Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-05811-4

RESEARCH ARTICLE - SPECIAL ISSUE - AI BASED HEALTH-RELATED COMPUTING FOR COVID-19 (AIHRC)

Correlation-Based Analysis of COVID-19 Virus Genome Versus Other Fatal Virus Genomes

Sidharth Purohit[1] · Suresh Chandra Satapathy[1] · S Sibi Chakkaravarthy[2] · Yu-Dong Zhang[3]

Received: 1 August 2020 / Accepted: 2 June 2021

Abstract
Virus attacks have had devastating effects on mankind. The prominent viruses such as Ebola virus (2012), SARS-CoV or Severe acute respiratory syndrome, Middle East respiratory syndrome-related coronavirus called as the MERS (EMC/2012), Spanish flu (H1N1 virus-1918) and the most recent COVID-19(SARS-CoV-2) are the ones that have created a difficult situation for the survival of the human race. Currently, throughout the world, a global pandemic situation has put economy, livelihood and human existence in a very pathetic situation. Most of the above-mentioned viruses exhibit some similar characteristics and genetic pattern. Analysing such characteristics and genetic pattern can help the researchers to get a deeper insight into the viruses and helps in finding appropriate medicine or cure. To address these issues, this paper proposes an experimental analysis of the above-mentioned viruses data using correlation methods. The virus data considered for the experimental analysis include the distribution of various amino acids, protein sequences, 3D modelling of viruses, pairwise alignment of proteins that comprise the DNA genome of the viruses. Furthermore, this comparative analysis can be used by the researchers and organizations like WHO(World Health Organization), computational biologists, genetic engineers to frame a layout for studying the DNA sequence distribution, percentage of GC (guanine–cytosine) protein which determines the heat stability of viruses. We have used the Biopython to illustrate the gene study of prominent viruses and have derived results and insights in the form of 3D modelling. The experimental results are more promising with an accuracy rate of 96% in overall virus relationship calculation.

Keywords Corona · MERS · SARS · Ebola · COVID-19 · Spanish Flu · DNA · Biopython · Genome

1 Introduction

Genetic engineering [1], also known as the genetic modification, is a strategy to change the DNA in an organism's genetic makeup. It can be done by tuning with one base pair C–G or A–T, removing a complete sequence of DNA, or forming a new copy of the gene. The techniques of snipping out the deoxyribonucleic acid (DNA) from an organism's genetic makeup and joining it to the deoxyribonucleic acid of other specific entity are common practices of genetic engineering. Researchers employ these techniques to enhance or modify the attributes of a living entity. Genetic engineering can be applied to any organism, from a virus (SARS, COVID-19, HIV and Ebola) to a cow. For instance, it can be used to study the impact of disease-causing viruses with a high mortality rate on organism and conditions that counter their survival, thereby helping in the survival of humanity. Traditional hybridization techniques used in plants and

✉ Suresh Chandra Satapathy
suresh.satapathyfcs@kiit.ac.in

Sidharth Purohit
1705078@kiit.ac.in

S Sibi Chakkaravarthy
sb.sibi@gmail.com

Yu-Dong Zhang
yudongzhang@ieee.org

[1] Kalinga Institute of Industrial Technology, Bhubaneswar, India

[2] Artificial Intelligence and Robotics (AIR) Research Center and School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India

[3] School of Informatics, University of Leicester, Leicester, UK

Published online: 24 June 2021

animals are often inclusive, leading to multiplication of unsuitable genes with suitable ones. The genetic engineering procedures involve the formation of re-collective DNA, also applications of cloning the gene and to transfer it, overcome the limitations of traditional methods and allows one to segregate and introduce only the desirable genes into the target organism. Genes describe blueprint to form proteins. They do not form it straightforwardly. The protein synthesis is accomplished via 2 common processes which are: translation and transcription. In the case of transcription, a DNA strand is taken up by molecules and is used as a guidebook to building RNA. RNA molecule acts as a tie-up between the DNA and protein fabrication, whereas while translation, the RNA molecule created in the previous procedure conveys data from the DNA to the protein building units in a cell.

