Chaos, Solitons and Fractals 138 (2020) 110018

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

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

Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model

Refat Khan Pathan[a], Munmun Biswas[a], Mayeen Uddin Khandaker[b,*]

[a] Department of Computer Science and Engineering, BGC Trust University Bangladesh, Chittagong-4381, Bangladesh
[b] Centre for Biomedical Physics, School of Healthcare and Medical Sciences, Sunway University, 47500 Bandar Sunway, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 22 May 2020
Accepted 12 June 2020
Available online 13 June 2020

Keywords:
SARS-Cov-2
Gene sequence
Mutation rate
Neural Network
LSTM model

ABSTRACT

SARS-CoV-2, a novel coronavirus mostly known as COVID-19 has created a global pandemic. The world is now immobilized by this infectious RNA virus. As of June 15, already more than 7.9 million people have been infected and 432k people died. This RNA virus has the ability to do the mutation in the human body. Accurate determination of mutation rates is essential to comprehend the evolution of this virus and to determine the risk of emergent infectious disease. This study explores the mutation rate of the whole genomic sequence gathered from the patient's dataset of different countries. The collected dataset is processed to determine the nucleotide mutation and codon mutation separately. Furthermore, based on the size of the dataset, the determined mutation rate is categorized for four different regions: China, Australia, the United States, and the rest of the World. It has been found that a huge amount of Thymine (T) and Adenine (A) are mutated to other nucleotides for all regions, but codons are not frequently mutating like nucleotides. A recurrent neural network-based Long Short Term Memory (LSTM) model has been applied to predict the future mutation rate of this virus. The LSTM model gives Root Mean Square Error (RMSE) of 0.06 in testing and 0.04 in training, which is an optimized value. Using this train and testing process, the nucleotide mutation rate of 400th patient in future time has been predicted. About 0.1% increment in mutation rate is found for mutating of nucleotides from T to C and G, C to G and G to T. While a decrement of 0.1% is seen for mutating of T to A, and A to C. It is found that this model can be used to predict day basis mutation rates if more patient data is available in updated time.

1. Introduction

The whole world is suffering by an ongoing pandemic due to Coronavirus disease brought by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. It was an outbreak from Wuhan, the capital of Hubei province in China during December 2019 [2]. The virus was identified on 7th January and observed that it is spread by human-to-human transmission via droplets or direct contact [3,4]. Its infection has been estimated to be a mean incubation period of 6.4 days and a basic reproduction number of 2.24-3.58. Since its identification, it has already been spread speedily over the whole globe, therefore the world health organization (WHO) had declared COVID-19 a global pandemic on 11th March 2020 [5].

The SARS-CoV-2 is a pathogenic human coronavirus under the Betacoronavirus genus. In the recent decade, the other two pathogenic species SARS-CoV and MERS-CoV were outbreaks in 2002 and 2012 in China and the Middle East, respectively [6–9]. The complete genomic sequence (Wuhan-HU1) of this large RNA virus (SARS-CoV-2) was first discovered in the laboratory of China on 10th January [10] and placed in the NCBI GenBank. The SARS-CoV-2 is a single positive-stranded RNA virus having non-segmented in nucleic acid sequence [11]. Although it is an RNA virus but for simplicity of understanding the gene sequence has been given as DNA type which means nucleobase Uracil (U) has been replaced by Thymine (T). The genomic sequence of SARS-CoV-2 virus shows about 79% and 50% similarity with the SAR-CoV and MARS-CoV, respectively [6].

SARS-CoV-2 performs mutation during replication of genomic information [12]. The mutation occurs due to some errors when copying RNA to a new cell. Mutations are mainly three kinds: Base substitution, Insertion, and Deletion. Further, in base substitutions, there are some more divisions: silent, nonsense, missense, and frameshift [13]. Micro-level alteration of mutation rate is also de-

* Corresponding author. Centre for Biomedical Physics, School of Healthcare and Medical Sciences, Sunway University, 47500 Bandar Sunway, Selangor, Malaysia.
E-mail addresses: mu_khandaker@yahoo.com, mayeenk@sunway.edu.my (M.U. Khandaker).

https://doi.org/10.1016/j.chaos.2020.110018
0960-0779/

tectable for virus infection in host immune systems and drastically change the virus characteristic and virulence. To understand viral evolution, the mutation rate is one of the crucial parameters [14]. Furthermore, it is one of the most important factors for the assessment of the risk of emergent infectious disease, like due to SARS-CoV-2. Therefore, an accurate estimation of this parameter finds a great significance [15,16].

