[25] It uses statistical modelling techniques to automatically choose the best width, number of frequent occurrences, and clarification for each motif, and finds various motifs in a group of closely related sequences based on the user-specified statistical confidence threshold.[1,25,26] Motif Alignment and Search Tool (MAST) is used to find the sequences for matches to a set of motifs and arranges the sequences by the best possible combination match to all other motifs.[27-29]

Selection Parameters

The statistical parameters such as E-value, sites, width, log-likelihood ratio (LLR), relative entropy, and Bayes Threshold are used to detect and analyse the sequential motifs within nucleotide or amino acid sequences.[25]

E-value

The E-value represents the statistical model significance of the motif. It is usually finding the most statistically significant (with low E-value) motifs first. The E-value is an estimate of the expected number of motifs with the given LLR (or higher), and with the same size and site count motifs, that one would find in a similarly sequence set of random sequences (sequences where each position is independent and letters are chosen according to the background letter frequencies).[26]

Log-likelihood Ratio

LLR is the ratio of the probability occurrences of the given motif in the mathematical model and the same motif probability given in the background model. The LLR of the motif can be estimated by the following relation (Eq. 1).[26,27]

$$LLR = ln\left[\frac{P(Ma)}{P(Mb)}\right] \qquad \text{(Eq.1)}$$

Where LLR: log-likelihood ratio; $P(M_a)$, $P(M_b)$: The probability occurrences of the given motif in the mathematical model and its probability given in the background model.

Relative Entropy

The relative entropy of the motif is directly related to the LLR of the motif which occurs in all ORFs/encoded proteins by chance in motif search. The relative entropy of the motif can be estimated by the following relation (Eq. 2).[25-27]

$$RE = \left\lfloor \frac{LLR}{\left(N * ln(2)\right)} \right\rfloor \qquad \text{(Eq. 2)}$$

Where, RE: relative entropy; LLR: Log-likelihood Ratio; N= Number of contributing sites to the motif construction.

Bayes Threshold

The Bayes Threshold is greater than (>1) one, which means that the detected motif in the motif model is more strongly supported by the background model.[25-27]

Methodology

In this study, a dataset-A of 172 genome sequences were used to study the phylogenetic analysis. From the dataset-A, we chose genome sequences based on phylogenetic cluster and represented a dataset-B. The dataset-B of 60 genome sequences was selected to have an insight into their evolutionary relationship among the Indian isolates including the Wuhan originated isolate and proteomics specifically in the identification of ORFs and sequential motifs.[27-29] The process methodology can be seen in Figure 2.

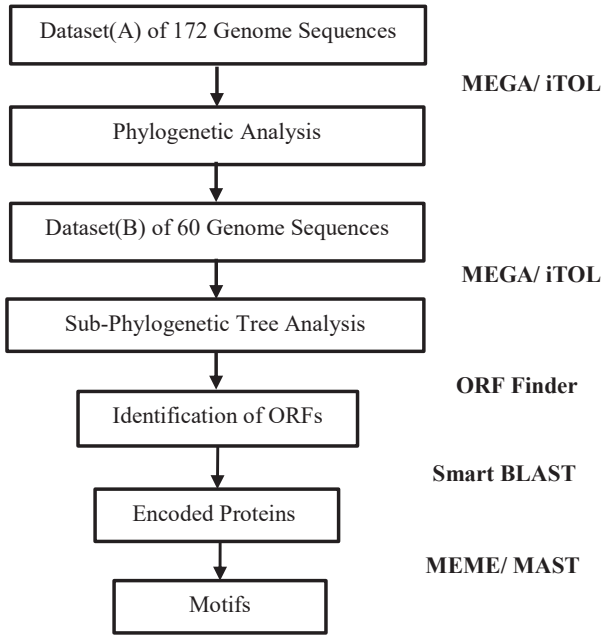

Figure 2. Methodology.

