ISSN:0975-9646

Study of COVID-19 and prediction of future model using genetic algorithm

Ravindra Nath[1, a)], **Vivek Pratap**[2]

[1]*University Institute of Engineering, Technology, CSJM University, Kanpur 208024, India*

[2]*Defence Materials and Stores Research and Development Establishment, Kanpur 208013, India*

a) rnkatiyar@gmail.com

Abstract: 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 proposes and explores the mutation rate of the whole genomic sequence gathered from the patient's dataset of different countries. 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. This study also focused on the challenges of uncertainty in data, evolving variants of the virus, and changes in human behavior. 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 a Root Mean Square Error (RMSE) of 0.06 in testing and 0.04 in training, which is an optimized value. Using this training and testing process, the nucleotide mutation rate of the 400th patient in future time has been predicted. About 0.1% increment in mutation rate is found for mutating nucleotides from T to C and G, C to G, and G to T. While, a decrement of 0.1% it's seen for mutating T to A, and A to C. It is found that this model can be used to predict day-based mutation rates if more patient data is available in updated time.

Keywords: SARS-CoV-2, Gene Sequence (GS), Mutation Rate (MR), Neural Network (NN), Based Long Short Term Memory (LSTM).

INTRODUCTION

The novel human coronavirus disease 2019 (COVID-19) was first reported in Wuhan, China, in 2019, and subsequently spread globally to become the fifth documented pandemic since the 1918 flu pandemic. By September 2021, almost two years after COVID-19 was first identified, there had been more than 200 million confirmed cases and over 4.6 million lives lost to the disease. Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment. However, some will become seriously ill and require medical attention. The whole world is suffering from an ongoing pandemic due to Coronavirus disease brought by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was an outbreak in Wuhan, the capital of Hubei province in China in December 2019 [1]. The virus was identified on 7th January and observed that it is spread by human-to-human transmission via droplets or direct contact. Its infection has been estimated to have a mean incubation period of 6.4 days and a basic reproduction number of 2.24-3.58. Since its identification, it has already spread speedily over the whole globe, therefore the World Health Organization (WHO) declared COVID-19 a global pandemic on 11th March 2020. The SARS-CoV-2 is a pathogenic human coronavirus under the Beta coronavirus 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. 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 and placed in the NCBI Gen-Bank [2]. The SARS-CoV-2 is a single positive-stranded RNA virus having a non-segmented nucleic acid sequence.

Although it is an RNA virus 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 the SARS-CoV-2 virus shows about 79% and 50% similarity with the SAR-CoV and MARS-CoV, respectively. SARS-CoV-2 performs mutation during replication of genomic information. 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. Micro-level alteration of mutation rate is also detectable for virus infection in host immune systems and drastically changes the virus characteristic and virulence. To understand viral evolution, the mutation rate is one of the crucial parameters. Furthermore, it is one of the most important factors for the assessment of the risk of emergent infectious diseases, like SARS-CoV-2. Therefore, an accurate estimation of this parameter finds great significance. In connection with this and following the current pandemic, many researchers and scientists are working relentlessly to understand the evolution of SARS-CoV-2. Asim et.al has performed a Phylogenetic analysis of the SARS-CoV-2 virus based on the spike gene of the genomic sequence.

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 finding 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 did not validate the findings with detailed statistical analysis. An analysis of the Gene signature of the SARS-CoV-2 virus has been performed by Ranajit and Sudeep [3]. 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 rates of the European genome and performed statistical analysis on the death/recovery ratio by using Graph-Pad Prism v8.4.0, Graph-Pad 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 [4]. 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 in signature from different variants of the SARS-CoV-2 virus [5]. 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 in Wuhan and India by using multiple sequence alignments. 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 based on the time series. Unfortunately, the current data shows that the SARS-CoV-2 virus is more infectious than the other harmful species of pathogenic human coronaviruses. 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 the SARS-CoV-2 virus, there is a critical need to invent proper vaccine and antibody-based therapy against this virus.