In connection to this and following the current pandemic, many researchers and scientists are working relentlessly to understand the evolution of SARS-CoV-2. Asim et. al have performed Phylogenetic analysis of SARS-CoV-2 virus based on the spike gene of the genomic sequence [17]. In this study, they described a detailed genomic sequence of the SARS-CoV-2 virus. They identified the factor of endemicity of SARS-CoV-2 and then focused on to find out the next reservoir of the SARS-CoV-2 virus. Based on the case study, the authors reported that all sequence of this virus is constituted in a single cluster without making any branching on it but not validated the findings with detailed statistical analysis. An analysis on Gene signature of SARS-CoV-2 virus has been performed by Ranajit and Sudeep [18]. They estimated the ancestry rate of the European genome from the reference population by applying a statistical tool qpAdm. Then they applied Pearson's correlation coefficient between various ancestry rate of European genome and performed statistical analysis on death/recovery ratio by using GraphPad Prism v8.4.0, GraphPad Software. In this study, they developed different linear regression models. Finally, they performed Genome-wide association analyses (GWAS) among European and East Asian genomes to examine single-nucleotide polymorphism (SNP) which is correlated to the infection of the SARS-CoV-2 virus. From the SNP association, they observed a huge difference in allele frequencies between European and Eastern Asian countries. Debaleena et al. analyzed the statistical changes of signature from different variant of SARS-CoV-2 virus [19]. They calculated diversity, non-synonymous, synonymous, and substitution rates for each gene of the nucleotide sequence by using DnaSP. They also employed time zone software for phylogenetic analysis and mutation detection for each gene. After that, they compared the sequence alignment of a protein of Wuhan and India by using multiple sequence alignment. Note that, in their study, the mutation rate was not calculated based on the patient's genomic sequence. However, the contemporary literature shows adequate studies on the genomic sequence but very few studies on the mutation rate. Therefore, the present study is designed to perform the mutation rate prediction for SARS-CoV-2 on the basis of the time series.

Unfortunately, the current data shows that the SARS-CoV-2 virus is highly infectious than the other harmful species of pathogenic human coronaviruses [20]. World populations are now suffering and are in great anxiety by observing the deadliest effect of this virus. But what can be done to stay healthy or avoid getting infected with the virus is still undiscovered. To stop SARS-CoV-2 virus, there is a critical need to invent proper vaccine and antibody based therapy against this virus [21]. Scientists and Researchers are trying their best to discover suitable drugs or vaccines to neutralize the effect of this virus on the human body, or at least in helping to create an effective resistance against the spreading out of this virus. For inventing proper drugs and vaccines against COVID-19 RNA viruses, genomic sequence and mutation analysis are crucially required [22]. In fact, accurate information on the viral mutation rate may play a vital role in the assessment of possible vaccination strategies.

In this regard, we performed a detailed study on the mutation rate of this virus using the available dataset in the NCBI GenBank. From this dataset, we have analyzed the Genomic sequence of 3408 patients from different countries for a period of 12th January to 11th May 2020. We focus specifically on the mutations that have developed freely on different dates (homoplasies) as these are likely possibilities for progressing adjustment of SARS-CoV2 to its novel human host. Specifically, we have calculated the base substitution mutation rates. Due to the lack of necessary information for insertion and deletion, we have considered those as substitution mutations to ensure that no nucleotide goes out of count. It is expected that the present analysis would help to understand the changing behavior of this virus in the human body and set up strategies to combat the epidemiological and evolutionary levels.

2. Dataset analysis and preprocessing

An adequate amount of gene dataset is currently available in the NCBI GenBank which contains the complete genome sequence of SARS-CoV-2. Among the many entities, we have filtered the gene sequence, date of collection, and country of the sample. All genes are taken from the human body who are affected by COVID-19. There are genes from almost 33 countries but China, Australia and the United States has a considerable number of patients' data. Though some countries like England, Italy, France, Spain, and Brazil has a very high mortality rate but for the lack of available data in the NCBI GenBank till 15th May, we were unable to calculate the mutation rates for these countries separately. Therefore, we have considered these countries along with others those have low gene data sequence available in the GenBank as the rest of the World category to cover as much region as possible. Table 1 shows the information about the gene dataset.

In this dataset, there are also some partial genes. So we filtered them and take only with the level of the complete genome. There is a reference gene sequence of length 29903. Finally, we have filtered the dataset by taking a maximum gene length of 29903 and a minimum of 29161, and avoided the copy sequences. With this filtering process, the total number of patients come down to 3068 from 3408, patients from China come down to 40 from 86, The United States come down to 1903 from 2103 and Australia come down to 918 from 925. Following the size of the available dataset, the mutation rate calculations have been set for four categories: China, the United States of America (USA), Australia and the rest of the World. Furthermore, the dataset is arranged in a suitable way to separately calculate the nucleotide mutation and codon mutation. The first filtered dataset is to find the nucleotide mutation rate. Then we have converted the four raw nucleotides (A = Adenine, T = Thymine, C = Cytosine and G = Guanine) into codon set. A codon consists of three nucleotides and forms a unit of genetic code in DNA or RNA. Information given in Table 2 is used to convert the gene sequence by their index number. For example, if three consecutive nucleotides are 'TTT' then it will be converted

Table 1
Number of patients in 33 countries.