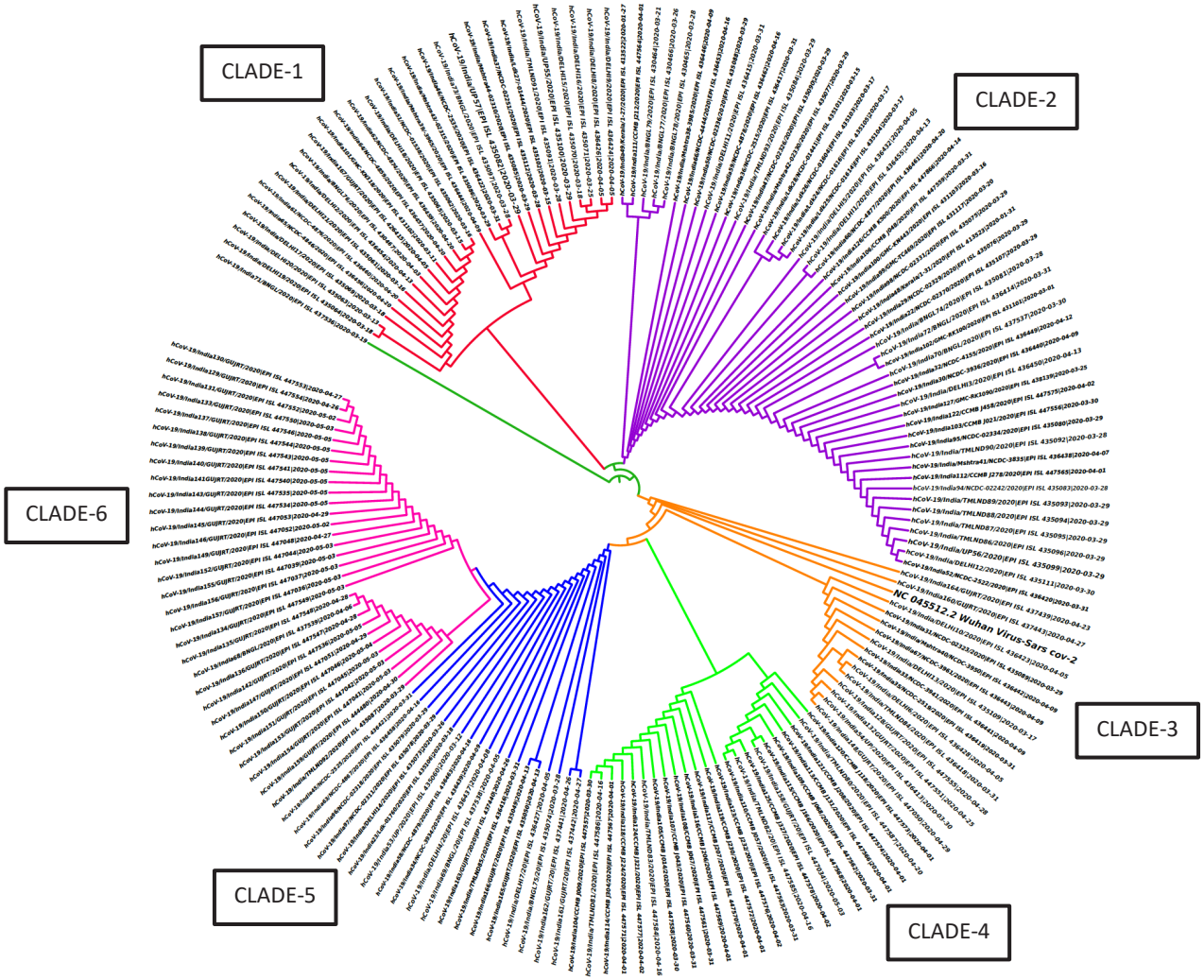

Figure 3. Phylogenetic Tree a Dataset (A) of 172 Genome Sequences of Indian isolates Including Wuhan Originaated Isolate.

Results and Discussion

SARS-CoV-2 Phylogenetic Analysis

The dataset-A of 172 genome sequences fed to the MEGA software to investigate the phylogenetic relationship based on the maximum likelihood phylogenetic protocol. A bagging of 1000 iterations were used to estimate the statistical stoutness of the phylogenetic analysis. The clustered set of sequences were identified to encode a substantial sequence similarity within the Indian isolates including the Wuhan originated isolate. The phylogenetic tree file generated in MEGA software and it was fed to an iTOL server and visualized as shown in Figure 3. In the phylogenetic analysis, clade represents the set of clustered sequences, which means that the set of genome sequences are grouped according to their sequence similarity. In this study, six distinguished clades were produced as in-group and one was observed under out-group and all the clades were coloured differently for easy analysis. In the phylogenetic analysis, the Wuhan originated isolate was identified in clade-3 along with Indian isolates (*EPI ISL 437439, EPI ISL 437443 and EPI ISL 436423*) particularly, from Gujarat and Delhi, which indicated that the Ind-Delhi/Gujarat isolates shared a close relationship with the isolate reported from Wuhan, China (Figure 4). The first contact might have travelled to Delhi on the 5th of April 2020 and the second and third contact might have travelled to Gujarat on the 23rd and 27th of April 2020 based on their travel history (Table 2).

In the phylogenetic analysis, a conventional way was applied to select the 10 genome sequences from each clade of phylogenetic tree with a dataset-B of 60 in order to have an insight into their evolutionary relationship among the Indian isolates including the Wuhan originated isolate. In this study, tree scale was used to investigate the percentage of variation occurring among clades or branches or genome sequences in a phylogenetic tree. However, a dataset-B of 60 fed to MEGA software and produced a tree file, which is deployed to an iTOL server and visualized with custom parameters such as colour clades, tree scale and boot strap value (Figure 5). From the evolutionary analysis, six distinguished clades were developed and coloured differently. The overall tree scale of the sub phylogenetic tree was 0.001. The six clades with the branch variation were identified as 0.0242, 0.062, 0.058, 0.025, 0.0001 and 0.012 respectively for clade-1, clade-2, clade3, clade-4, clade-5 and clade-6.

