Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Genome-wide identification and prediction of SARS-CoV-2 mutations show an abundance of variants: Integrated study of bioinformatics and deep neural learning

Md Shahadat Hossain [a,1], A.Q.M. Sala Uddin Pathan [b,1], Md Nur Islam [a], Mahafujul Islam Quadery Tonmoy [a], Mahmudul Islam Rakib [b], Md Adnan Munim [a], Otun Saha [c], Atqiya Fariha [a], Hasan Al Reza [d], Maitreyee Roy [e], Newaz Mohammed Bahadur [f], Md Mizanur Rahaman [c,*]

[a] *Department of Biotechnology & Genetic Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh*
[b] *Department of Computer Science and Telecommunication Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh*
[c] *Department of Microbiology, University of Dhaka, Dhaka, Bangladesh*
[d] *Department of Genetic Engineering and Biotechnology, University of Dhaka, Dhaka, Bangladesh*
[e] *School of Optometry and Vision Science, Faculty of Medicine and Health, University of New South Wales, Bangladesh*
[f] *Department of Applied Chemistry and Chemical Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh*

ARTICLE INFO

Keywords:
Genomic data
SARS-CoV-2
Mutation
Mutation rate
COVID-19

ABSTRACT

Genomic data analysis is a fundamental system for monitoring pathogen evolution and the outbreak of infectious diseases. Based on bioinformatics and deep learning, this study was designed to identify the genomic variability of SARS-CoV-2 worldwide and predict the impending mutation rate. Analysis of 259044 SARS-CoV-2 isolates identified 3334545 mutations with an average of 14.01 mutations per isolate. Globally, single nucleotide polymorphism (SNP) is the most prevalent mutational event. The prevalence of C > T (52.67%) was noticed as a major alteration across the world followed by the G > T (14.59%) and A > G (11.13%). Strains from India showed the highest number of mutations (48) followed by Scotland, USA, Netherlands, Norway, and France having up to 36 mutations. D416G, F106F, P314L, UTR:C241T, L93L, A222V, A199A, V30L, and A220V mutations were found as the most frequent mutations. D1118H, S194L, R262H, M809L, P314L, A8D, S220G, A890D, G1433C, T1456I, R233C, F263S, L111K, A54T, A74V, L183A, A316T, V212F, L46C, V48G, Q57H, W131R, G172V, Q185H, and Y206S missense mutations were found to largely decrease the structural stability of the corresponding proteins. Conversely, D3L, L5F, and S97I were found to largely increase the structural stability of the corresponding proteins. Multi-nucleotide mutations GGG > AAC, CC > TT, TG > CA, and AT > TA have come up in our analysis which are in the top 20 mutational cohort. Future mutation rate analysis predicts a 17%, 7%, and 3% increment of C > T, A > G, and A > T, respectively in the future. Conversely, 7%, 7%, and 6% decrement is estimated for T > C, G > A, and G > T mutations, respectively. T > G\A, C > G\A, and A > T\C are not anticipated in the future. Since SARS-CoV-2 is mutating continuously, our findings will facilitate the tracking of mutations and help to map the progression of the COVID-19 intensity worldwide.

1. Introduction

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the causal agent for the ongoing pandemic of non-acronym coronavirus disease 2019 (COVID-19) which is in part a respiratory disease showing pneumonia-like symptoms and was first recorded in December 2019 in Wuhan, China [1]. Spike glycoprotein (S), membrane (M), envelope (E), and nucleocapsid (N) proteins are four important structural proteins encoded by structural genes located in the region preceding the 3′ end of the genome, whereas several non-structural proteins (NSP)

[*] Corresponding author. Department of Microbiology, University of Dhaka, Dhaka, 1200, Bangladesh.
E-mail address: razu002@du.ac.bd (M.M. Rahaman).
[1] Contributed equally to this study.

https://doi.org/10.1016/j.imu.2021.100798
Received 19 September 2021; Received in revised form 6 November 2021; Accepted 15 November 2021
Available online 18 November 2021

encoded by genes located in the 5′-region, namely NSP1 to NSP16 [2]. Sixteen non-structural proteins are released from pp1a and pp1ab after proteolytic cleavage by two cysteine proteases found in nsp3 (Papain--like protease) and nsp5 (chymotrypsin-like protease) [3]. RNA-dependent RNA polymerase (RdRp), also known as NSP12, is an important part of the replication and transcription machinery of SARS-CoV-2 [4]. These non-structural proteins play various roles in viral replication, transcription, morphogenesis, and evasion of the host immune system [2]. SARS-CoV-2 is highly transmissible due to its unique structural properties, such as a flat sialic acid binding domain, excellent binding interactions with the ACE2 entrance receptor, and the ability to use Furin and other protease cell receptors [5].

Mutation is the fundamental process that leads to the emergence of genetic diversity. Mutation can be occurred due to certain errors in SARS-CoV-2 genome during the replication process, specifically the time of copying RNA to a new cell [6]. Single nucleotide polymorphism (SNP), Insertion, and Deletion are the main type of mutations. There are numerous deletions observed in the SARS-CoV-2 variants from different geographic locations. In a Singapore case cluster, a SARS-CoV-2 variant with a 382-nucleotide deletion was detected from January to February 2020 and also detected in a traveler who returned from Wuhan, China, to Taiwan in February 2020 [7,8], which is responsible for decollate the ORF7b and preventing the transcription of ORF8 by eliminating the transcription-regulatory sequence of ORF8 and this variant was not detected after the march, 2020 though it was successfully transmitted. ORF8 deletion in multiple SARS-CoV-2 isolates identified in the cases of Bangladesh (345 nucleotides), Australia (138 nucleotides), and Spain (62 nucleotides) [8]. Most mutations found in SARS-CoV-2 genomes are predicted to be neutral or moderately deleterious, as highly deleterious alterations are swiftly purged. In some contexts, a minimal number of mutations are expected to affect viral phenotype in a manner that can affect several aspects of viral biology including pathogenicity, transmissibility, infectivity, and antigenicity [9]. Even though it is important not to confound mutations present in growing lineages that can affect viral biology [10], fitness-enhancing mutations were initially identified to have emerged within a few months of SARS-CoV-2 evolution. For instance, the D614G mutation in spike protein was figured out to be increasing in frequency in April 2020 and to have emerged several times globally [11]. Several studies documented that D614G reveals a moderate benefit to SARS-CoV-2 for infectivity and transmissibility [11,12]. However, the mutation rate of the SARS-CoV-2, which is the basis of the COVID-19 pandemic, is crucial for precise interpretations of SARS-CoV-2 demographic evolution and possible molecular adaptations. In viral evolution, the mutation rate is one of the critical parameters and the micro-level alteration in the mutation rate can change the characteristics of the virus and virulence dramatically in the host immune systems [13]. Therefore, a precise assessment of the mutation rate becomes significant to assume the risk of emerging infectious diseases like SARS- CoV-2 [14]. In addition, genomic sequence and mutation analysis are critical for the invention of efficacious drugs and vaccines against this RNA virus [15].

Deep neural learning has been successfully applied to a wide range of challenging prediction issues faced in the real world, such as time-series forecasting, drug discovery, medical image analysis, and diseases and disease subtypes classification [16,17]. Since it provides more accurate forecasts, deep learning is frequently considered as the most promising strategy for dealing with the noisy and chaotic nature of time series forecasting challenges [17,18]. Long short-term memory (LSTM), a specific recurring neural network structure (RNN), is one of the most effective and widely used deep learning algorithm [19]. LSTM algorithms can readily collect information on sequence patterns but they are meant to deal with periodic correlations and only employ features from the training dataset [17,18]. LSTM algorithm is well-suited for processing, classifying, and making predictions based on time series data [20]. LSTM has already been utilized in several COVID-19 studies, including mutational analysis [21–26].