DNA - - Transcript to → RNA - - Translated to → Protein

Nucleotide molecules are common fundamental formation units for both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). It is to be noted that the proteins are not built of the nucleotides but by amino acid sequences. The processes translation and transcription convert the DNA nucleotides to amino acid sequences, for fabrication of the intended protein. Such phenomena occur in each of the prokaryotic and eukaryotic cells. A single DNA strand comprises a chain of connecting nucleotides, containing one sugar, one of the nitrogen bonded bases(purines or pyrimidines) and one phosphate group. Four nitrogenous bases present in the deoxyribonucleic acid are: guanine or G base, adenine or A base, thymine or T base, cytosine or the C base. DNA occurs in a double helical form, where two strands are interlocked and bonded by the nitrogenous bases present in them. The base molecules formed this way are known as base pairs. A–T and G–C do not bond with each other and only form conjugate bonds, i.e. A with T and G with C. DNA and RNA are different from each other as the former has a deoxyribose sugar and the later has a ribose sugar. Although out of the total four nitrogenous bases, three are similar in both deoxyribonucleic acid and ribonucleic acid. ribonucleic acid molecules get a U or uracil in place of the T base. When transcription occurs, U replaces the position of T resulting in the formation of compatible base pairs between A (adenine) and U (uracil).

Transcription [2] is the mechanism of copying genetic material data from one strand of DNA into the RNA strand. The principle of complementary governs the process, except to the fact that the A or adenine forms the base pair bond with U or uracil instead of T or thymine. Like the utility in replication, the DNA is also used as a transcription template. The data present in DNA are written again or are transcribed into the ribonucleic acid molecule. Every nitrogenous base of a DNA provides data for protein fabrication. Every strand of DNA has a particular bases sequence. These determined sequences provide the data for the synthesis of intended protein. The bases of deoxyribonucleic acid are transcribed to corresponding bases of a strand of a ribonucleic acid via transcription. Also, the data of a deoxyribonucleic acid molecule are passed to a novel single strand of ribonucleic acid molecule which carries these data to those cells where proteins are synthesized. The ribonucleic acid molecules employed for this purpose are also called mRNA or messenger RNA.

Translation [3] is the mechanism where the data carried by messenger ribonucleic acid molecules are used to fabricate proteins. The sequence and order of amino acids are determined by the sequence of the messenger ribonucleic acid (mRNA). The particular nucleotide sequences in the messenger ribonucleic acid molecules yield the details for the protein synthesis. Ribonucleic acid is composed of several nucleotides, whereas protein is constructed from various amino acid units. When translation occurs, the data in a strand of a ribonucleic acid are translated from ribonucleic acid gene protocol to polypeptide-based gene protocol, so the nucleotides sequences get translated to amino acid sequences.

DNA sequencing [4] is the genetic engineering technique which deals with finding out the sequences of nucleic acid and the nucleotide arrangement in a given deoxyribonucleic acid. It comprises all the methods or technologies that are used in finding out the order of occurrence of the 4 bases A, G, C, T discussed above, in a deoxyribonucleic acid. The advancement of evolving DNA sequencing methods has escalated biomedical research and discoveries. The mutated and the healthy DNA sequences are compared and diagnosed to find cures for various ailments like cancers, polio, AIDS, etc. The faster way of DNA sequencing allows quicker and more specific medical examination. These quicker techniques speed up DNA sequencing by using high-power distributed technologies. Hence, in this paper, an analysis of the viral DNA sequences, their heat stability and computational pattern recognition has been studied, analysed and visually shown. The key contribution of this paper is given below. A design of novel vectorizer based on term frequency (TF) and inverse term frequency (ITF) which is much faster than the existing methods. The manual assessment of weights on humongous FASTA sequence of DNA genome data and their k-mers takes time; it can have disparities and is inconsistent. To deal with this, an algorithm to assign the weights on the basis of likability and occurrence of a given k-mer gene sequence in the whole DNA is employed in this research; it is called the TF-IDF score.

TF-IDF score helps in determining how rarer is a given text or a word in a given document. More essentially, the score helps in assuring and finding the optimal aptness of a given *codone* or a protein sequence in the whole DNA

genome structure. The authors of this manuscript have used the TF-IDF weights on the DNA genome sequences of fatal viruses because it has helped us in finding the most common repeating k-mer in the DNA sequence. The lesser is the weight, there is much more likability for that k-mer to be present in the given sequence and vice versa. K-mer determination is essential in this pattern recognition task, and it helps in finding the DNA pattern that are unique to a given virus and has helped us in successfully identifying the k-mer sequences with higher search proportions. It has further helped us to optimize the selection of appropriate combination for similarity-based analysis, which is the base of this research. The contribution of this paper is listed below:

- Applying k-mers algorithm to transform the different length DNA sequence into the equal-length vectors.
- Two different levels of analysis: (i) the genome correlation-based analysis and (ii) the genome alignment-based analysis.
- An exclusive comparison study and three-dimensional (3D) visualization of the viruses DNA sequence.
- A comparative analysis on DNA sequence distribution, percentage of GC (guanine–cytosine) protein which determines the heat stability of viruses, etc.