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 help to create an effective resistance against the spreading of this virus. For inventing proper drugs and vaccines against COVID-19 RNA viruses, genomic sequence and mutation analysis are crucially required [6]. Accurate information on the viral mutation rate may play a vital role in the assessment of possible vaccination strategies. In this regard a detailed study on the mutation rate of this virus using the available dataset in the NCBI Gen Bank. From this dataset, Authors have analyzed the Genomic sequence of 3408 patients from different countries for the period of 12th January to 11th May 2020. Specifically, on the mutations that have developed freely on the different dates (homoplasies) as these are likely possibilities for progressing adjustment of SARS-CoV2 to its novel human host. Specifically, the base substitution mutation rates have been calculated. Due to the lack of necessary information for insertion and deletion, consider those as substitution mutations to ensure that no nucleotide goes out of count. It expected that the present analysis will help to understand the changing behavior of this virus in the human body and set up strategies to combat the epidemiological and evolutionary levels. History of This virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing, or breathe. These particles range from larger respiratory droplets to smaller aerosols. This case is infected by breathing in the virus near someone who has COVID-19, or by touching a contaminated surface and then your eyes, nose, or mouth. The virus spreads more easily indoors and in crowded settings.

Variant of COVID-19

Viruses are always changing, and that can cause a new variant, or strain, of a virus to form. A variant usually doesn't affect how the virus works. But sometimes they make it act in different ways. Scientists around the world are tracking changes in the virus that causes COVID-19 [7]. Their research is helping experts understand whether certain COVID-19 variants spread faster than others, how they might affect your health, and how effective different vaccines might be against them. Coronaviruses have all their genetic material in something called RNA (ribonucleic Acid). RNA has some DNA similarities, but they aren't the same. When viruses infect you, they attach to your cells, get inside them, and make copies of their RNA, which helps them spread. If there's a copying mistake, the RNA gets changed. Scientists call those changes mutations. These changes happen randomly and by accident. It's a normal part of what happens to viruses as they multiply and spread. Because the changes are random, they may make little to no difference in a person's health. Other times, they may cause disease.

Major variants of covid-19

As of July 2021, there are four dominant variants of SARS-CoV-2 spreading among global populations.

- ✓ The Alpha Variant (formerly called the UK Variant and officially referred to as B.1.1.7), first found in London and Kent
- ✓ The Beta Variant (formerly called the South Africa Variant and officially referred to as B.1.351)
- ✓ The Gamma Variant (formerly called the Brazil Variant and officially referred to as P.1)
- ✓ The Delta Variant (formerly called the India Variant and officially referred to as B.1.617.2).
- ✓ The currently designated variants are;
- ✓ The Lambda Variant (formerly called the Peru Variant and officially referred to as C.37)
- ✓ The Mu Variant (formerly called the Colombia Variant and officially referred to as B.1.621)
- ✓ Omicron Variant

Variants of concern (VOC)

A SARS-CoV-2 variant that meets the definition of a VOI, through a comparative assessment [7], has been demonstrated to be associated with one or more of the following changes at a degree of global public health significance:

- ✓ Increase in transmissibility or detrimental change in COVID-19 epidemiology;
- ✓ Increase in virulence or change in clinical disease presentation;
- ✓ Decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics.

Dataset Analysis and Pre-processing

An adequate amount of gene dataset is currently available in the NCBI Gen-Bank which contains the complete genome sequence of SARS-CoV-2. Among the many entities filtered the gene sequence, date of collection, and country of the sample. All genes are taken from the human body that affected by COVID-19 [1]. There are genes from almost 33 countries but China, Australia, and the United States have a considerable number of patients' data. Though some countries like England, Italy, France, Spain, and Brazil have a very high mortality rate but to the lack of available data in the NCBI Gen-Bank till 15 May 2022, unable to calculate the mutation rates for these countries separately [8].