In this current study, we aimed to perform a large scale study over the SARS-CoV-2 genome to identify the base substitution mutations along with the rate of the mutation using the available dataset in the Global Initiative on Sharing All Influenza Data (GISAID) database to address a crucial knowledge gap in the underlying biology of SARS-CoV-2 and estimate how predictable its evolution can be. From GISAID, we have analyzed the complete genome sequence of 259044 viral samples isolated from different countries for a certain period of December 2019 to December 2020. We concentrate particularly on mutations which have freely evolved on different dates because these are the possible opportunities for further adaptation of SARS-CoV-2 to its human host. We also predicted the effect of the identified missense mutations on their corresponding protein to trace the deleterious mutations that may interfere in viral infectivity and transmission. Afterwards, based on the findings from the mutational analysis, we intended to predict the mutation rate of the virus for future times through the deep learning approach lied on artificial recurrent neural network (RNN) called Long Short Term Memory (LSTM). It is anticipated that the present study will contribute to understand the evolving nature of SARS-CoV-2 in the human body which can ultimately determine how natural selection fixes or eliminates new SARS-CoV-2 variants from wild populations and will help to establish strategies to tackle the epidemiological and evolutionary levels. It may also aid in understanding some of the possible genetic constraints of the SARS-CoV-2 which are critical in the construction of evolution proof antiviral drugs and vaccines.

1.1. Related works

To forecast the COVID-19 pandemic, many researchers used the standard forecasting approaches along with statistical modeling. Based on the ARIMA (Autoregressive Integrated Moving Average) statistical analysis model, Ceylan [27] anticipated the pattern of COVID-19 prevalence in France, Spain, and Italy. The ARIMA technique is not optimal with such studies as COVID-19 data are nonlinear and nonstationary [28]. Car et al. [29] used an MLP-ANN (Multilayer Perceptron-Artificial Neural Network) model to predict the number of dead, recovered, and infected COVID-19 patients throughout the world. Salgotra et al. [30] employed GP (genetic programming) to estimate the possible impact of COVID-19 on confirmed and death cases in the most infected 15 countries between January (2020) and May (2020). Due to noisy time series data along with a lack of training data and appropriate features, the COVID-19 studies that relied only on machine-learning models suffered underfitting or overfitting concerns. These studies are limited to retrospective analysis and/or solely forecasting short-term trends [31–34]. Deep neural learning can readily address the aforementioned problems and provide more accurate predictions [17,35]. LSTM is one of the most successful deep learning algorithm and has already been effectively utilized in several COVID-19 studies including mutational analysis [21–26]. Based on the LSTM model, Pathan et al. [26] demonstrated that Thymine (T) and Adenine (A) are mostly mutated to other nucleotides whereas codon alteration is not as frequent as nucleotides. They also found a 0.1% increment in mutation rate for mutating of nucleotides from Thymine (T) to Cytosine (C) and Guanine (G), Cytosine (C) to Guanine (G) and Guanine (G) to Thymine (T). Conversely, they found 0.1% decrement for mutating of T to A, and A to C. This study used only 3068 SARS-CoV-2 samples with 2453 training and 614 testing data for the analysis. Mercatelli and Giorgi [36] observed an average of 7.23 mutations per SARS-CoV-2 isolates with a prevalence of SNP as the major mutational type. This study is also based on only 48,635 SARS-CoV-2 isolates. Based on the LSTM model, Chimmula and Zhand [23] predicted that around June 2020 COVID-19 outbreak will be ended in Canada. Their estimation was somewhat close, as the number of COVID-19 patients declined in May 2020 before a second wave occurred. Similarly, several studies predicted the spread of SARS-CoV-2 in India, New Zealand, Egypt, Kingdom of Saudi Arabia, United Arab of Emirates, Kuwait, Bahrain, Oman, and Qatar [37–40]. Maio et al. [41]

evaluated 140,000 SARS-CoV-2 genomes and observed that two specific mutations G to U (uracil) and C to U are equally amplified and substantially higher than all other mutation rates which cause the majority of mutations in the SARS-CoV-2 genome. The accuracy of the LSTM model increases with the increase number of training data as it can provide good memory by getting rid of the gradient vanishing and gradient explosion problems [42]. In this present study, based on bioinformatics and LSTM, we investigated the genetic variability of 259044 SARS-CoV-2 isolates alongside predicting the mutation rate of the virus in the future time and the effect of missense mutations on the corresponding protein was examined as well. The prediction accuracy of the LSTM model was 97% with 207236 training and 51808 testing data.

2. Methods

2.1. *Genomic data retrieval of SARS-CoV-2*

To investigate the genetic variations of SARS-CoV-2 virus genomes, we retrieved 259044 complete genome sequences covering all clades and submitted from December 01, 2019 to December 31, 2020 across all countries from the GISAID database [43]. The full-length sequences (>29000bp), as well as high nucleotide coverage (<1% Ns; <0.05% unique amino acid mutations), were considered for retrieving the SARS-CoV-2 genome sequences. Genomes associated with human infection were taken and the poor coverage (>5% Ns) genomes were excluded from the list. The reference genome sequence (NC_045512.2) of SARS-CoV-2 [44] containing 29,903bp in length was downloaded in FASTA format from National Center for Biotechnological Information (NCBI) Reference Sequence (RefSeq) database (https://www.ncbi.nlm.nih.gov/refseq/).

2.2. *Genomic variability analysis of SARS-CoV-2*

A Generic Feature Format (GFF3) annotation file associated to the SARS-CoV-2 reference sequence was downloaded from the NCBI SARS-CoV-2 Resources (https://www.ncbi.nlm.nih.gov/sars-cov-2/) and used to show the genomic sequences for all protein sequences of SARS-CoV-2. The ORF1 polyprotein was divided into its Non-structural proteins (NSPs) constituent such as NSP12. In the genome annotation, viral RNA-dependent RNA polymerase encoding NSP12 was considered as two regions called NSP12a and NSP12b. Alignment of total 259044 SARS-CoV-2 genome sequences was done by using the NUCMER v3.1 algorithm [45] where NC_045512.2 was considered as reference sequence. An R script [36] was utilized through R (version 4.0.5) software [46] to convert the output of the alignment to an annotated variant list that contain all the mutational events in nucleotide and protein level. The annotated variant list was then loaded through an in-house R script and checked whether there is any IUPAC (International Union of Pure and Applied Chemistry) code other than A, C, G, and T. If there is a different code, the list is fixed by removing it. Finally, the reference sequence was loaded along with the GFF3 annotation file of the reference sequence. The NUCMER object was then categorized according to SNPs, insertions and deletions. All the SNPs were then merged together in a separate file. Similarly, all the insertions and deletions were separated. Afterwards, we merged the NUCMER object again according to neighboring events (SNPs, insertions and deletions). Changes in query protein sequence according to variants were observed over the GFF3 annotation file of the reference sequence. Afterwards, Multiple Sequence Alignment (MSA) of the 10,000 SARS-CoV-2 sequences were done by the MAFFT [47] online server and a phylogenetic tree was then constructed by the same software to track the geographical distributions of SARS-CoV-2 isolates based on their mutational type. Two-cycle progressive method (FFT-NS-2) was used to perform MSA. In this approach, at first low quality all-pairwise distances are computed rapidly and then build a provisional MSA from where refined distances are derived and finally, the second progressive alignment is conducted [48]. Alongside, a scoring matrix was configured to 200PAM/k = 2 where gap opening penalty and offset value were set to 1.53 and 0.123, respectively. The aligned sequences were then utilized to construct a phylogenetic tree based on a distance based method "Neighbor-Joining" with a bootstrap consensus tree inferred from 1000 resampling. The phylogenetic tree was visualized and annotated by the NCBI Tree Viewer (https://www.ncbi.nlm.nih.gov/tools/treeviewer/). Finally, we determined the effect of the identified missense mutations on their corresponding protein through the PredictSNP [49], MAPP [50], Polyphen-1 [51], Polyphen-2 [51], SIFT [52], and SNAP [53] web servers. Mutations that were predicted as deleterious by all of these tools were considered as deleterious for their respective protein. The identified deleterious missense mutations were then further analyzed by the I-Mutant3.0 server (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi) to determine their impact on the structural stability of the corresponding proteins. Based on "DDG (Delta Delta G) Ternary Classification" prediction method, we classified the mutations in three categories: mutations having neutral ($-0.5 \leq$ DDG value ≤ 0.5) effect on protein stability, mutations that largely decrease the protein stability (DDG value < -0.5), and mutations that largely increase the protein stability (DDG value > 0.5).