In the case of clade-1, the evolutionary branch scale showed a variation of 0.0242, by means that 2.42% variation occurred from a common hypothetical ancestor to the clade-1, Indian isolates (Figure 5). In the case of clade-2, the evolutionary branch scale showed a variation of 0.062, by means that 6.2% variation occurred from a common hypothetical ancestor to the clade-2, Indian isolates, which indicates that the diversity is more in clade-2 than clade-1 from a common hypothetical ancestor. This is while all ten isolates from *Telangana* have been closely related to each other and that the transmission of virus occurred from the 31st of March to the 1st of April 2020 based on their travel history.

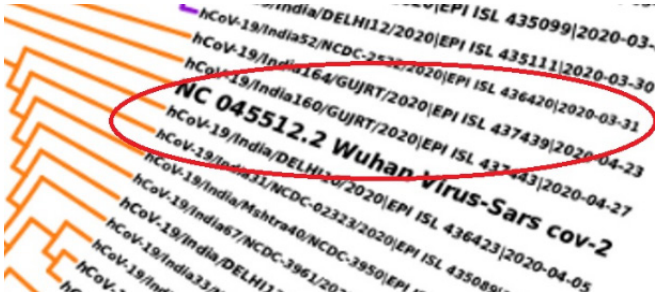

Figure 4. Clade 3 of Phylogenetic Tree With Indian Isolates (Delhi/Gujarat) and Wuhan Originated Isolate.

In the case of clade-3, the evolutionary branch scale showed a variation of 0.058, by means that 5.8% variation occurred from a common hypothetical ancestor to the clade-3, Indian isolates, which indicates that the diversity is more in clade-3 than clade-1 and close to clade-2 from a common hypothetical ancestor (Figure 5). However, all six Indian isolates of *Maharastra-1*, all three isolates of *Uttar Pradesh-3,* all five isolates of *Tamil Nadu-1*, one isolate of *Gujarat-2* and one isolate of *Delhi-2* are closely related to the Wuhan originated isolate which was first reported in December 2019. The primary contacts might have travelled to Maharashtra from the Wuhan originated isolate on the 29th of March 2020, then travelled to Tamil Nadu on the same day. The isolates *Delhi-2* and *Gujarat-2* might have been infected from Maharashtra isolates.

In the case of clade-4, the evolutionary branch scale showed a variation of 0.025, by means that 2.5% variation occurred from a common hypothetical ancestor.

In the case of clade-5, evolutionary branch scale showed a variation of 0.0001 that means that only 0.01% variation occurred from a common hypothetical ancestor to the clade-5 Indian isolates, which indicates the fact that less diversity occurred from a common hypothetical ancestor to the clade-5 Indian isolate (Figure 5). It has been observed that all three isolates *of Ladak-3,* all three isolates *of Maharastra-8,* and one isolate *of Delhi-3* are closely related to each other and might be close to a common hypothetical ancestor. In the case of clade-6, evolutionary branch scale showed a variation of 0.012, which means that 1.2% variation occurred from a common hypothetical ancestor to the clade-5 Indian isolates. It has been observed that all eight isolates *of Ladak-4*, all 10 isolates *of Gujarat-4,* one isolate *of Karnataka-1* and one isolate of *Uttar Pradesh-4* are closely related to each other. It indicated that the isolate *(Uttar Pradesh-4)* might be contacted with other isolates of this cluster.

The variation in clades of phylogenetic tree (Figure 5) indicated that the emerging heterogeneity occurred within the SARS-CoV-2 isolates across India. Clade-5 Indian isolates (*Ladak-3, Maharastra-8, Delhi-3*) have very close relationships to a common hypothetical ancestor with a less diversity of just 0.01% as tree scale. Phylogenetic analysis revealed that the Indian isolates belonged to six different clusters. The variation in clades of phylogenetic tree indicated that the emerging heterogeneity occurred within the SARS-CoV-2 isolates across India. A previous study also characterized the distinct and predominant cluster of SARS-CoV-2 across India.[30]