Country	Number of Patient	Country	Number of Patient
Australia	925	Malaysia	4
Brazil	3	Nepal	1
China	60	Netherlands	1
Colombia	1	Pakistan	2
Czech Republic	7	Peru	1
Finland	1	Puerto Rico	13
France	1	Serbia	1
Germany	1	South Africa	1
Greece	4	South Korea	1
Hong Kong	20	Spain	12
India	46	Sri Lanka	4
Iran	2	Sweden	1
Israel	2	Taiwan	22
Italy	2	Thailand	23
Japan	5	Turkey	1
Kazakhstan	4	United States	2103
Vietnam	2		

Table 2
Codon indexing to change raw nucleotides.

	T	C	A	G	
T	1. TTT	5. TCT	9. TAT	13. TGT	T
	2. TTC	6. TCC	10. TAC	14. TGC	C
	3. TTA	7. TCA	11. TAA	15. TGA	A
	4. TTG	8. TCG	12. TAG	16. TGG	G
C	17. CTT	21. CCT	25. CAT	29. CGT	T
	18. CTC	22. CCC	26. CAC	30. CGC	C
	19. CTA	23. CCA	27. CAA	31. CGA	A
	20. CTG	24. CCG	28. CAG	32. CGG	G
A	33. ATT	37. ACT	41. AAT	45. AGT	T
	34. ATC	38. ACC	42. AAC	46. AGC	C
	35. ATA	39. ACA	43. AAA	47. AGA	A
	36. ATG	40. ACG	44. AAG	48. AGG	G
G	49. GTT	53. GCT	57. GAT	61. GGT	T
	50. GTC	54. GCC	58. GAC	62. GGC	C
	51. GTA	55. GCA	59. GAA	63. GGA	A
	52. GTG	56. GCG	60. GAG	64. GGG	G

Fig. 1. Nucleotide to codon indexing.

to 1, 'GCT' is converted into 53, and so on. The conversion process has been shown in Fig. 1. This process is important to understand the mutation in the codon sequence of COVID-19. Also, it helps to lower the computational complexity.

3. Gene mutation

Gene mutates for many reasons. When RNA tries to copy genetic codes form DNA it may cause some error which causes mutation. Also, errors in DNA replication, recombination, and chemical damage in DNA or RNA causes mutation. There are basically three types of mutations: base substitutions, deletions, and insertions. From this dataset, we can find out the three kinds of substitution mutation which are silent, missense, and nonsense. A silent mutation is the change of codon by which the resulting amino acid remains unchanged. If the resulting amino acid changes then it is called a missense mutation. On the other hand, when changing codon produces the stop signal for gene translation which causes a nonfunctional protein then it is called a nonsense mutation. These three types of substitution mutation of the observed dataset have been shown in Fig. 2, where the missense rate is 34.3%, the nonsense mutation rate is 6.7% and the silent mutation rate is 0.8%.

3.1. Nucleotide mutation

If the mutation type is missense then it can be said that the change of nucleotide has affected the protein generation, which

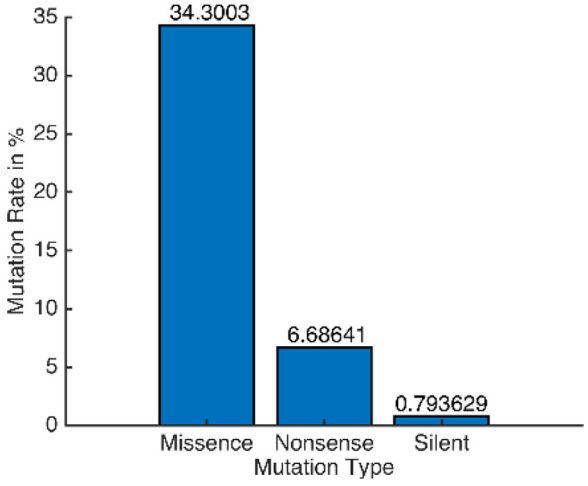

Fig. 2. Substitution Mutation Rate.