2.3. *Deep neural learning and future mutation prediction*

A dataset having all the nucleotide mutation data from December 01, 2019 to December 31, 2020 was prepared and processed to predict future mutations based on machine learning approach. In this regard, each sample with single or multiple mutations was separated and labeled. For each mutation (A > T\C\G, T > A\C\G, C > A\T\G, and G > A\T\C), the mutation rate (%) was calculated for each sample by the following equation,

$$M^{i}_{X\neq Y}(X\rightarrow Y)=\frac{L^{i}_{X\neq Y}(X\rightarrow Y)\times 100}{N^{i}}$$

Here, M is the mutation rate (%) for the mutation X to Y (X, Y = any one of A, T, C, and G) in sample i, L is the number of occurrences of a certain mutation in sample i and N^i is the length of that sample.

Afterwards, the dataset that contained percentage based mutation data was used to select around 100000 genome samples randomly to run the model properly within our limited resources. The selected data were later divided to 80/20% as train and test sets. The train set was scaled by MinMaxScaler () function, a function of Scikit-learn machine learning library, and a time series generator was defined for the prediction of future mutations [26].

An artificial Recurrent Neural Network (RNN) called Long Short Term Memory (LSTM) Network was used to build the mutation prediction model. The model was trained with a TimeseriessGenerator, a tool of keras API used to automatically transform a univariate or multivariate time series dataset into a supervised learning problem, and compared against the Test set. The input layer of the model got the prepared set of training data with 200 neurons. Then it has been through a dense layer of 200 neurons with relu activation layer. After that 0.15 dropout has been used. A dense of 12 neurons has been used as an output layer. The model was trained in 100 epochs. Adam optimized and MSE (Mean Squared Error) loss function was used to train the model. Finally, the mutation rate (%) was predicted for the future 2000 SARS-CoV-2 variants.

3. Results

3.1. *Overall mutational profile of SARS-CoV-2 genome*

Worldwide a total number of 259044 complete genome sequences of SARS-CoV-2 were taken to investigate the overall genetic variations over the NC_045512.2 Wuhan reference genome and this analysis identifies a total of 3334545 mutations. Most of the samples possessed more than one mutation where 17221 samples were found to contain 18 mutations followed by the 17, 16, and 19 mutations for 17084, 14983, and 14801 samples, respectively. The highest number of 48 mutations were observed for 1 sample (Fig. 1). Average mutations per sample was calculated as 14.01. Among the 259044 complete genome sequences of SARS-CoV-2, the top 20 most mutated samples were identified, and we found that a sample from India had the maximum number of mutations (48). Out of 48 mutations, 11, 33, and 4 mutations were accounted as missense, silent and extragenic mutations, respectively. Samples from Scotland, United States of America (USA), Netherlands, Norway, and France were found to have up to 36 mutations (Fig. 2). 15 missense mutations along with 24 silent and 4 extragenic mutations were identified in the SARS-CoV-2 isolate collected from Scotland. The highest 21 missense mutations were identified in the isolate collected from Norway. The number of mutations identified in the isolates from India, Scotland, USA, Netherlands, Norway, and France are enlisted in Table S1. A total of 152 missense mutations were identified from these mostly mutated samples (Table S2). The effect of these 152 missense mutations on viral proteins is described in the “Mutational effect on viral proteins” section. After the onset of SARS-CoV-2 at Wuhan in China, 4 mutations were found in December 2019 (Fig. S1) and surprisingly from there, the number of mutations reached 25 in January 2020 (Fig. S2). We noticed that this number was increasing exponentially and reached 48 within December 2020. From the monthly basis of sequence analysis, we observed that some of the samples from March, August, November, and December 2020 have 40 mutations whereas 35 mutations were observed in June, July and October (Fig. S1-S13).

3.2. Analysis of nature of SARS-CoV-2 mutations

The nature of each of 3334545 mutations had been analyzed and a high prevalence of single nucleotide polymorphisms (SNPs) was identified worldwide over brief cases of deletion/insertion (indels). We observed a total of 1745775 (52.35% of the total) SNPs (missense mutations) where 1234456 (37.02% of the total) silent (synonymous) SNPs were found to fall over the coding regions. Moreover, 337340 (10.11%), and 10220 (0.30%) mutational events were identified in extragenic regions (5′ and 3′ untranslated region of the SARS-CoV-2 RNA sequence) and as deletion, respectively. Very small amount of stop codon creating SNP 4746 (0.14%) were observed followed by the in-frame deletions, in-frame insertion, and insertion which account for 1122, 260, and 518 of all the investigated mutational cases (Fig. 3). A similar type of mutational profile was also observed from monthly basis of sequence (December 2019 to December 2020) analysis which demonstrates that SNPs are the major mutational event followed by the silent SNP and extragenic mutations (Fig. S1-S13), assuming a conserved molecular mechanism for SARA-CoV-2 mutational evolution.

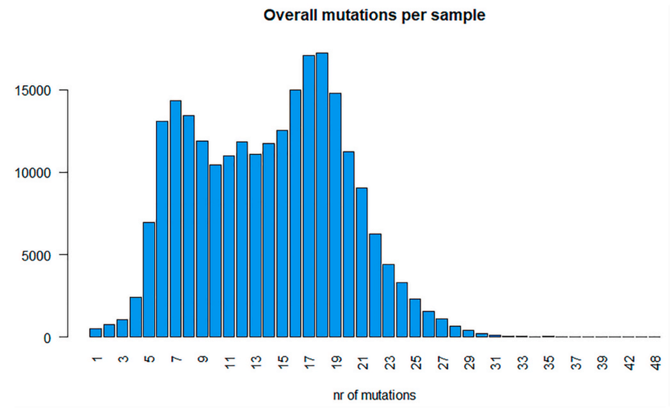

Fig. 1. The number of mutational events occurred from December 2019 to December 2020 for all the 259044 SARS-CoV-2 genome samples. The X-axis represents the number of mutations where Y-axis represents the number of SARS-CoV-2 genome samples. The highest 17221, 17084, 14983, and 14801 samples had 18, 17, 16, and 19 mutations per sample, respectively. One sample had the most mutations, with 48 in total.