The following manuscript is assembled as follows: Sect. 2 reviews the state-of-the-art techniques available in gene data analysis, also the literature survey of this manuscript. Section 3 proposes the methodology for this paper. Section 4 showcases the experimental setup and dataset collection. Section 5 shows the experimental results obtained for the virus genomes. Finally, Sect. 6 concludes the paper.

2 Literature Survey

Viruses are the only type of microorganism that cannot replicate without a host cell. It is to be noted that, after they contact a host cell, a virus injects its genetic composition into the targeted host and takes over its vital functions including the cell division and multiplication. Computational knowledge of the deoxyribonucleic acid sequences and pattern distributions of a virus has become essential for bio-medical fields. They are used in developing intelligent systems for medical diagnosis, prognosis and effective treatment. This knowledge representation helps the researcher to find the relationship between various viruses aiming at finding the treatment or cure for viral borne diseases. Furthermore, computational algorithms on DNA sequencing has also showed up a good performance in terms of outcomes. In the current section, we inspect the existing state-of-the-art (SOTA) computational techniques used by the researchers in past for DNA sequencing and analysis.

Campbell (2020) proposed a novel next-generation DNA sequencing technique. The process used in their paper is to find patterns within the DNA sequences of viruses that have affected the mankind adversely for decades. They claimed that the DNA sequence is a single format through which a wide domain of biological phenomena can be presented for high output data gathering [5]. The next level of analysis carried out using statistical interference plays a vital role in analysing the DNA sequences. Bullard Purdom et al. [6] used Illumina Genome Analyser (IGA), which are robust tools for analysing a broad range of biological and medical queries. IGA helps to understand the high output sequencing technologies which uses probabilistic and mathematical algorithmic methods to provide accurate and useful insights from the complex and informative datasets. Irizarry et al. [7, 8] proposed a novel method for the analysis of gene expression data. Their method explained about analysing the pairwise alignments and helped in deriving meaningful pattern from the genome data. Craig Venter initiated the human genome project [9] towards the contribution of DNA sequencing techniques.

Cosine similarities are a fundamentally strong tools used to find the similarity in corpus-based data and has showed good performance insights even with the large dataset. Srikanth Maturu [10] compared the clinical trial data of 10 different virus sequences comprising more than 25,000 proteins in each sequence. Their experimentation analysis claimed that the cosine-similarity-based algorithm is as good as the BLASTP [11] algorithm and the CD-HIT algorithm. Further, it also noticed that the CD-HIT algorithm uses a short word filter [12], which is not accurate, to compute similarity between any two sequences. The KD tree- and ball tree-based approaches [13] used for genome sequences as a part of approximation nearest neighbour-based techniques had some flaws, for a KD tree all the instances of a sequence that are less than the median are placed in the left partition, and all that are greater than or equal to the median are placed in the right partition. Now for each partition, a KD tree is further constructed using the remaining $d-1$ dimensions recursively. Every leaf node of a given KD tree stores a set of instances. For a given query q, the KD tree is traversed from the root down to a leaf node by comparing values of the query with the median at each split corresponding to a dimension. All the instances in the leaf node are returned as the nearest neighbour to the query. This method is an approximate nearest-neighbour search method because it can miss some instance when the instance is placed into another partition.

Similarly, for the ball tree, a d-dimensional set of instances D are considered. Instead of splitting the points based on some median, they are split by computing distances between two centroids. Initially, two centroids are chosen; for every instance, distance to the two centroids is computed

and is assigned to the cluster with the smaller distance forming two clusters. If the distances are equal, then the instance is assigned to the cluster chosen randomly. In turn for each cluster, two more centroids are chosen and split again. This is continued recursively until each cluster is containing a specified number of points or the number of clusters allowed has reached. For a given query q, recursively computed to an end of a cluster to which it belongs and all the instances in that cluster are returned as the nearest neighbour to that query. Again, some instance can miss when the instance is placed into another cluster. Newberg et al. [14] utilized the hidden Markov model (HMM) and hidden Boltzmann model (HBM) for analysing error statistics in the genome data. In their method, the frequency of every possible k-mer occurrence in each sequence is noted. They used HMM effectively to perform subtraction of random background gene information for highlighting the role of exacting evolution and thereby reduced the influence of random neutral mutations. Furthermore, the cosine similarity function is being employed to compute the pairwise distance between composition vectors of the sequences. The distance matrix acquired by this process can be used to construct a phylogenetic tree using algorithms like UPGMA and neighbour joining [15].