may change the behavior of the virus. Also, it is hard to identify the cure's gene sequence. The missense nucleotide mutation rate has been calculated by the given algorithm in Fig. 3. After using this algorithm Eq. (1) has been used to convert the values in percentage.

$$MutationRate = \left(\frac{mutation}{lg \times gs} \right) \times 100 \tag{1}$$

Here, *MutationRate* is the final output array, *mutation* is the output array sized 4×4 containing raw values after applying the algorithm, *lg* is the length of a dataset which is 3068 for the full dataset, 40 for China, 918 for Australia and 1903 for the USA, *gs*

Input: Dataset with patient in rows and nucleotide in columns.
Output: A 4×4 mutation matrix.

```
let mutation[1:4,1:4]=0
for i = 1 to (len (dataset)) do
  for j = 1 to (len (ReferenceGene)) do
    let D1= dataset[i][j]
    let D2= reference[j]
    if D1!=D2 then
      mutation[D1][D2] ← mutation[D1][D2]+1
    end if
  end for
end for
```

Fig. 3. Algorithm for calculating nucleotide mutation rate.

is the length of reference gene sequence which is 29903 in this dataset.

In this process, we have calculated the nucleotide mutation rate for the prepared dataset. The mutation rate for China has been shown in Fig. 4(a). It shows that a huge percent of Thymine (T) are being mutated to other nucleotides but not producing the same amount of T again. Also, a huge amount of Adenine (A) is mutated to other nucleotides. Comparing to T and A, Cytosine (C) and Guanine (G) were not changed much.

After that, the mutation rate has been calculated for Australia and the USA, and shown in Fig. 4(b) and (c). This is clear that all rates have a common factor of having the high mutation rate of T and A. But there is a significant increase in the mutation rate compared to China. This clearly indicates that this virus is very much active in changing its gene sequence. Finally, the nucleotide mutation for the full dataset of 33 countries has been shown in Fig. 4(d). It shows that C and G mutation rates are almost equal to the USA because there are more data of USA than any other countries. But some changes in T and A can be seen for the dataset from the rest of the World. These values vary on the availability of the data from different countries.

3.2. Codon mutation

The second processed and converted dataset that were prepared previously has been used here to calculate the codon mutation rate, and shown in Fig. 5. Changes in nucleotide cause changes in codon set, which later affects the protein directly. We have used the same algorithm shown in Fig. 3 for detecting the codon mu-