3.3. Analysis of SARS-CoV-2 mutations according to their type

To observe the pervasiveness of SNP transitions (purine to purine and pyrimidine to pyrimidine) and/or SNP trans-version (purine to pyrimidine and pyrimidine to purine), SARS-CoV-2 mutations were classified based on their types. Worldwide the most common transition event was C > T transition, accounting for 1756440 (52.67%) of all the observed 3334545 SARS-CoV-2 mutations. The second most common mutation type was marked as G > T trans-version with 486610 occurrences (14.59%) where A > G transition was noticed as the third most common mutation type with 371334 (11.13%) cases. Transition (T > C, and G > A) and trans-version (G > C, C > G, C > A, and A > T) were remarked as 4th, 5th, 6th, 8th, 9th, and 10th common events of SARS-CoV-2 mutational evolution. A particular type of peculiar multi-nucleotide mutation (substitution of a GGG triplet with AAC) was observed as 7th common mutational type worldwide with 67596 incidents. Besides this, multi-nucleotide mutations like CC > TT, TG > CA, and AT > TA were also identified. The deletion of the ATG codon is the most common indel for SARS-CoV-2 and was found as the 16th common mutational type with 1231 events (Fig. S14). Our monthly basis mutation type analysis also revealed that the substitution of C with T is the most prominent alteration in every month from January 2020 to December 2020 worldwide (Fig. S2-S13). Moreover, a phylogenetic tree was constructed to track the geographical distributions of SARS-CoV-2 isolates based on their mutational type. The phylogenetic tree was clustered into three distinct clusters (Fig. 4). Isolates positioned in Cluster 1, Cluster 2, and Cluster 3 were found to show 47.63%, 50.16%, and 54.23% C > T transition in their genome. Conversely, G > T trans-version was noticed as 12.39%, 13.57%, and 15.01% for the isolates in Cluster 1, Cluster 2, and Cluster 3, respectively. 9.83% A > G transition was noticed for the isolates in Cluster 1, whereas 11.78% and 13.28% were observed for Cluster 2 and Cluster 3. Most of the countries in Europe are dominated by Cluster 1 and Cluster 2 followed by the USA, Africa, and Asia. On the contrary, Asia is dominated by Cluster 3 followed by the USA, Europe and Africa.

3.4. Analysis of SARS-CoV-2 mutations in nucleotide and protein level

We investigated the effect of each genetic variation on the viral protein sequences. The most predominant mutations were observed in the 23403th (A > G transition), 3037th (C > T transition), 14408th (C > T transition), and 241th (C > T transition) nucleotide position of SARS-CoV-2 genome (Fig. 5). The A23403G mutation causes a change from Aspartate (D) to Glycine (G) in protein position 614 (spike protein) which is responsible for the initial entry of the virus through the ACE2 receptor and is associated with the severity of COVID 19 [11]. The C14408T mutation substitute Proline (P) with Leucine (L) in position 314 of non-structural protein 12b (NSP12b), a RNA-dependent RNA polymerase (RdRp). Conversely, C3037T (F106F) mutation was found as a synonymous mutation in the region encoding NSP3, a viral predicted phosphoesterase whereas C241T mutation fall over non-coding regions (5′ UTR) (Fig. 6). Other common identified mutations are GGG28881AAC (RG203KR, in the nucleocapsid protein), C22227T (L93L, in the membrane protein), G29645T (A222V, spike protein), G21255C (A199A, in the NSP16), C28932T (V30L, in the ORF10), and T445C (A220V, in the nucleocapsid protein) (Figs. 5 and 6). Furthermore, the monthly basis of sequence analysis revealed that D614G (spike protein), F106F (NSP3), P314L (NSP12b), and 5′ UTR:241 mutations are at the top of the mutation analysis chart of every month from March 2020 to December 2020 (Fig. S4-S13).

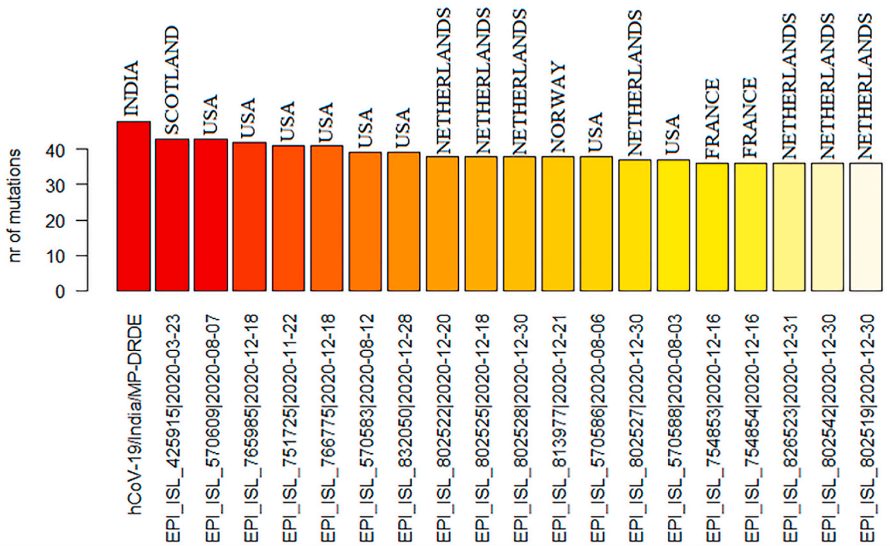

Fig. 2. Distribution of the most mutated SARS-CoV-2 samples throughout the world from December 2019 to December2020. The X-axis represents the most mutated SARS-CoV-2 samples and the country from where they were collected. The Y-axis represents the number of mutations identified in the samples. The sample from India had the highest 48 mutations where each of all other samples processed more than 36 mutations.

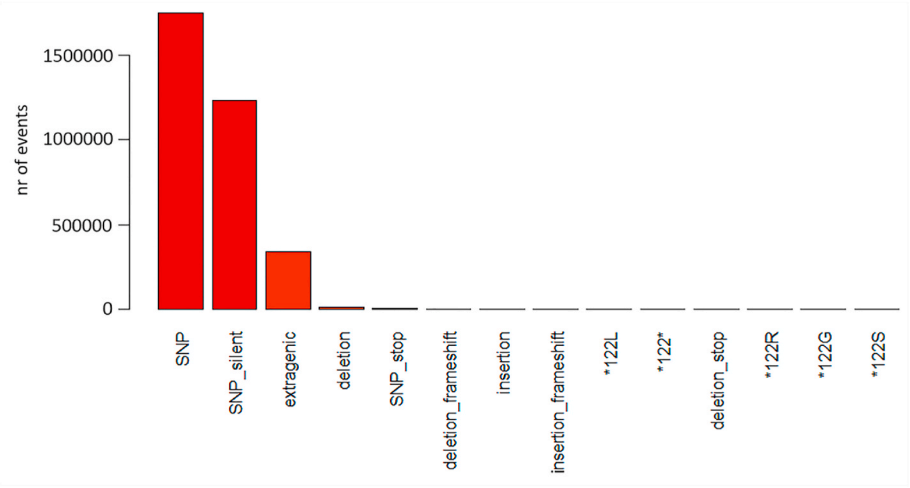

Fig. 3. Most frequent classes of SARS-CoV-2 mutations throughout the world. The Y-axis indicates the number of mutational events occurred in SARS-CoV-2 samples whereas X-axis indicates the classes of mutations observed for the mutational events. SNP is the most prevalent class of mutation followed by the silent and extragenic mutation.