From the literature, it is inferred that although the established alignment-based approaches are providing good insights when the sequences under observation are firmly related and stable aligned. But if the sequences are divergent, a stable alignment is not generated; thus, the applications of sequence alignment are restricted. Some other constraints of the alignment-based approaches are their computationally expensive and time drawing; thus, they are limited when dealing with multiply scaled DNA sequence data. With the advancement of next-generation sequencing technologies, huge amount of sequence data is getting generated, especially now during a global pandemic situation. The dimensions of this sequence data give rise to provocations on the established alignment-based techniques in the comparative analysis. To address this issue, this article proposes the term frequency (TF)- and inverse term frequency (ITF) [16]-based cosine similarity method which uses the alignment-free algorithm called k-mers [17] algorithm which is more accurate than the short word filter [18].

3 Methodology Employed

Figure 1 shows the design and flowchart of the proposed methodology. We first load the DNA sequence given in FASTA format [19], and then we go for two main phases: (i) genome correlation-based analysis or the (ii) genome alignment-based analysis. The objective of the overall analysis is to find a mathematical and graphical correlation between all the considered viruses and find similarity between their genome sequences that can facilitate in treatment (cure), drug assessment and vaccine preparation.

3.1 Genome Correlation-Based Analysis

For correlation-based analysis, we find k-mers of the DNA sequence and the k-mers results are converted into a string. The reason behind string conversion is the mutable nature of the object (string), and it helps to remove the redundant information. K-mers are built upon to construct DNA sequence, improvise the gene data, identify species given microscopic sample and create enervated vaccines. The concept of k-mers is that the sub-sequence of length 'k' contained within a biological sequence is formed from the A, C, T and G nucleotides such that the sequence CGTA would have 4 monomers (C, G, T and A), 3 two-mers (CG, GT and TA), 2 three-mers (CGT and GTA) and 1 four-mer (CGTA). It thus can be stated that for a sequence of length if 'Z' is the nucleotide sequence length, then it shall be having Z-$k+1$ k-mers and pk total possible k-mers, where 'p' is the number of possible monomers.

A sub-sequence k-mer of length 6 is considered and used for mathematical analysis. It is also noted that a hexanucleotide bias of two or more phylogenetically similar species has a lot of identical features. This is due to the exact cause of variation in hexanucleotide, but it has been conceptualized and established as an outcome of the preservation of genetic stoutness at the molecular extent.

3.1.1 Selection of k-mers Value

Disadvantages of using a small value of k in the k-mers algorithm can be leading to the overlapping of the k-mers to overlap and leading to complications for constructing the De Bruijn graph [20]. There are also risks where a large number of vertices merge in the graph leading into the k-mers which increases the complexity. Thus, this orientation of the genome gets more arduous due to increased level of path ambiguity. The data often get lost when the k-mers become small and insignificant. For instance, the possibility of ACT GACCTCTG is lesser than the sequence ATGG and thus naturally has a large volume of information. If k-mers are small, then the areas where small repeats occur often get harder to be resolved. In the subsequence ACTGACCTCTG, the number of re-occurrences of TC will vanish if a k-mer size is less than 6 is used.