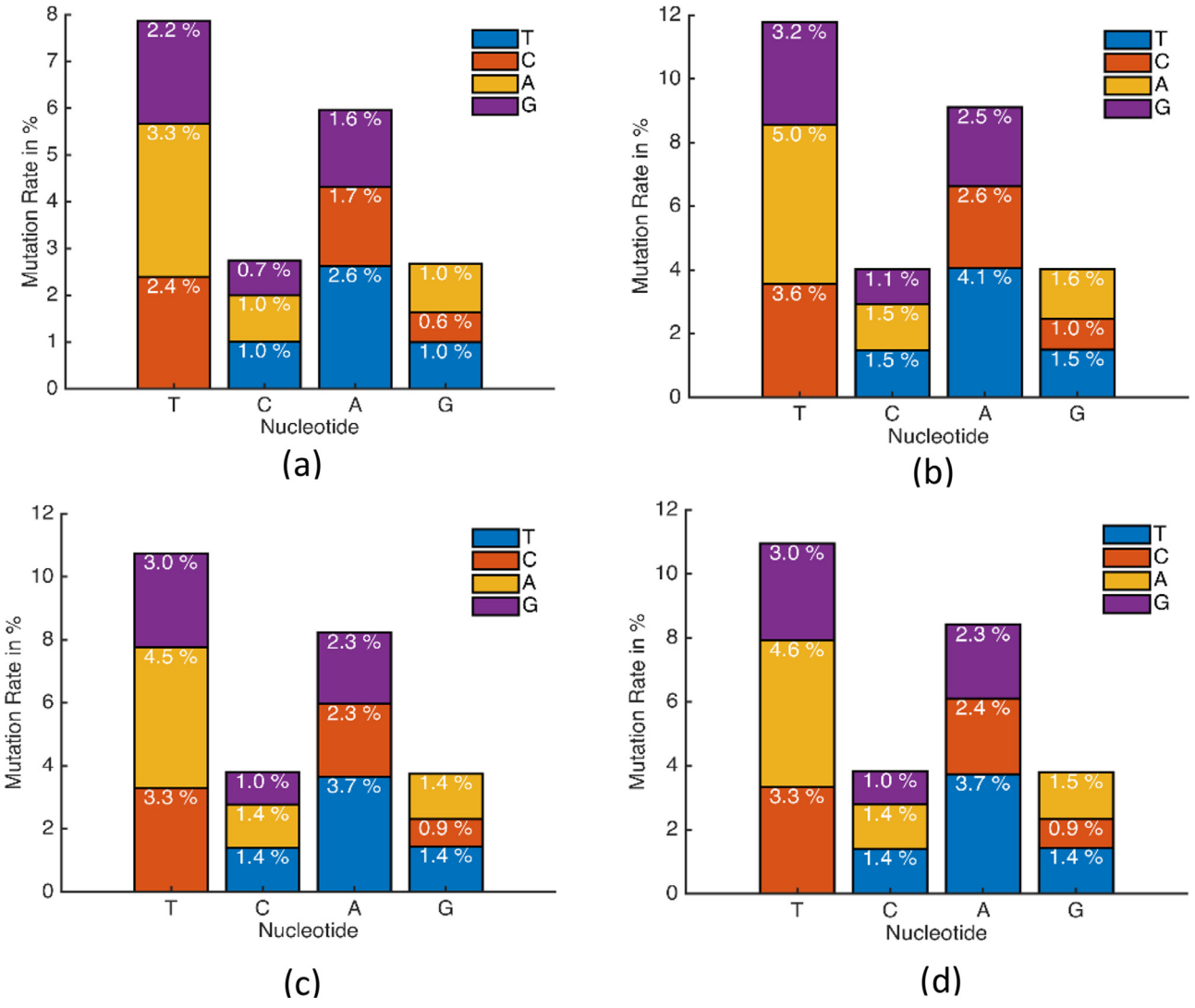

Fig. 4. Nucleotide mutation rate for (a) China, (b) Australia, (c) The USA, (d) Rest of the World.

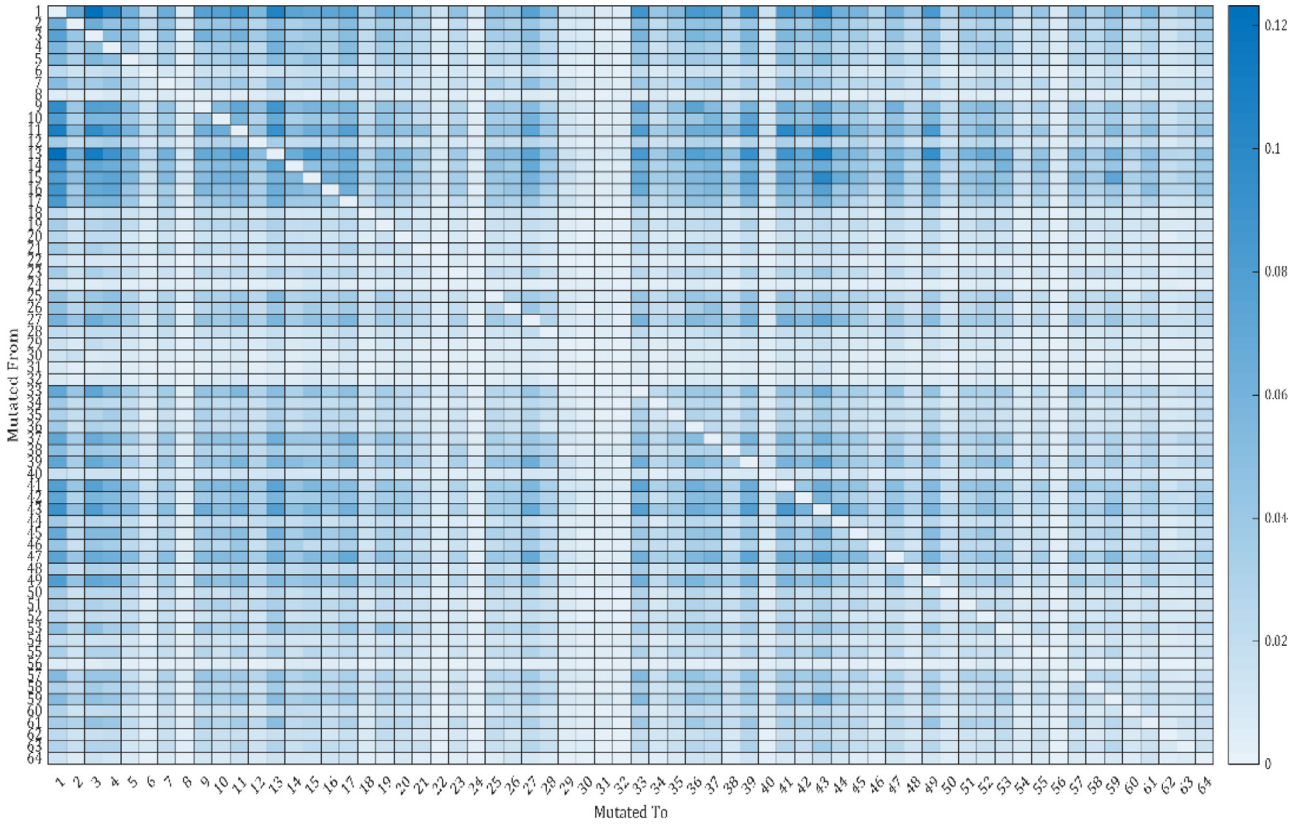

Fig. 5. Codon mutation rate for the full dataset. X and Y axis ticks are numbered following the sequence shown in Table 2.