3.5. *Mutational effect on viral proteins*

We identified 152 missense mutations throughout the genome of SARS-CoV-2. The effect of these mutations on viral proteins was determined and preliminary identified 46 missense mutations which may have a deleterious effect on their corresponding protein (Table S2). D1118H and C1243Y mutations have a deleterious effect on Spike protein. D3L, P13L, S183Y, S186Y, S194L, R262H, and D377Y have a deleterious effect on Nucleocapsid protein. A538T, M809L, P314L, A8D, S220G, A88V and L629F were found to have a deleterious effect on RNA-dependent RNA polymerase. Missense mutations [A890D, T1063I, G1433C, and T1456I], [N266I and H268L], R233C, [L89F, and V212F], F263S, [D315Y, A316T, and C444Y], [L183A, and G185E], T85I, L111K, and [A54T, and A74V] have a damaging effect on NSP3, NSP14, NSP6, NSP5, NSP15, NSP13, NSP16, NSP2, NSP4, and NSP8 proteins, respectively. Mutations [L5F, and R80I], S97I, and [L46C, V48G, G49I, Q57H, W131R, G172V, Q185H, and Y206S] were identified to have a deleterious effect on ORF7a, ORF8, and ORF3a proteins, respectively. Moreover, these 46 missense mutations were further analyzed to determine their effect on the respective proteins' stability (Table S3). Out of 46, 25 missense mutations (D1118H, S194L, R262H, M809L, P314L, A8D, S220G, A890D, G1433C, T1456I, R233C, F263S, L111K, A54T, A74V, L183A, A316T, V212F, L46C, V48G, Q57H, W131R, G172V, Q185H, and Y206S) were found to largely decrease (DDG value < -0.5) the structural stability of the corresponding proteins. Conversely, 3 missense mutations (D3L, L5F, and S97I) were found to largely increase (DDG value > 0.5) the structural stability of the corresponding proteins. The rest of the mutations were identified to have a neutral ($-0.5 \leq$ DDG value ≤ 0.5) effect on the protein stability.

3.6. *Prediction of mutation rate of future SARS-CoV-2 variants*

Long Short Term Memory (LSTM) Network was used to build the mutation prediction model through the training and testing process of the COVID-19 patient's sample. With 97% prediction accuracy, the LSTM model can predict the mutation rate of 250 future variants for

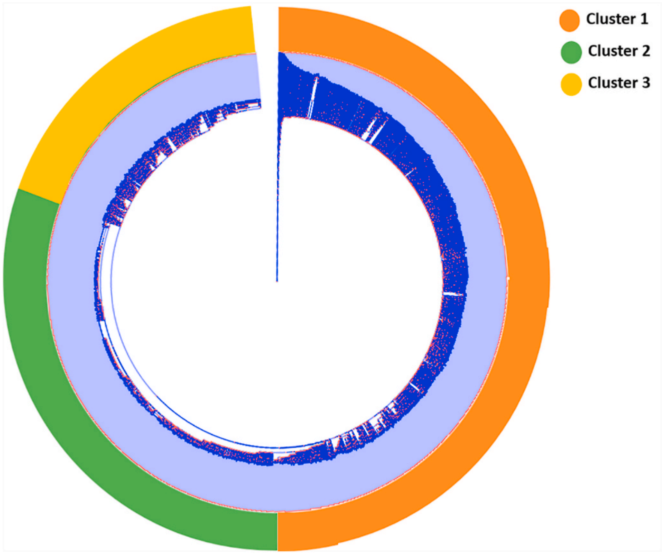

Fig. 4. Phylogenetic tree representing the distribution of the 10,000 SARS-CoV-2 isolates based on their mutational type. Variations that distinguish the branches of the tree are indicated as Cluster 1, Cluster 2, and Cluster 3. Deep blue color indicates the branches of the tree. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

98800 training samples at a time. To maintain the model size, we have added the immediately predicted samples at the bottom of the training samples and deducted the top 250 samples every time. Through this process, we predicted the nucleotide mutation rate (%) for 2000 future variants with 97% prediction accuracy. From this analysis, we observed that the nucleotide mutation rate between the 1st (Fig. 7) and 2000th (Fig. 8) future variants are highly deviated from each other. We noticed that the substitution of C with T is predominant in future variants and continuously increasing with the expansion of variants number and increased about 17% in the case of the 2000th variant (56%) than 1st (39%), suggesting a possible higher increment of C > T in future time. On the contrary, replacement of T with C was found to decrease about 7% in 2000th variant (2%) than 1st (9%), propounding a possible decrement of T > C in future time. In case of G > A and G > T substitutions, we perceived about 6% and 7% decrement respectively in 2000th variant when compared to 1st whereas 7% and 3% increment were respectively observed for A > G and A > T substitutions. We did not observe any noticeable alterations as T > G\A, C > G\A, and A > T\C from 1st to 2000th future variants (Fig. 9).

4. Discussion

Since SARS-CoV-2 is an RNA virus, it is evolving continuously in human populations over time which is fueling its massive worldwide transmission. Due to the genetic diversity of the virus along with the patient's genomic variations, the severity of COVID-19 greatly varies from patient to patient. A large proportion of patients either remain asymptomatic or show mild to moderate symptoms [54]. According to a cohort study, the average age of the patients who died after hospitalization was in 70s, with previous medical issues such as diabetes and obesity [55]. This difference in disease severity from one person to another is associated with multiple factors depending on the virus level, host genetic factors, and the host health condition level such as hypertension, diabetes, obesity and liver dysfunction [56–60]. The genomic sequence data offers a great opportunities to study the molecular changes in the expanding viral population by providing new insights into the mode of spread, diversity during the pandemics and the dynamics of evolutions [61]. The current study was intended to explore the genome wide accumulation of viral mutations at different time points to identify mutations which are occurring globally and to predict the mutation rate of the virus for future times through the artificial Recurrent Neural Network (RNN) called Long Short Term Memory (LSTM).

Mutational profiles of the 259044 SARS-CoV-2 isolates from December 2019 to December 2020 recognized a total of 3334545 mutations with an average of 14.01 mutations per sample. Each of the 17221, 17084, 14983, and 14801 samples were found to contain 18, 17, 16 and 19 mutations, respectively. Among the mostly mutated 20 samples, Indian sample had the maximum number of mutations of 48 followed by the samples (having up to 36 mutation) from Scotland, USA, Netherlands, Norway, and France. The occurrence of such a high amount of mutations in a large number of samples indicating a faster molecular evolution of SARS-CoV-2 which is probably responsible for making it more deadly in terms of time. Alongside this, we notice that the number of mutation is exponentially increased every month from the emergence of SARS-CoV-2 till December 2020 (48 mutations) which suggests that the virus maintains its mutating nature and continuously evolves in a random pattern. From the phylogenetic tree analysis, we observed that most of the countries in Europe, Asia, Africa and USA are dominated by the isolates having around 50% C > T transition, 14% G > T trans-version and 11% A > G transition. Analysis of the nature of SARS-CoV-2 mutations confirms a conserved molecular mechanism for