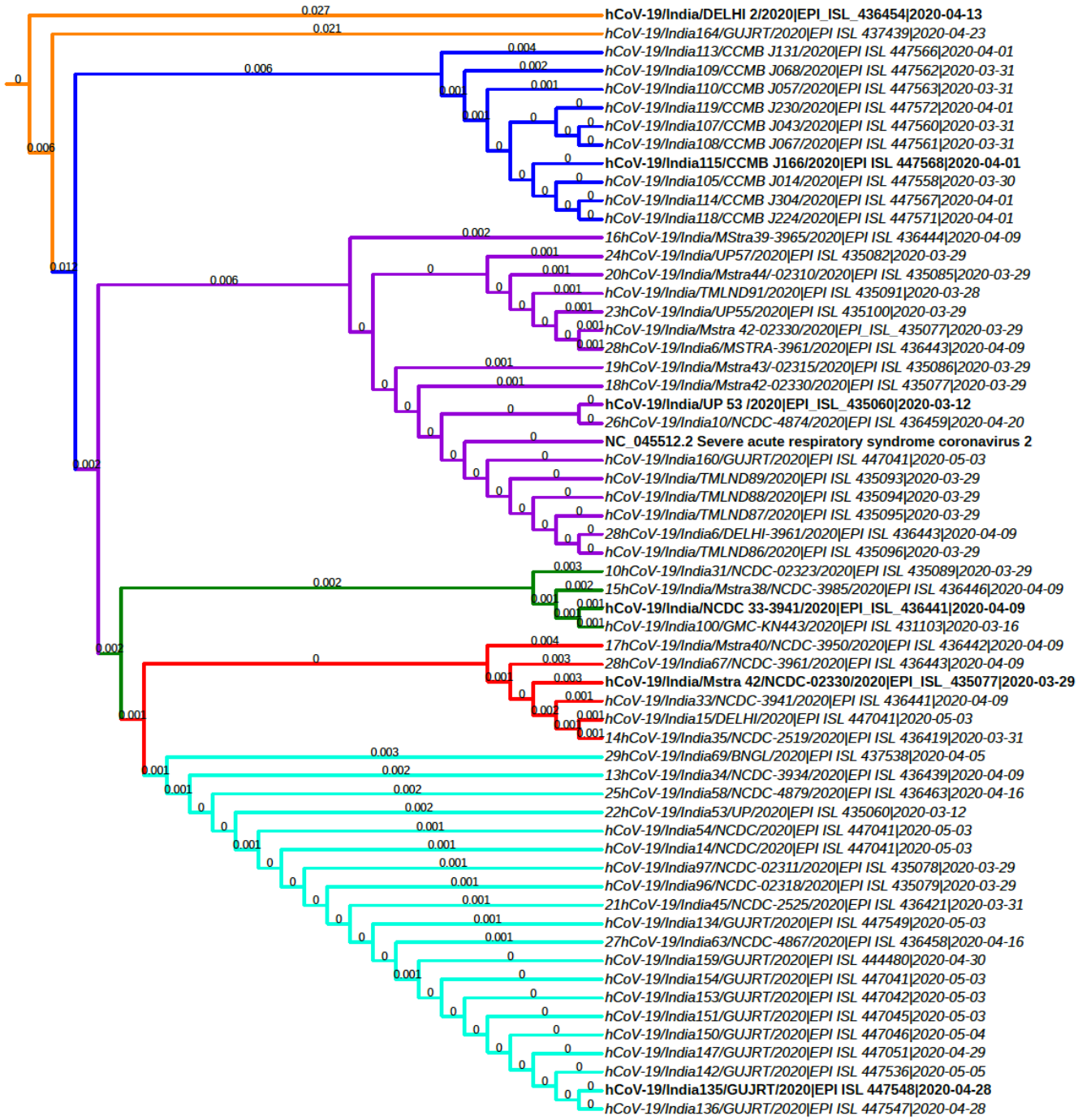

Figure 5: Sub-phylogenetic Tree With a Dataset (B) of 60 Genome Sequences of Indian Isolates Including Wuhan Originaated Isolate.

Further analysis was performed with 60 genome sequences in order to have insight into their genomics and proteomics specifically in the identification of ORFs for seven genomes of Indian isolates including the Wuhan originated isolate.

Identification of ORFs and Encoded Proteins

In molecular genomics, ORF is the part of a reading frame that can be translated into potential protein encoding segments. In this study, we considered the ORFs which code minimum 100 amino acids in length. However, the 100 amino acid length ORFs were encoded as possible proteins in the building blocks of viral particles. In this study, four common ORFs such as ORF7, ORF44, ORF25and ORF79 were identified with amino acid length of 121, 222, 1282 and 419, respectively among 60 Indian isolates using ORF Finder. The Molecular Weight (MW) of each ORF was calculated by an expert in protein analysis (Expasy) tool. ORF25 encodes 1282 amino acids and has a molecular weight of around 142 kDa, indicating that it may encode structural proteins such as viral particle surface, spike proteins because of its significant length and molecular weight when compared to other ORFs.

The smart BLAST was used to identify the encoded proteins against protein sequence databases. For this study, we identified four encoded proteins such as ORF7a protein

Table 3. Predictive Amino Acid Length and Molecular Weight of Proteins Encoded by ORFs of SARS-CoV-2 Indian Isolates

ORFs	Proteins Encoded	Length (AA)	*MW (kDa)	Database Accession	Identity
7	ORF7a protein	121	13.74	QJA41754.1	99.17%
44	Membrane glycoprotein	222	25.18	QJR92940.1	99.55%
25	Surface glycoprotein	1282	142.24	QKC53229.1	99.75%
79	Nucleocapsid phosphoprotein	419	45.62	QLC94852.1	99.76%

*Molecular Weight (MW) was calculated by online tool by Expasy (https://web.expasy.org/compute_pi/).