Hence, by extending the dimensions of the k-mers, the vertex shall decrease leading to boosting the process of fabrication of the genome, as there shall be a fewer path to traverse in the graph. The extremely large k-mers size also has more risk of non-overlaps with another k-mer by k-1 factor which leads to disjoints in the reads, thus having a higher amount of smaller contig parts. The sufficiently

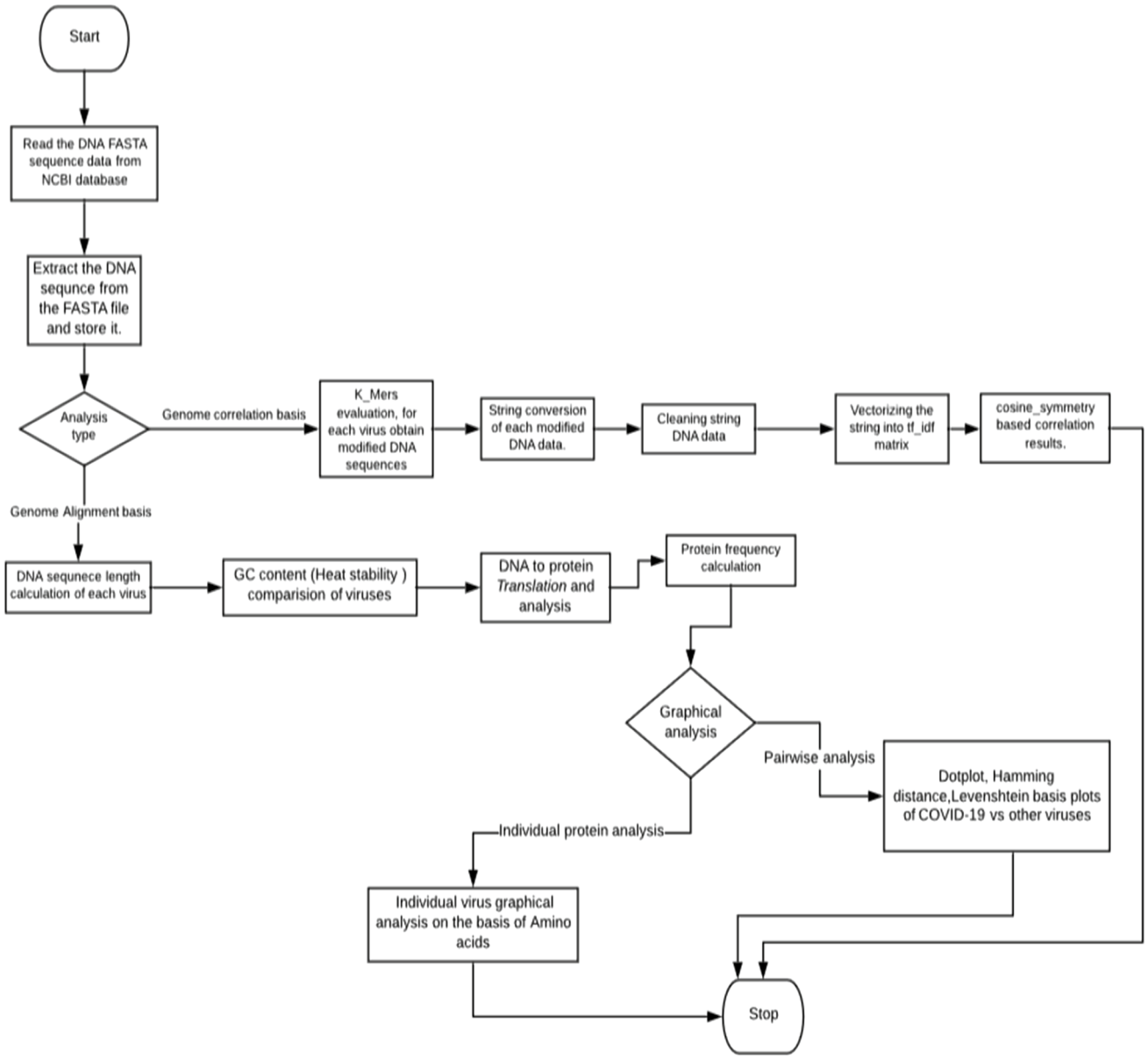

Fig.1 Flowchart of the proposed methodology

large k-mer sizes helps in dealing with the problem of small repeated patterns.

Then, we apply TF-IDF [16]-based Vectorizer methods which help in finding the frequency distribution of the topmost occurring elements (strings) in each hexamerized (k-mers = 6) gene sequence. In information retrieval and gene data analysis, TF-IDF is a probabilistic, mathematical modelling method that helps to reflect how essential a word is to a gene sequence in a given dataset. It is employed in the weighting factor calculation in the information retrieval-based tasks, text mining and user modelling. The TF-IDF value increases equivalently to the number of times a pattern comes in the gene sequence and helps to adjust for the fact that some k-mers patterns appear more frequently in general.