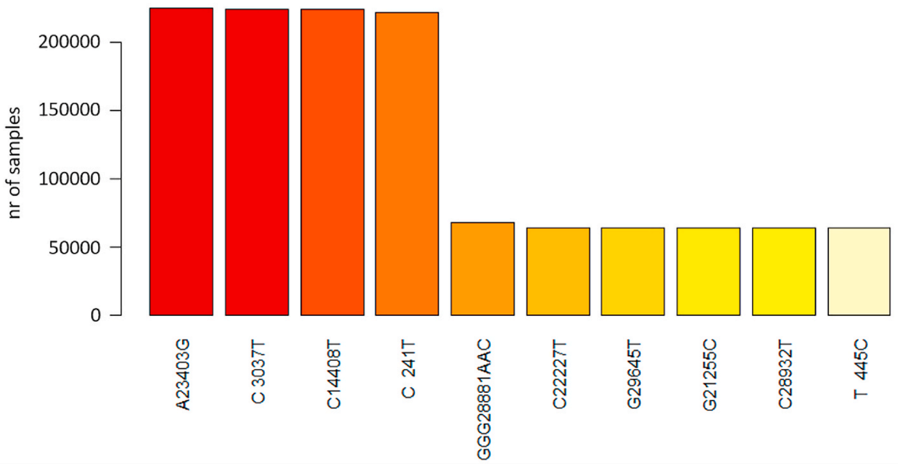

Fig. 5. Worldwide distribution of most frequent SARS-CoV-2 mutational events (annotated as nucleotide coordinates over the reference genome). The Y-axis represents the SARS-CoV-2 samples where X-axis represents the most frequent nucleotide substitution events found for the samples. The A23403G, C3037T, C14408T, and C241T are the most widespread mutations globally. A substantial multi-nucleotide mutation (GGG > AAC) is noticeable at the 28881 position.

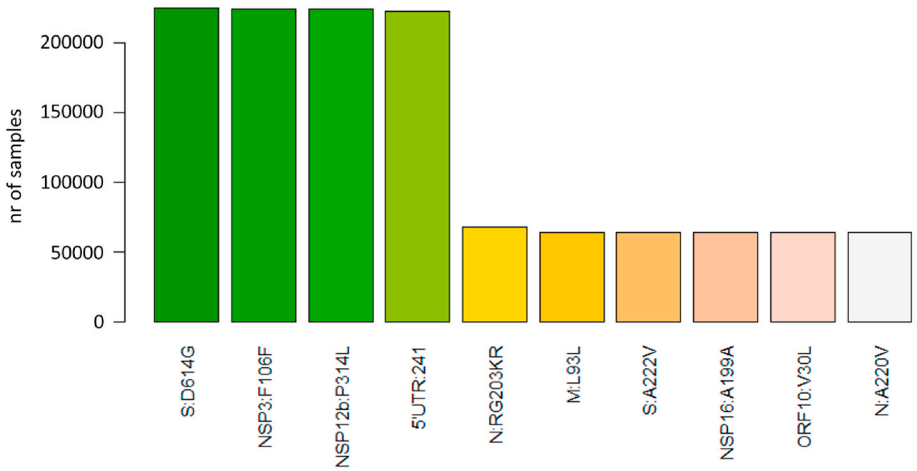

Fig. 6. Worldwide distribution of most frequent SARS-CoV-2 mutational events (annotated as amino acid coordinates over the reference genome). The Y-axis represents the SARS-CoV-2 samples where X-axis represents the most frequent amino acid substitution events found for the samples. The D614G, F106F, P314L, and C241T are the most widespread mutations globally. Here, S, N, NSP, ORF, and M mean the spike protein, nucleocapsid protein, non-structural protein, open reading frame, and membrane protein, respectively.

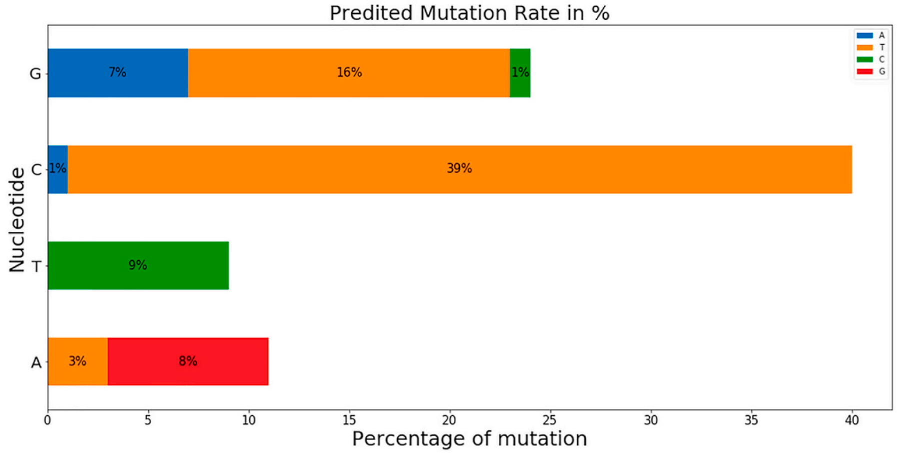

Fig. 7. Predicted nucleotide mutation rate for 1st future SARS-CoV-2 variant. The Y-axis represents the nucleotides that are substituted by the other nucleotides A (blue in color), T (orange in color), C (green in color), and G (red in color). The X-axis represents the mutation rate in percentage for the nucleotide substitution. The most predominant nucleotide substitution event is C to T which is accounted for 39%. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

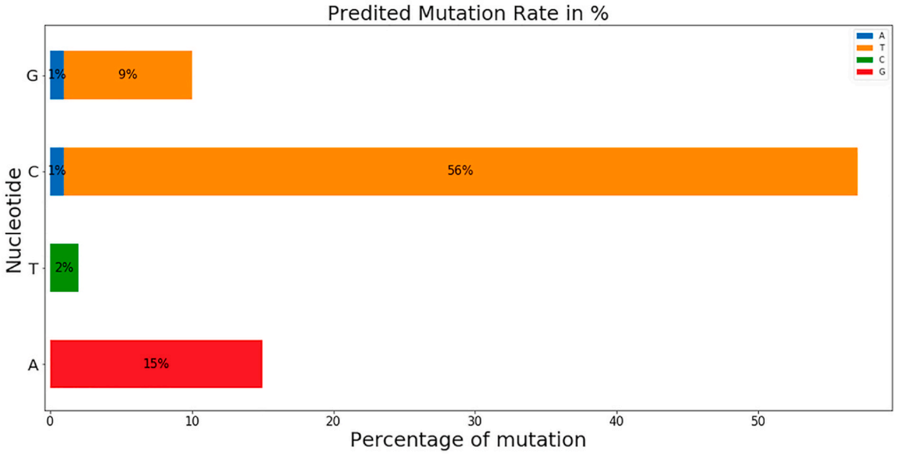

Fig. 8. Predicted nucleotide mutation rate for 2000th future SARS-CoV-2 variant. The Y-axis represents the nucleotides that are substituted by the other nucleotides A (blue in color), T (orange in color), C (green in color), and G (red in color). The X-axis represents the mutation rate in percentage for the nucleotide substitution. The most predominant nucleotide substitution event is C to T which is accounted for 56%. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

SARA-CoV-2 mutational evolution as the missense mutations (52.35%) are the most prevalent mutational events in terms of long time, followed by the silent SNPs (37.02%) and extragenic SNPs (10.12%). In our large scale study, previously reported D614G, and P314L missense mutations are also identified as the most prevalent mutation in the viral genome [36]. The P314L mutation in RdRp is associated with the D614G mutation and may favor the SARS-CoV-2 by enhancing its transmission ability [62]. Furthermore, we identified 152 missense mutations throughout the viral genome where 46 mutations were predicted as deleterious that may affect the viral protein structure, hence altering the stability of protein-protein interactions which ultimately can affect the viral entry into the host [11]. F106F mutation is found as predominantly occurring silent mutation in NSP3, suggesting a possible role in mRNA processing which might alter the nature of the viral protein [36,63].