(QJA41754.1), membrane glycoprotein (QJR92940.1), surface glycoprotein (QJR92940.1) and nucleocapsid phosphoprotein (QLC94852.1) for their corresponding ORFs of 7, 44, 25 and 79, respectively (Table 3). In Figure 6, yellow colour represents the query ORF sequence, and the blue colour represents the significant matches to query ORFs. In this analysis, significant matches have been found in the case of a, b, c and d against protein sequence databases with an identity about 99.17% to 99.76% of query coverage of four ORFs. The detected four encoded proteins might be useful for drug and vaccine targets of SARS-CoV-2 of Indian isolates in the current pandemic situation.

Identification and Analysis of Sequential Motifs

Conserved motifs were identified within the four ORFs of 60 genome sequences. A total of six common motifs and their

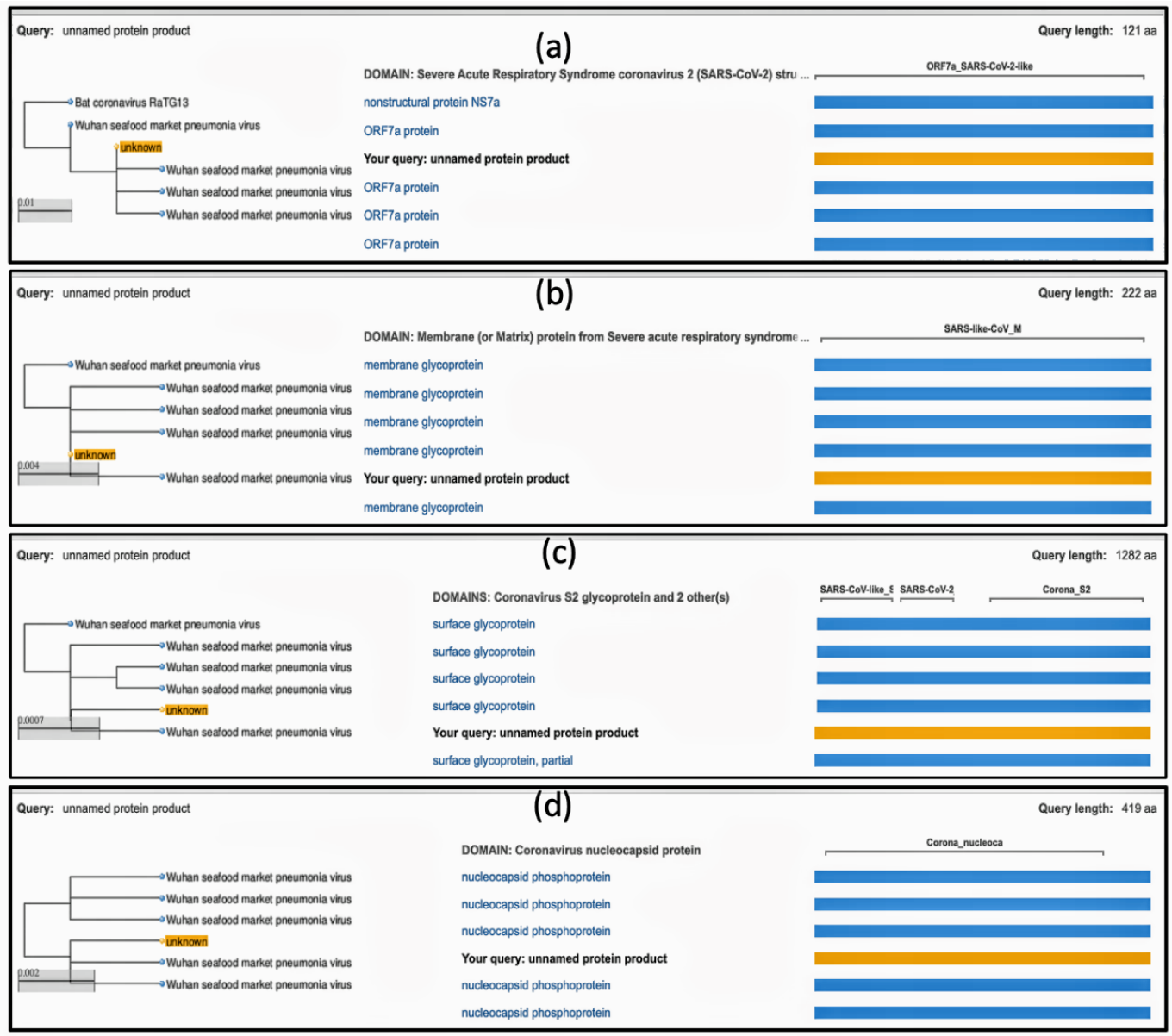

Figure 6. Predictive Amino Acid Length and Encoded Proteins in Smart BLAST Analysis. *Note*. Yellow colour: Query ORF and Blue colour: More significant matches to query ORFs