tation rate. A small change has been made in the receiving array where array size was 4 × 4 for nucleotide but here the array size is 64 × 64 for codon mutation. After finding the codon mutations, Eq. (2) has been used to get the rates in percentage.

$$CodonMutation = \left(\frac{mutation}{\lg \times gs}\right) \times 100 \tag{2}$$

Here, *CodonMutation* is the final output array, *mutation* is the output array sized 64 × 64 containing raw values after applying the algorithm, *lg* is the length of dataset which is 3068, *gs* is the length of the reference gene sequence which is 9967 in this converted dataset.

The codon mutation rate for the full dataset has been shown in Fig. 5. From the obtained value it is clear that codons are not frequently mutating like nucleotides. The diagonal values are 0 because that point codons are not changing comparing with reference gene sequence and highest codon mutation rate is 0.12%.

4. Predicting nucleotide mutation rate

In processed nucleotide mutation dataset, we have gene data from 12th January 2020 to 11th May 2020 discontinuously. These dates are in sorted ascending order which makes it easy to consider this as a time series dataset. In one particular date, this dataset has one or more patients. 3068 patients are in this dataset for 62 days. By taking all the patient, we can find a time series dataset for patients shown in Fig. 6.

To obtain a day basis time-series dataset, we have averaged the mutation rates for different patients in the same date. So the dataset becomes smaller and dates are in non-sequentially increasing order and the mutation rates for 62 days have shown in Fig. 7. The low date availability makes it difficult to train a model in such a small amount of data.

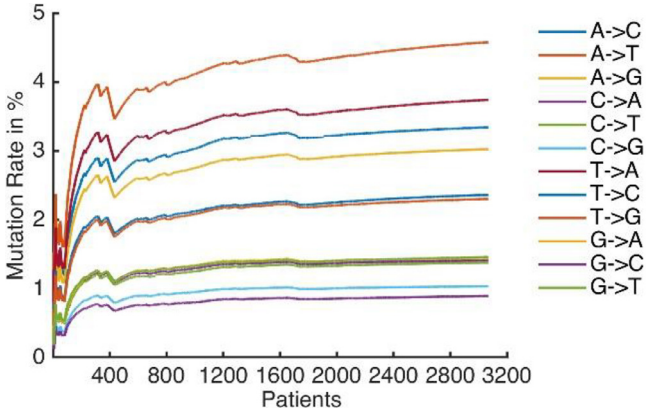

Fig. 6. Time series dataset based on patients.

Table 3
RNN data preparation.

Data - Shape(12 × 12)	Label - Shape(1 × 12)
Mutation set {1, 2, 3 .....11, 12}	Mutation set 13
Mutation set {2, 3, 4 .....12, 13}	Mutation set 14
Mutation set {3, 4, 5 .....13, 14}	Mutation set 15
Mutation set {n-12, n-11, n-10 ..... n-2,n-1}	Mutation set n

Long Short Term Memory network which is a type of recurrent neural network (RNN) has been used in deep learning. Data has been organized as shown in Table 3 where each set has a mutation rate of 12 patients.

All of the data have been divided to 80/20% as training and testing. Therefore, we get 2453 data for training and 614 for testing. An LSTM model has been created with keras, a deep learning API of

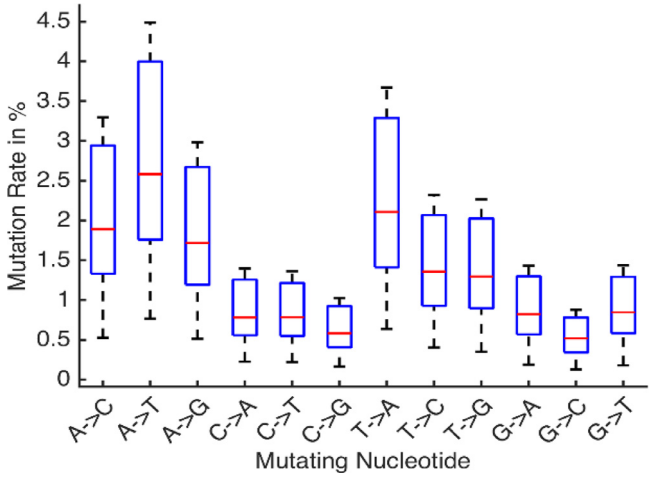

Fig. 7. Mutation rates for 62 days.