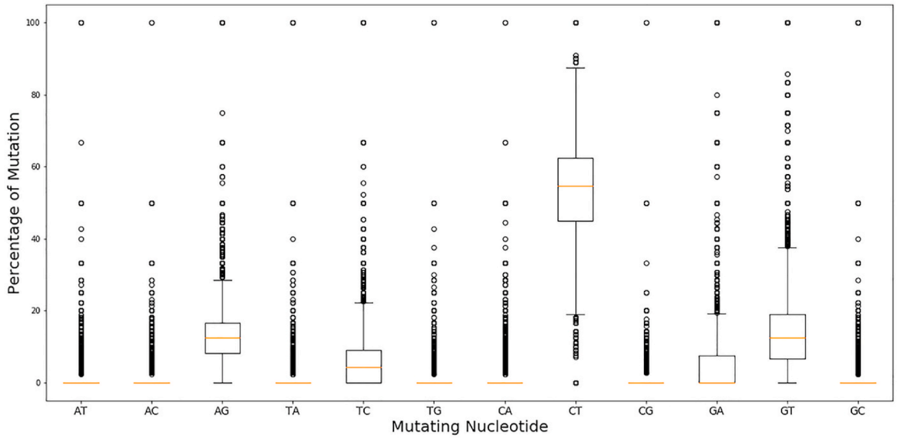

Fig. 9. Overall predicted nucleotide mutation rate for 2000 future SARS-CoV-2 variants. The X-axis represents the mutating nucleotides wherein every case the first nucleotide is substituted by the second nucleotide. The X-axis represents the mutation rate in percentage for the nucleotide substitution. For all the predicted 2000 future variants the most prevalent nucleotide substitution event is C to T followed by the G to T, A to G, T to C, and G to A.

Moreover, the 5′ UTR:C241T mutation might be associated with the transcription and replication rates of SARS-CoV-2 as it is found to occur most prominently [36,64]. Along with the previously found GGG > AAC mutation [36], our mutational type analysis identify some another multi-nucleotide mutations CC > TT, TG > CA, and AT > TA which are in the top 20 mutational type and should be monitored for the future as the GGG > AAC (R203K and G204R) reported to be associated with the insertion of a lysine in SR domain of N protein which might affect the phosphorylation [65]. Besides D416G, F106F, P314L, and 5′ UTR: C241T, our large scale analysis also identify C22227T;L93L (membrane protein), G29645T;A222V (spike protein), G21255C;A199A (NSP16), C28932T;V30L (ORF10), and T445C;A220V (nucleocapsid protein) mutations which are in top 10 mutations found in our investigation and should get importance in evaluating their role in the efficiency of SARS-CoV-2 transmission. D1118H, (S194L, and R262H), (M809L, P314L, A8D, and S220G), (A890D, G1433C, and T1456I), R233C, F263S, L111K, (A54T, and A74V), L183A, A316T, V212F, and (L46C, V48G, Q57H, W131R, G172V, Q185H, and Y206S) missense mutations were found to largely decrease the structural stability of the spike, nucleocapsid, RNA-dependent RNA polymerase, NSP3, NSP6, NSP15, NSP4, NSP8, NSP16, NSP13, NSP5, and ORF3a proteins, respectively and suggests that these missense mutations might decrease the infectivity of the virus. On the contrary, D3L, L5F, and S97I missense mutations were found to largely increase the structural stability of the nucleocapsid, ORF7a, and ORF8 proteins, respectively and suggests that these mutations might increase the viral infectivity. Whatever, since SARS-CoV-2 is continuously mutating, new strains will surely emerge due to natural selection. Our Long Short Term Memory (LSTM) recurrent neural network based future mutation prediction model claim that the virus will continue to mutate at a high rate which might give evolutionary advantage to the virus. Mutation rate analysis for future time for 2000 patients predicted a 17% increment of C > T with the expansion of patients number which suggests a possible higher increment of C > T in future time. Furthermore, 7% and 3% increment are respectively observed for A > G and A > T substitutions. At the same time, decrement is also noticed for T > C (7%), G > A (6%), and G > T (7%) mutations in future time. The appearance of new mutations may affect the development of new therapies and can even impair the adaptation of current therapies to get rid of the new SARS-CoV-2 variants. The appearance of new mutations may increase the viral transmission. For example, we notice from our monthly sequence analysis that after the emergence of coronavirus in December 2019, the virus with a large number of mutations (up to 30) were identified within 1 year in India, Scotland, USA, Netherlands, Norway, Israel, Italy, England and France, suggesting the fastest spread of SARS-CoV-2 subpopulation worldwide. Continuous tracking of mutations is the key to map the spread of the virus between individuals and throughout the world.

5. Conclusion

The COVID-19 has challenged the globe not just in regard to global health but also psychological and economic health. This challenge has been taken up by the scientific community to investigate the virus and its pathogenicity along with clinical management. In this present study through integrated strategies of bioinformatics and deep neural learning, we explained the genetic variability of SARS-CoV-2 strains as well as the mutation rate of the virus in the future and the impact of the identified mutations on the corresponding protein. This study paves a new way of predicting SARS-CoV-2 mutations including the evolution of the virus. This current methodology might be applied to study the frequency, recurrence, and possible effects of insertions and deletions. It would be interesting if the current study will be extended at the amino acid level to predict amino acid changing mutations for the future variants of SARS-CoV-2. The current approach can be applied to forecast the SARS-CoV-2 spread in different countries for comparative analysis. Even though we only evaluated genomic data of SARS-CoV-2, the same methodology might be implemented for other viruses. To the best our knowledge, this study will facilitate the tracking of mutations and help to map the spread of the virus worldwide. Extensive studies will be required to explore whether the identified mutations can exert any effect on the COVID-19 severity. There is still an opportunity to further improve the forecasting accuracy of this study by restructuring forecasting methods by adding more data.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contribution

M.S.H., M.M.R. developed the hypothesis. S.U.P., M.N.I., M.S.H., M.I.Q.T., M.I.R. performed the study. M.I.Q.T., S.U.P., A.F., O.S. wrote the manuscript. M.S.H., S.U.P., H.R., M.R., N.M.B., M.M.R. reviewed the manuscript. All authors approved the manuscript.

Declaration of competing interest

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.

Acknowledgments

The authors acknowledge the Department of Biotechnology and Genetic Engineering (BGE), Noakhali Science and Technology University and Computational Biology and Chemistry lab (CBC) for providing the research facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2021.100798.

References

[1] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R. A novel coronavirus from patients with pneumonia in China, 2019. New England journal of medicine; 2020.

[2] Yadav R, Chaudhary JK, Jain N, Chaudhary PK, Khanra S, Dhamija P, Sharma A, Kumar A, Handu S. Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19. Cells 2021;10:821.

[3] V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol 2021;19:155–70.

[4] Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020;18:1–9.

[5] Seyran M, Takayama K, Uversky VN, Lundstrom K, Palù G, Sherchan SP, et al. The structural basis of accelerated host cell entry by SARS-CoV-2. FEBS J 2020;288:5010–20.

[6] Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020;18:1–9.

[7] Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: evidence for virus evolution. J Med Virol 2020;92:455–9.

[8] Rubino S, Kelvin N, Bermejo-Martin JF, Kelvin D. As COVID-19 cases, deaths and fatality rates surge in Italy, underlying causes require investigation. J Infect Develop Countries 2020;14:265–7.

[9] Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, Ludden C, Reeve R, Rambaut A, Peacock SJ. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol 2021;19:409–24.