The tabular analysis shown in Table 1 is the technique employed to clean a DNA sequence data and converting it into a string of k-mer form, where we have considered a hexamer. It is used for a descriptive analysis of DNA sequence and formulating a mechanism for the synthesis of the genes. This tabulation has helped in the analysis in terms of tf_idf-based matrix format. Similarly, an analysis for all the 10 viruses genome is being made and they are then converted to a cosine-based similarity matrix, which then provided input to a tf_idf-based vectorizer. The role of a vectorizer is to convert the repeating data to probabilistic representation.

Table 1 Modified sequence of the virus-COVID-19 in the k-mers form

Virus DNA genome considered	Modified sequence with optimum k-mer, $K=6$	String conversion of the sequence	Cleaned hexamers to be used in frequency matrix
COVID-19	[Seq('attaaa',SingleLetterAlphabet()), Seq('ttaaag', SingleLetterAlphabet()), Seq('taaagg', SingleLetterAlphabet()), Seq('aaaggt', SingleLetterAlphabet()), Seq('aaggtt', SingleLetterAlphabet()), Seq('aggttt', SingleLetterAlphabet()) .....]	"[Seq('attaaa',SingleLetterAlphabet()), Seq('ttaaag', SingleLetterAlphabet()), Seq('taaagg', SingleLetterAlphabet()), Seq('aaaggt', SingleLetterAlphabet()), Seq('aaggtt', SingleLetterAlphabet()),....]"	attaaa ttaaag taaagg aaaggt aaggtt aggttt ggttta gtttat tttata ttatac tatacc atacct taccttt accttc ccttcc cttccc ttccca tcccag cccagg ccaggt caggta

The task is performed by considering the frequency of occurrence of the sub-sequence (hexamers in our case) and allocating weights to them, thereby forming a matrix.

3.1.2 Term Frequency

Let us consider a simple example "AGTTCT", a set of DNA gene sequence and we want to organize based on gene sequence pattern which has the most relevance to a query. We use the k-mer concept, where $K=6$. We can analyse simply by eliminating nitrogen bases that do not contain all characters in the order-A G T T C T, but this is time consuming, redundant and has disparities. To further distinguish them, we might count the number of times each such term occurs in each gene sequence; the number of times a pattern occurs in a DNA sequence is called its term frequency. However, in this case where the length of the DNA sequence varies greatly and the adjustments are often made, the weight of a term that occurs is consistent with the term frequency.

3.1.3 Inverse Document Frequency

When a base pair say "GC" is common, the term frequency will wrongly be biased to this base pair without giving enough weight to other essential base pair letters in the DNA sequence. These base pairs are biased and thus not a good demarcation for classification between useful and non-useful gene patterns. Thus, the *inverse document frequency* factor is used which helps in reducing the weight of patterns that occur very frequently in the gene sequence and increases the weight of those occurring seldom.

Mathematical equation used to evaluate TF-IDF for a term m of an amino acid l in a gene sequence can be given as:

$$\mathrm{TF-IDF}(m,l) = \mathrm{idf}(m) * \mathrm{tf}(m,l) \tag{1}$$

idf term is given by:

$$\mathrm{idf(m)} = 1 + \log\left(\frac{s}{\mathrm{df}(m)}\right) \tag{2}$$

Here, we have s as the number of amino acids present in a gene sequence, df (m), is the amino acid frequency of m. Here, df is the count of total amino acids in the gene sequence which contain the term m.

Finally, this term frequency-based vectorized matrix is used to find cosine similarity between the virus genomes. Cosine similarity is employed as a metric to determine the interdependence, similarity and correlation between any two sets of viruses. The given metric is dynamic and can easily be customized in accordance with change in DNA sample data, example mutation strands occurring on COVID-19 virus. Since COVID-19 is a mutating in nature, we have

deployed this metric to determine the correlation between corona virus and the other fatal viruses. The metric is very sensitive to shift in invariant of an input and has value in the range [−1,1] or [0,1] depending upon whether the input is negative or non-negative vector. It is an inner-product normalization representation of two vectors that helps in determining the similarity angle between any 2 DNA sequences, irrespective of the change in its sequence type, or structure.