Therefore, consider these countries along with others that have low gene data sequences available in the Gen-Bank as the rest of the World category to cover as many regions as possible. In this dataset, there are also some partial genes. So, filtered them and took them only with the level of the complete genome. There is a reference gene sequence of length 29903. Finally, 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 came down to 3068 from 3408, patients from China came down to 40 from 86, The United States came down to 1903 from 2103 and Australia came 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 [8]. 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 transform the four raw nucleotides (A = Adenine, T = Thymine, C = Cytosine, and G = Guanine) into a codon set. A codon consists of three nucleotides and forms a unit of genetic code in DNA or RNA.

Gene Mutation

Gene mutates for many reasons. When RNA tries to copy genetic codes from DNA it may cause some error which causes mutation. Also, errors in DNA replication, recombination, and chemical damage in DNA or RNA cause mutation. There are three types of mutations: base substitutions, deletions, and insertions. From this dataset, 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 a changing codon produces the stop signal for gene translation which causes a non-functional protein then it is called a nonsense mutation. These three types of substitution mutation of the observed dataset where the missense rate is 34.3%, the nonsense mutation rate is 6.7% and the silent mutation rate is 0.8%.

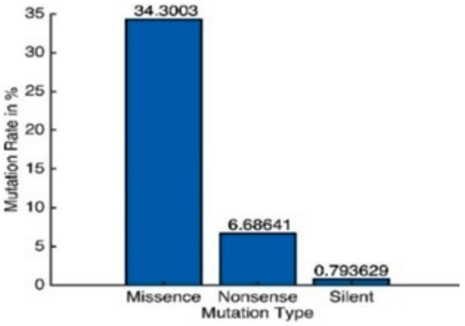

Fig. 1 Mutation rate of missense nonsense and silent mutations

Nucleotide mutation

If the mutation type is missense then it can be said that the change of nucleotide has affected the protein generation, which 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 this process, we have calculated the nucleotide mutation rate for the prepared dataset.

The mutation rate for China has been shown. It shows that a huge percentage of Thymine (T) is being mutated to other nucleotides but not producing the same amount of T again. Also, a huge amount of Adenine (A) is mutated into other nucleotides. Compared to T and A, Cytosine (C) and Guanine (G) did not change much.

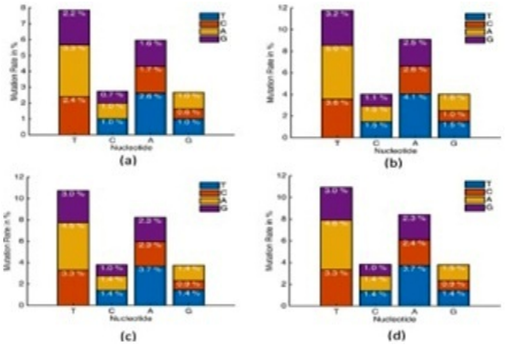

Fig. 2 Nucleotide mutation rate

Methods

The main objective is to study the data of various variants and find patterns between them. To find patterns we will use three algorithms:

• Longest Common Subsequence (LCS) • Binary String • Rabin-Karp

After finding patterns between different datasets and doing analysis predict the future variants and their infectivity.

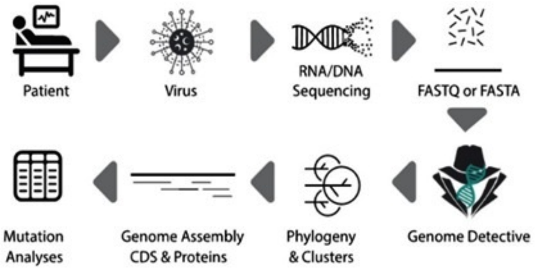

Fig. 3 Future variants and their infectivity

Genetic Algorithm

A genetic algorithm is a search heuristic that is inspired by Charles Darwin’s theory of natural evolution. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation [9-10].