python and the structure has shown in Fig. 8 to train the dataset. First, the input layer got the prepared set of training data with 500 neurons. Then it has been through a dense layer of 250 neurons with relu activation layer. After that 0.25 dropout has been used. Another dense layer of 150 neurons has been used with relu activation. Then again 0.25 dropout is used. Finally, dense of 12 neurons has been used as an output layer with adam optimizer. This model gives Root Mean Square Error (RMSE) of 0.06 in testing and 0.04 in training.

After the train and testing process, the model found to be working in the expected level. So we used the last 12 patients' mutation rate to predict one future patient's mutation rate and then take that patient and again make 12 patients' mutation rate by 11 old and 1 new patient and predicted. By this procedure, we have predicted 400 patients' mutation rate for future time, as shown in Fig. 9.

The nucleotide mutation rate of 400th patient in future time has been shown in Fig. 10. A little increment of mutation rate can be seen. If more continuous data can be found from different locations and date then this method can be applied to find the mutation rate for one particular date in the future.

Input Layer	Input:	[(None,12,12)]
	Output:	[(None,12,12)]
LSTM	Input:	(None,12,12)
	Output:	(None,500)
Dense	Input:	(None,500)
	Output:	(None,250)
Activation	Input:	(None,250)
	Output:	(None,250)
Dropout	Input:	(None,250)
	Output:	(None,250)
Dense	Input:	(None,250)
	Output:	(None,150)
Activation	Input:	(None,150)
	Output:	(None,150)
Dropout	Input:	(None,150)
	Output:	(None,150)
Dense	Input:	(None,150)
	Output:	(None,12)

Fig. 8. LSTM architecture for creating model.

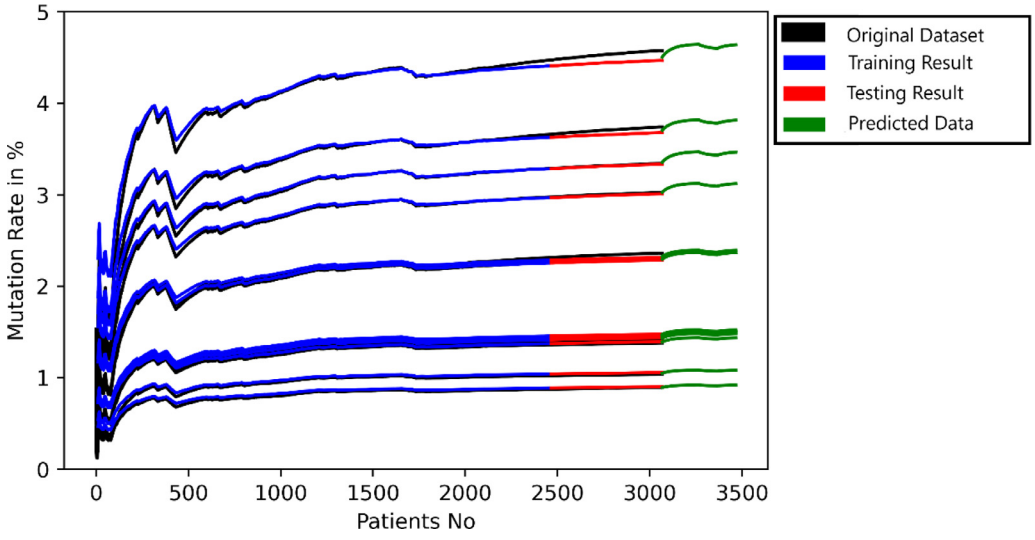

Fig. 9. Prediction of nucleotide mutation rate.

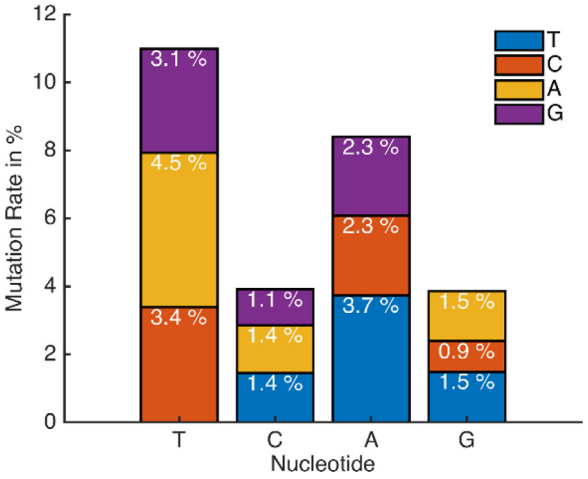

Fig. 10. Predicted mutation rate for 400th future patient.