Table 4. Identifications of Six Motifs in MEME/MAST Server for Four Encoded Proteins of SARS-CoV-2 Indian Isolates

Motif	E-Value[a]	Sites (N)[b]	Width[c]	LLR[d]	RE[e]	Bayes Threshold[f]
YKKWPW	4.0	2	6	38	27.2	9.98014
CPDGIIW	9.3	4	7	67	24	9.71396
IMQCCMRGCCVCLKECCSCG	3.6E+01	2	20	105	75.5	9.93958
WFHAJH	7.4E+02	2	6	35	25.1	9.98014
RGDFCGKGHH	8.1E+02	2	10	54	38.8	9.96867
RKRIGNY	8.9E+02	2	7	53	25.5	9.83963

Abbreviations: LLR, log likelihood ratio; RE, relative entropy.
[a] E-value: the most statistically significant (low E-value) motifs first.
[b] The number of sites contributing to the construction of the motif.
[c] The width of the motif.
[c] The logarithm of the ratio of the probability of the occurrences of the motif given the motif model versus their probability given the background model.
[d] The relative entropy of the motif [RE = LLR / (N * ln (2))].
[e] Bayes Threshold > 1 means that the detected motif in motif model is more strongly supported by the background model.

widths were identified as YKKWPW-M1 (6), CPDGIIW-M2 (7), IMQCCMRGCCVCLKECCSCG-M3 (20), WFHAJH-M4 (6), RGDFCGKGHH-M5 (10) and RKRIGNY-M6 (7), respectively, by means that, the short length amino acid sequences repeated within the ORFs of SARS-CoV-2 Indian isolates and Wuhan originated isolate (Table 4). The pictorial representation of motifs and their consensus can be seen in Figure 7 and Figure 8. The E-value represents the statistical significance of each motif. It typically finds the most statistically significant (low E-value) motifs first. The E-values for the conserved six motifs were M1 (4.0), M2 (9.3), M3 (36), M4 (740), M5 (810) and M6 (890), respectively (Table 4). The E-value of the first M1 (4.0), and second M2 (9.3) motifs was less than 10, which indicates that these motifs are statistically significant than others.

The number of contributing sites to the construction of the six motifs were M1 (2), M2 (4), M3 (2), M4 (2), M5 (2) and M6 (2) respectively. The sites value of the second motif M2 (4), was more than the others, which indicates that the contribution of this motif was four times more in all four encoded proteins or ORFs. The LLR of the six motifs M1 (38), M2 (67), M3 (105), M4 (35), M5 (54) and M6 (53), respectively have been presented in Table 4. The probability of the occurrence of the third motif M3 (105) is more than the others. However, the occurrence of this motif is more in all ORFs due to its large length of amino acids. The relative entropy of the six motifs were M1 (27.2), M2 (24), M3 (75.5), M4 (25.1), M5 (38.8) and M6 (25.5), respectively. The RE of the third motif M3 (75.5) is more than the others. However, the probability of occurrence of this motif M3 in all ORFs is more due to its high LLR of 105 and sites of 2 in contributing as the most significant motif's formation than the other motifs. The Bayes Threshold was greater than one (>1), means that the detected motif in the motif model is strongly supported by the background model. The Bayes Threshold of the six motifs was M1 (9.98014), M2 (9.71396), M3 (9.93958), M4 (9.98014), M5 (9.96867) and

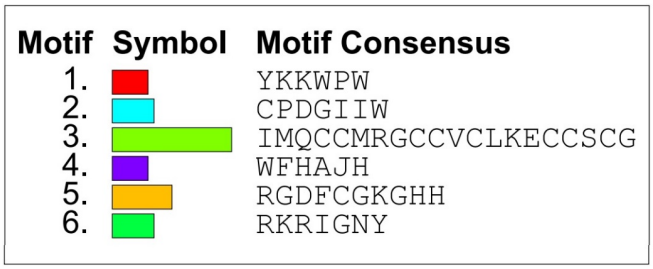

Figure 8. Motifs Consensus Within the ORFs/Encoded Proteins.