Natural Selection

A genetic algorithm is a search heuristic that is inspired by Charles Darwin's theory of natural evolution. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction to produce offspring of the next generation [9-10].

Natural Selection

The process of natural selection starts with the selection of fittest individuals from a population. They produce offspring which inherit the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance of surviving. This process keeps on iterating and in the end, a generation with the fittest individuals will be found. This notion can be applied to a search problem. In this case, consider a set of solutions for a problem and select the set of best ones out of them. Five phases are considered in a genetic algorithm [11].

1. Initial population
2. Fitness function
3. Selection
4. Crossover
5. Mutation

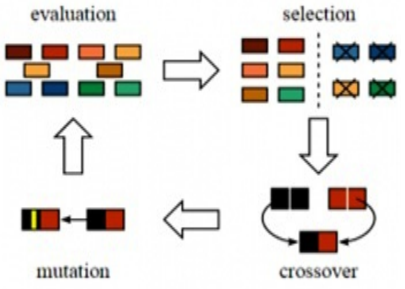

Fig. 4 Phases in a genetic algorithm

Initial Population

The process begins with a set of individuals which is called a Population. Each individual is a solution to the problem want to solve. An individual is characterized by a set of parameters (variables) known as Genes. Genes are joined into a string to form a Chromosome (solution). In a genetic algorithm, the set of genes of an individual is represented using a string, in terms of an alphabet. Usually, binary values are used (a string of 1s and 0s). Encoding the genes in a chromosome and population initialization is the first step in the Genetic Algorithm Process. Population is a subset of solutions in the current generation. Population P can also be defined as a set of chromosomes. The initial population P(0), which is the first generation is usually created randomly. In an iterative process, populations P(t) at generation t (t =1,2,….) are constituted.

Fitness Function

The fitness function determines how fit an individual is (the ability of an individual to compete with other individuals). It gives a fitness score to each individual. The probability that an individual will be selected for reproduction is based on its fitness score. In genetic algorithms, each solution is generally represented as a string of binary numbers, known as a chromosome. These solutions come up with the best set of solutions to solve a given problem. Each solution, therefore, needs to be awarded a score, to indicate how close it came to meeting the overall specification of the desired solution. This score is generated by applying the fitness function to the test, or results obtained from the tested solution.

Use of fitness function for a given problem

Each problem has its fitness function. The fitness function that should be used depends on the given problem. Coming up with a fitness function for the given problem is the hardest part when it comes to formulating a problem using genetic algorithms. There is no hard and fast rule that a particular function should be used in a particular problem. However, certain functions have been adopted by data scientists regarding certain types of problems. Typically, for classification tasks where supervised learning is used, error measures such as Euclidean distance and Manhattan distance have been widely used as the fitness function. For optimization problems, basic functions such as a sum of a set of calculated parameters related to the problem domain can be used as the fitness function [12]. Let's go through a few example problems and their related fitness functions.

Crossover

Crossover is the most important point in a genetic algorithm. For each pair of parents to be mated, a crossover point is chosen at random from within the genes. Crossover is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next. The crossover is sexual reproduction. Two strings are picked from the mating pool at random to crossover to produce superior off spring[13]. The method chosen depends on the Encoding Method.

Different types of crossover Single Point Crossover

A crossover point on the parent organism string is selected. All data beyond that point in the organism string is swapped between the two parent organisms. Strings are characterized by positional bias.

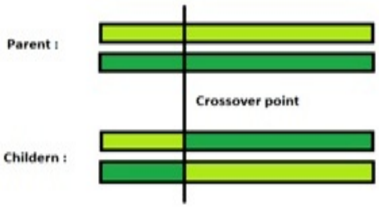

Fig. 5 Single point crossover

Two-Point Crossover

This is a specific case of an N-point Crossover technique. Two random points are chosen on the individual chromosomes (strings) and the genetic material is exchanged at these points.