4.1. Comparison with available literature

It is worth mentioning that the present study is unique in regard to our methodology of mutation rate analyses. We have developed a suitable program/algorithm (See in Fig. 3), and then used this algorithm together with the recurrent neural networking concept. By using the Long Short Term Memory (LSTM) model, we have analyzed the mutation rate of SARS-CoV-2 obtained from the current patient and predict a scenario for the future patient. Note that there are many varieties of gene mutation studies in literature, however as far as our knowledge, no earlier study detect/analyses the mutation rate for Covid 19 using a similar technique to this study. Since only mutation analyses together with different techniques is available for other types of viruses/diseases, therefore this study did not produce any comparison for the obtained results.

5. Conclusion

The COVID-19 pandemic has almost immobilized the world in this twenty-first century. The great spreading power mixing with mutation turns this virus very difficult and deadly, and the cumulative incidence of COVID-19 is rapidly increasing day-by-day. Lockdown has limited the spreading power of this virus temporarily but the mutation power cannot be controlled till now as no reliable vaccine has invented yet. In this research, we have explained the nucleotide mutation rate and pattern in the codon mutation set. A RNN-based LSTM model has been created to predict the future rate of mutation in person's body if affected with COVID-19. Also, we have explained this LSTM-RNN model for time series prediction based on patients' nucleotide mutation rate, and predicted 400th patient's mutation rate in future time. By analyzing more patient data in updated time, this model can be used to predict day basis mutation rates. The situation may change if a reliable way of cure would be invented. Also in this paper, the mutation rate is limited to base substitution only, insertion and deletion rate can be determined in further research.

Funding information

This research received no funding

Declaration of Competing Interest

The authors declare no competing financial interest

CRediT authorship contribution statement

Refat Khan Pathan: Data curation, Formal analysis, Writing - original draft. **Munmun Biswas:** Conceptualization, Visualization. **Mayeen Uddin Khandaker:** Writing - review & editing.

Acknowledgement

Technical supports from the BGC Trust University computer club has been acknowledged

References

[1] 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(4):536-544. doi:10.1038/s41564-020-0695-z.
[2] Laboratory testing of 2019 novel coronavirus (2019-nCoV) in suspected human cases: interim guidance, 17 January 2020 [Internet]. Available from: https://apps.who.int/iris/handle/10665/330676
[3] Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet North Am Ed 2020;395(10223):470–3.
[4] Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Bio-medica 2020;91(1):157–60.
[5] ... Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Cheng Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet North Am Ed 2020;395(10223):497–506.
[6] ... Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Bi Y. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet North Am Ed 2020;395(10224):565–74.
[7] ... Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Sheng J. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 2020.
[8] Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol 2010;84(7):3134–46.
[9] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019;17(3):181–92.
[10] Yang J. Inhibition of SARS-CoV-2 replication by acidizing and RNA lyase-modified carbon nanotubes combined with photodynamic thermal effect. J Explor Res Pharmacol 2020:1–6.
[11] Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. In: Coronaviruses. New York, NY: Humana Press; 2015. p. 1–23.
[12] ... Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Zella D. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020;18(1):1–9.
[13] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Mutations: types and causes. Mol Cell Biol 2000;4.
[14] Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol 2010;84(19):9733–48.
[15] Pfeiffer JK, Kirkegaard K. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. PLoS Pathog 2005;1(2).
[16] Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 2006;439(7074):344–8.
[17] Biswas A, Bhattacharjee U, Chakrabarti AK, Tewari DN, Banu H, Dutta S. Emergence of Novel Coronavirus and COVID-19: whether to stay or die out? Crit Rev Microbiol 2020:1–12.
[18] Das R, Ghate SD. Investigating the likely association between genetic ancestry and COVID-19 manifestation. medRxiv 2020.
[19] Bhowmik D, Pal S, Lahiri A, Talukdar A, Paul S. Emergence of multiple variants of SARS-CoV-2 with signature structural changes. bioRxiv 2020.
[20] ... Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367(6483):1260–3.
[21] Kumar GV, Jeyanthi V, Ramakrishnan S. A short review on antibody therapy for COVID-19. New Microbes New Infect 2020:100682.
[22] Ojosnegros S, Beerenwinkel N. Models of RNA virus evolution and their roles in vaccine design. Immunome Res. 2010;6(S2):S5.