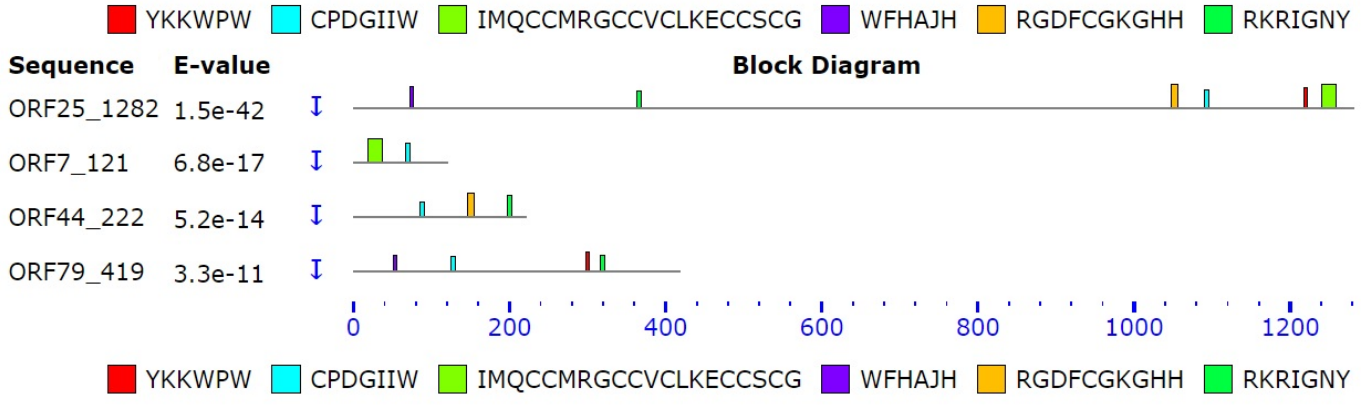

Figure 7. Identified Motifs Within Encoded Proteins/ORFs for SARS-CoV-2 Indian Isolates. Note. Different colours represent the distinguished six motifs with in the ORFs.

M6 (9.83963), respectively. The Bayes Threshold values were close to 10 for the six motifs, which indicates that the detected motifs in the motif model have been strongly supported by the background model.

The six motifs are ranked as M3 (1), M5 (2), M1 (3), M2 (4), M4 (5) and M6 (6) respectively. These are based on six parameters such as low E-value, large site number, large width, high likelihood ratio, high relative entropy, and high Bayes threshold values. The detected six motifs may be useful for the identification of drug and vaccine candidates to effectively control the SARS-CoV-2 in the current pandemic situation.

Conclusions

In this study, the whole genome sequence of SARS-CoV-2 Indian isolates were investigated, were sorted in to six different clades from phylogenetic cluster analysis. The variation in all six clades were observed, which indicates that an emerging heterogeneity within the SARS-CoV-2 isolates across India. However, four ORF encoded proteins were identified such as *ORF7a protein, Membrane glycoprotein, Surface glycoprotein* and *Nucleocapsid phosphoproteins* of the SARS-CoV-2 Indian isolates. Moreover, six highly conserved motifs also detected within ORFs such as YKKWPW-M1 (6), CPDGIIW-M2(7), IMQCCMRGCCVCLKECCSCG-M3 (20), WFHAJH-M4 (6), RGDFCGKGHH-M5 (10) and RKRIGNY-M6 (7) respectively. The encoded proteins and detected sequential motifs might be useful for screening drug and vaccine candidates of the SARS-CoV-2 Indian isolates in the current pandemic situation.

Authors' Contributions

KRG described the work plan to carry out this research. NRV collected the genome sequences and executed the phylogenetic analysis study. DD supported the bioinformatics tools and software's to accomplish the study. SJ and HRE both supervised the study. All authors contributed to the study. All authors read and approved the final manuscript.

Conflict of Interest Disclosures

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors are thankful to Koneru Lakshmaiah Education Foundation (KL Deemed to be University) for providing sophisticated laboratory facilities in order to carry out this research.