[10] MacLean OA, Orton RJ, Singer JB, Robertson DL. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evolution 2020;6. veaa034.

[11] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182:812–27. e819.

[12] Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, Southgate J, Johnson R, Jackson B, Nascimento FF. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 2021;184:64–75. e11.

[13] Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol 2010;84:9733–48.

[14] Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 2006;439:344–8.

[15] Ojosnegros S, Beerenwinkel N. Models of RNA virus evolution and their roles in vaccine design. Immunome Res 2010;6:1–14.

[16] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:20170387.

[17] Li J, Dai Q, Ye R. A novel double incremental learning algorithm for time series prediction. Neural Comput Appl 2019;31:6055–77.

[18] Zou W, Xia Y. Back propagation bidirectional extreme learning machine for traffic flow time series prediction. Neural Comput Appl 2019;31:7401–14.

[19] R. DiPietro, G.D. Hager, Deep learning: RNNs and LSTM, Handbook of medical image computing and computer assisted intervention, Elsevier2020, pp. 503-519.

[20] Singh NK, Suprabhath KS. HAR using Bi-directional LSTM with RNN, 2021 international conference on emerging techniques in computational intelligence (ICETCI). IEEE; 2021. p. 153–8.

[21] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM, Chaos. Solitons & Fractals 2020;140:110212.

[22] Yan B, Tang X, Liu B, Wang J, Zhou Y, Zheng G, Zou Q, Lu Y, Tu W. An improved method for the fitting and prediction of the number of covid-19 confirmed cases based on lstm, arXiv preprint arXiv:2005.03446. 2020.

[23] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos. Solitons & Fractals 2020;135:109864.

[24] Pereira IG, Guerin JM, Silva Júnior AG, Garcia GS, Piscitelli P, Miani A, Distante C, Gonçalves LMG. Forecasting Covid-19 dynamics in Brazil: a data driven approach. Int J Environ Res Publ Health 2020;17:5115.

[25] Wang P, Zheng X, Ai G, Liu D, Zhu B. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: case studies in Russia, Peru and Iran, Chaos. Solitons & Fractals 2020;140:110214.

[26] Pathan RK, Biswas M, Khandaker MU. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model, Chaos. Solitons & Fractals 2020;138:110018.

[27] Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020;729:138817.

[28] Du Y, Cai Y, Chen M, Xu W, Yuan H, Li T. A novel divide-and-conquer model for CPI prediction using ARIMA, Gray Model and BPNN. Procedia Comput. Sci. 2014;31:842–51.

[29] Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V. Modeling the spread of COVID-19 infection using a multilayer perceptron. Computational and mathematical methods in medicine 2020:2020.

[30] Salgotra R, Gandomi M, Gandomi AH. Evolutionary modelling of the COVID-19 pandemic in fifteen most affected countries, Chaos. Solitons & Fractals 2020;140:110118.

[31] Sun J, Chen X, Zhang Z, Lai S, Zhao B, Liu H, Wang S, Huan W, Zhao R, Ng MTA. Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. Sci Rep 2020;10:1–10.

[32] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman J, Yan P, Chowell Gb. Real-time forecasts of the COVID-19 epidemic in China from february 5th to february 24th, 2020. Infect. Dis. Model. 2020;5:256–63.

[33] Jia L, Li K, Jiang Y, Guo X. Prediction and analysis of coronavirus disease 2019. 2020. arXiv preprint arXiv:2003.05447.

[34] Yang Z, Zeng Z, Wang K, Wong S-S, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020;12:165.

[35] Zheng J, Fu X, Zhang G. Research on exchange rate forecasting based on deep belief network. Neural Comput Appl 2019;31:573–82.

[36] Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol 2020;11:1800.

[37] Chandra R, Jain A, Chauhan DS. Deep learning via LSTM models for COVID-19 infection forecasting in India, arXiv preprint arXiv:2101.11881. 2021.

[38] Kumar S, Sharma R, Tsunoda T, Kumarevel T, Sharma A. Forecasting the spread of COVID-19 using LSTM network. BMC Bioinf 2021;22:1–9.

[39] Ghany KKA, Zawbaa HM, Sabri HM. COVID-19 prediction using LSTM algorithm: GCC case study. Informat Med Unlocked 2021;23:100566.

[40] Saba AI, Elsheikh AH. Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Saf Environ Protect 2020;141:1–8.

[41] De Maio N, Walker CR, Turakhia Y, Lanfear R, Corbett-Detig R, Goldman N. Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol. Evol. 2021;13:evab087.

[42] Ma R, Zheng X, Wang P, Liu H, Zhang C. The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method. Sci Rep 2021;11:1–14.

[43] Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data–from vision to reality. Euro Surveill 2017;22:30494.

[44] Gorbalenya AE, Baker SC, Baric RS, de Groot -RJ, Drosten C, Gulyaeva AA, Haagmans B, Lauber C, Leontovich A, Neuman B. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol 2020;5:536–44.

[45] Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 2002;30:2478–83.

[46] Team RC. R: a language and environment for statistical computing [Internet]. R Foundation for Statistical Computing; 2018. http://www.r-project.org/.

[47] Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings Bioinf 2019;20:1160–6.

[48] K. Katoh, G. Asimenos, H. Toh, Multiple alignment of DNA sequences with MAFFT, Bioinformatics for DNA sequence analysis, Springer2009, pp. 39-64.

[49] Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 2014;10:e1003440.

[50] Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 2005;15:978–86.

[51] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.

[52] Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 2012;40:W452–7.

[53] Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, De Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24:2938–9.

[54] Callaway E, Ledford H, Mallapaty S. Six months of coronavirus: the mysteries scientists are still racing to solve. Nature 2020;583:178–9.

[55] Fajnzylber J, Regan J, Coxen K, Corry H, Wong C, Rosenthal A, Worrall D, Giguel F, Piechocka-Trocha A, Atyeo C. SARS-CoV-2 viral load is associated with increased disease severity and mortality. Nat Commun 2020;11:1–9.

[56] Tang D, Comish P, Kang R. The hallmarks of COVID-19 disease. PLoS Pathog 2020;16:e1008536.

[57] Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, Ogishi M, Sabli IK, Hodeib S, Korol C. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science 2020:370.

[58] De La Cruz M, Nunes DP, Bhardwaj V, Subramanyan D, Zaworski C, Roy P, Roy HK. Colonic epithelial angiotensin-converting enzyme 2 (ACE2) expression in blacks and whites: potential implications for pathogenesis Covid-19 racial disparities. J Racial Ethnic Health Dispar 2021:1–7.

[59] Guilger-Casagrande M, de Barros CT, Antunes VA, de Araujo DR, Lima R. Perspectives and challenges in the fight against COVID-19: the role of genetic variability. Front Cell Infect Microbiol 2021;11:150.

[60] Trump S, Lukassen S, Anker MS, Chua RL, Liebig J, Thürmann L, Corman VM, Binder M, Loske J, Klasa C. Hypertension delays viral clearance and exacerbates airway hyperinflammation in patients with COVID-19. Nat Biotechnol 2021;39:705–16.

[61] Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene reports 2020;19:100682.

[62] Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G-W. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun Biol 2021;4:1–14.

[63] Dickson ET, Hyman P. Brenner's encyclopedia of genetics. second ed. ScienceDirect: Elsevier; 2013.

[64] Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell 2020;181:914–21. e910.

[65] Ayub MI. Reporting two SARS-CoV-2 strains based on a unique trinucleotide-bloc mutation and their potential pathogenic difference. 2020.