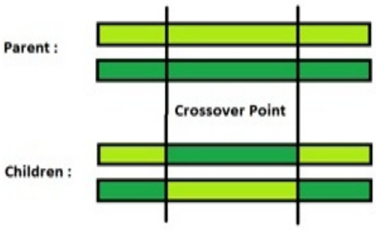

Fig. 6 Two point crossover

Uniform Crossover

Each gene (bit) is selected randomly from one of the corresponding genes of the parent chromosomes. Use the tossing of a coin as an example technique. The crossover between two good solutions may not always yield a better or as good a solution[9]. Since parents are good, the probability of the child being good is high. If the offspring is not good (poor solution), it will be removed in the next iteration during "Selection". The new offspring are added to the population.

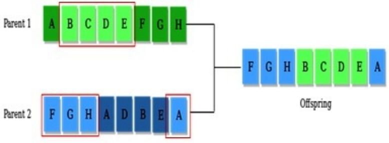

Fig. 7 Uniform crossover

Mutation

In certain new offspring formed, some of their genes can be subjected to a mutation with a low random probability. This implies that some of the bits in the bit string can be flipped. Typically, mutation modifies the properties of a genome just a little. In our case, it picks one gene from the seven to change [13-14]. Then, it randomly selects a single bit in the gene. Finally, it flips that bit. Mutation occurs to maintain variety within the population and prevent premature convergence.

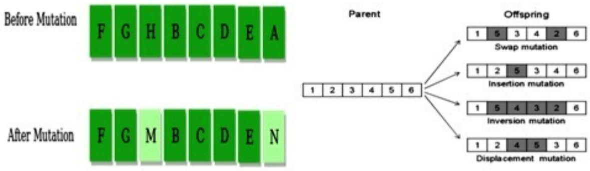

Fig. 8 Mutations

Termination

The algorithm terminates if the population has converged (does not produce offspring that are significantly different from the previous generation). Then it is said that the genetic algorithm has provided a set of solutions to our problem.

RESULTS

The sequence of phases is repeated to produce individuals in each new generation who are better than the previous generation.

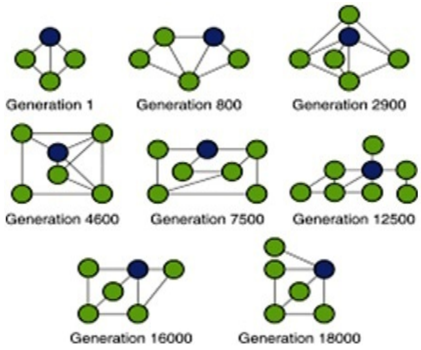

Fig. 10 Produce new generation of previous generation

Here is the code.

```
PSEUDOCODE
START GENERATE THE INTIAL POPULATION
COMPUTE FITNESS
REPEAT
SELECTION
CROSSOVER
MUTATION
COMPARE FITNESS
UNTIL POPULATION HAS CONVERGED
STOP
```

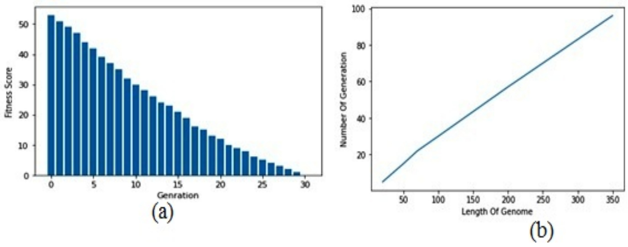

Fig. 11(a) Fitness score of different mutationsand (b) Genome found after this no of mutations

Table. 1 Variation in Fitness score in different generation

Length of string	Generation 0	Generation 10	Generation 20	Generation 30	Generation 40	Generation 50
50	22	5	0	0	0	0
100	54	31	20	0	0	0
200	122	92	65	42	24	9
300	192	156	123	94	69	49
400	259	219	179	146	117	91
500	334	290	248	211	175	147
1000	692	631	574	519	472	430