References

1. Ghosh AK, Xi K, Johnson ME, Baker SC, Mesecar AD. Progress in anti-SARS coronavirus chemistry, biology and chemotherapy. Annu Rep Med Chem. 2007;41:183-196. doi:10.1016/s0065-7743(06)41011-3.
2. Zhao Z, Zhang F, Xu M, et al. Description and clinical treatment of an early outbreak of severe acute respiratory syndrome (SARS) in Guangzhou, PR China. J Med Microbiol. 2003;52(Pt 8):715-720. doi:10.1099/jmm.0.05320-0.
3. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507-513. doi:10.1016/s0140-6736(20)30211-7.
4. Zhong NS, Zheng BJ, Li YM, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. Lancet. 2003;362(9393):1353-1358. doi:10.1016/s0140-6736(03)14630-2.
5. Alsahafi AJ, Cheng AC. The epidemiology of Middle East respiratory syndrome coronavirus in the Kingdom of Saudi Arabia, 2012-2015. Int J Infect Dis. 2016;45:1-4. doi:10.1016/j.ijid.2016.02.004.
6. Hu B, Ge X, Wang LF, Shi Z. Bat origin of human coronaviruses. Virol J. 2015;12(1):221. doi:10.1186/s12985-015-0422-1.
7. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-733. doi:10.1056/NEJMoa2001017.
8. Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle. J Med Virol. 2020;92(4):401-402. doi:10.1002/jmv.25678.
9. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-273. doi:10.1038/s41586-020-2012-7.
10. Culp WC Jr. Coronavirus disease 2019: in-home isolation room construction. A A Pract. 2020;14(6):e01218. doi:10.1213/xaa.0000000000001218.
11. Mannam MR, Devineni SR, Pavuluri CM, Chamarthi NR, Kottapalli RS. Urea and thiourea derivatives of 3-(trifluoromethyl)-5,6,7,8-tetrahydro-[1, 2, 4]triazolo[4,3-a] pyrazine: synthesis, characterization, antimicrobial activity and docking studies. Phosphorus Sulfur Silicon Relat Elem. 2019;194(9):922-932. doi:10.1080/10426507.2019.1577845.
12. Maddali NK, Viswanath IK, Murthy YL, et al. Design, synthesis and molecular docking studies of quinazolin-4-ones linked to 1,2,3-triazol hybrids as Mycobacterium tuberculosis H37Rv inhibitors besides antimicrobial activity. Med Chem Res. 2019;28(4):559-570. doi:10.1007/s00044-019-02313-9.
13. Dasari SR, Tondepu S, Vadali LR, Seelam N. Design, synthesis and molecular modeling of nonsteroidal anti-inflammatory drugs tagged substituted 1,2,3-triazole derivatives and evaluation of their biological activities. J Heterocycl Chem. 2019;56(4):1318-1329. doi:10.1002/jhet.3503.
14. Bodige S, Ravula P, Gulipalli KC, et al. Design, synthesis, antitubercular and antibacterial activities of pyrrolo[3,2-b] pyridine-3-carboxamide linked 2-methoxypyridine derivatives and in silico docking studies. Synth Commun. 2019;49(17):2219-2234. doi:10.1080/00397911.2019.1618874.
15. Malothu N, Kulandaivelu U, Jojula M, Gunda SK, Akkinepally RR. Synthesis, antimycobacterial evaluation and docking studies of some 7-methyl-5,6,7,8-tetrahydropyrido[4',3':4,5] thieno[2,3-d]pyrimidin-4(3H)-ones. Chem Pharm Bull (Tokyo). 2018;66(10):923-931. doi:10.1248/cpb.c17-00999.
16. Konidena LNS, Boda SK, Chettu SK, et al. Synthesis, biological evaluation and molecular docking studies of novel 2-(2-cyanophenyl)-N-phenylacetamide derivatives. Res Chem Intermed. 2018;44(9):5467-5481. doi:10.1007/s11164-018-3434-9.
17. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017;1(1):33-46. doi:10.1002/gch2.1018.
18. Glaser F, Pupko T, Paz I, et al. ConSurf: identification

of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics. 2003;19(1):163-164. doi:10.1093/bioinformatics/19.1.163.

19. Srideepthi R, Krishna MSR, Suneetha P, Krishna RS, Karthikeyan S. Genome-wide identification, characterization and expression analysis of non-RD receptor like kinase gene family under *Colletotrichum truncatum* stress conditions in hot pepper. Genetica. 2020;148(5-6):283-296. doi:10.1007/s10709-020-00104-4.
20. Srideepthi R, Lakshmisahitya U, Peddakasim D, Suneetha P, Krishna MS. Morphological, pathological and molecular diversity of *Colletotrichum capsici* inciting fruit rot in Chilli (*Capsicum annuum* L.). Res J Biotech. 2017;12:14-21.
21. Tabassum Khan N. MEGA - Core of Phylogenetic Analysis in Molecular Evolutionary Genetics. J Phylogenetics Evol Biol. 2017;5(2):1000183. doi:10.4172/2329-9002.1000183.
22. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256-W259. doi:10.1093/nar/gkz239.
23. Yella VR, Bansal M. DNA structural features of eukaryotic TATA-containing and TATA-less promoters. FEBS Open Bio. 2017;7(3):324-334. doi:10.1002/2211-5463.12166.
24. Koteswara Reddy G, Nagamalleswara Rao K, Yarrakula K. Insights into structure and function of 30S ribosomal protein S2 (30S2) in *Chlamydophila pneumoniae*: a potent target of pneumonia. Comput Biol Chem. 2017;66:11-20. doi:10.1016/j.compbiolchem.2016.10.014.
25. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006;34(Web Server issue):W369-373. doi:10.1093/nar/gkl198.
26. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39-49. doi:10.1093/nar/gkv416.
27. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 1998;14(1):48-54. doi:10.1093/bioinformatics/14.1.48.
28. Yella VR, Kumar A, Bansal M. Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. Sci Rep. 2018;8(1):4520. doi:10.1038/s41598-018-22129-8.
29. Yella VR, Bhimsaria D, Ghoshdastidar D, Rodríguez-Martínez JA, Ansari AZ, Bansal M. Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. Nucleic Acids Res. 2018;46(22):11883-11897. doi:10.1093/nar/gky1057.
30. Banu S, Jolly B, Mukherjee P, et al. A distinct phylogenetic cluster of Indian severe acute respiratory syndrome coronavirus 2 isolates. Open Forum Infect Dis. 2020;7(11):ofaa434. doi:10.1093/ofid/ofaa434.