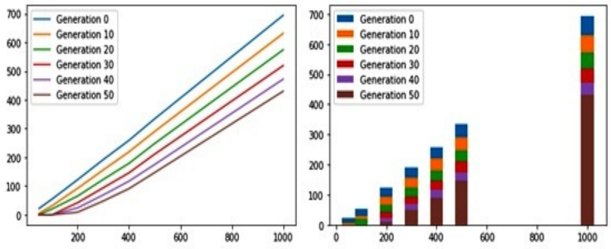

Fig. 12 Generation graph of mutations

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. The lockdown has limited the spreading power of this virus temporarily, but the mutation power cannot be controlled till now as no reliable vaccine has been invented yet. In this research, we have used a genetic algorithm which is a heuristic that is inspired by Charles Darwin's theory of natural evolution. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction to produce offspring of the next generation. Tried to replicate the gene mutation phenomenon as to how many generations it will take to again have major variants of concern and it will help in developing medication before the rise of new variants.

Acknowledgement

This study is part of Doctoral research work, and both authors read and approved the final manuscript.

Conflict of interest

All the authors declare that there are no conflicts of interest regarding the publication of this manuscript's study.

Funding Source

This work did not provide any source of financial support.

References

[1] Gorbalenya A.E., Baker S.C., Baric R.S., de Groot R.J., Drosten C., Gulyaeva A.A., et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat Microbiol. 5 536–544 **(2020)**.

[2] Leitmeyer K., et al. Laboratory testing of 2019 novel coronavirus (2019-nCoV) in suspected human cases: Interim Guidance, W. H. Organization, **(2020)**.

[3] Das R. and Ghate S. D., Investigating the likely association between genetic ancestry and COVID-19 manifestation. Medrxiv **(2020)**.

[4] Valls G. C. and Chalk A. M., Bioinformatics and Computational Biology, Encyclopedia of Data Warehousing and Mining: Second Edition: Section: Bioinformatics, 160-162 **(2009)**.

[5] Bhowmik D., Pal S., Lahiri A., Talukdar A., and Paul S., Emergence of multiple variants of SARS-CoV-2 with signature structural changes, BioRxiv, 1-26 **(2020)**.

[6] Wang C., Horby P. W., Hayden F. G., Gao G. F., A novel coronavirus outbreak of global health concern, Lancet, 395 10223 470-473 **(2020)**.

[7] Nath R., Katiyar N., Saini A. K., and Agarwal A., Study of covid-19 and prediction of future model using machine learning, **(2022)**.

[8] Cucinotta D., Vanelli M., WHO declares COVID-19 a pandemic, Acta BioMed., 91 1 157-160 **(2020)**.

[9] Kelly J. D. (Jr.) and Davis L., A Hybrid Genetic Algorithm for Classification, Proceedings of the 12th IJCAI 91 645–650 **(1991)**.

[10] Janikow C. Z., A knowledge-intensive genetic algorithm for supervised learning, Machine Learning Kluwer Academic Publishers, Boston, 13 189-228 **(1993)**.

[11] Galletly J. E., An overview of genetic algorithms, Kybernetes, 21 6 26-30 **(1992)**.

[12] Shapiro J., Genetic algorithms in machine learning: ACAI, Machine Learning and Its Applications 146-168 **(1999)**.

[13] Katiyar N., Nath R., and Katiyar S., Genetic algorithm approach to find the estimated value of HMM parameters for NS5 Methyltransferase Protein., Biomed. Pharmacol. J. 14 3 1567–1579 **(2021)**.

[14] Krishna K., and Murty M. N., Genetic K-means algorithm, IEEE Trans. Syst. Man, Cybern., Part B, 29 3 433-439 **(1999)**.