Cosine similarity is a mathematical concept of finding the similarity between two nonzero vectors. It elucidates the cosine of the angle between the considered vectors. This similarity helps in knowing how similar two given gene sequences are, and what is the level of commonness between them. The cosine of 0° is 1, and it is less than 1 for any angle in the interval (0, π] radian. The cosine similarity if 1 determines the overlapping tendency of two gene sequences. Also, by trigonometry, we know that the two vectors inclined at π/2 rad for each other have a similarity of 0 and the diametrically obverse will be having a similarity of −1, independent of the magnitude. For two viruses converted into vectors I and J by tf_idf equations, the cosine similarity can be calculated as:

$$\text{cosine\_similarity} = \frac{I \cdot J}{\|I\|\|J\|} \tag{3}$$

In the proposed genome sequence analysis, each term is assigned a different dimension. Here, every gene sequence is determined by a mathematical vector, and the value of each dimension correlates with the number of times the term appears in the given gene sequence. These bounds can be applied for the specified number of dimensions, and it is to note that the cosine similarity is most commonly used in high-dimensional positive spaces [21]. For instance, in the gene sequence data with over 25,000 character sequences, counting the most abundant occurring pattern in the string is a complex task. To ease with this process, techniques like cosine similarity in treating the gene sequence as a vector and then facilitates the entire processing.

Furthermore, the angle between the COVID-19 [22, 23] DNA represented vector and all other virus DNA vectors was also computed. As a result, the more the angle between the vectors, the less the similarity between the viruses.

3.2 Genome Alignment-Based Analysis

For genome alignment-based analysis, we find the DNA sequence length, GC content of the DNA (to find heat stability of the viruses), also DNA translation to proteins and amino acids. Amino acids count is taken into consideration to count the frequency of each amino acid in each protein sequence. Further, this counting helps in the graphical visualization and analysis of the pairwise alignment-based comparison in terms of the global alignment. We used dot plot-based techniques to visualize the similarity quotient between the COVID-19 and other virus genomes. The usage of PDP parser is being used here to find the 3D model representation of the viruses, and this 3d model helps in understanding the structure of the virus, finding patterns in the DNA molecule, the representation of DNA and helical strands of the viral genome. Hamming distance (HD) between two viruses say P and Q can be given as:

$$\begin{aligned} &\text{Hamming\_Distance (HD)} = \sum_{s=1}^{z} |P_s - Q_s| \\ &P = Q \Rightarrow \text{HD} = 0 \\ &P \neq Q \Rightarrow \text{HD} = 1 \end{aligned} \tag{4}$$

where z is the length of the larger out of the two compared virus lengths. The dot plot-based graphical implementation is based on the Hamming distance between the two compared viruses.

4 Experimental Setup

The proposed analysis model is experimentally tested in a testbed using a dedicated server with Intel Xeon (6th Gen) processor, 28 GB RAM, NVIDIA Titan X Pascal GPU unit, 2 TB Hard drive running on Ubuntu server. The proposed analytical model is implemented using Biopython. Biopython implementation contains parsers for various bioinformatics file formats like the BLAST, FASTA and Genbank. It provides support from bioinformatics vendors like NCBI or Expasy interfaces. Essential tools like the conventional sequence-based classes and clustering methods and also some of the data structures like KD tree are provided by this framework. The dataset for the experimentation has been taken from the NCBI (National Center for Biotechnology Information) [24] in the FASTA file format. The total size of the dataset is around 1.1 GB. The algorithm for the entire process flow is given below:

Algorithm

Algorithm

Input: FASTA file (DNA Sequence)
Output: Visualize (DNA sequence, protein sequence, similarity values)
Begin
Read the DNA sequence
Find the GC content from the given DNA sequence
Translate each DNA sequence into protein
Count for Amino Acids in the protein sequence
Distribute the protein sequence of each virus
Find the pair-wise alignment
Set Kmers size = 6
Apply term and inverse term frequency
Calculate Cosine similarity
End

5 Result Analysis

The algorithms discussed in Sect. 4 are being employed to compute the virus protein structures. The main objective is to do a computational comparison for showcasing the relationship between the COVID19 and the other prominent viruses. The algorithmic representation gives a clear illustration of the workflow and the analysis. The data, graphs and 3D models presented below in this are completed after the analysis using the novel algorithms introduced in the previous section.

The genetic code establishes the relationship between the A, C, G, T DNA base sequences in a gene and the protein sequence it holds. Figure 2 showcases the standard codon table. The codon is a trinucleotide sequence of DNA that corresponds to an amino acid. The three-stop codon marks the end of the sequence of the protein. One start codon, called as the "AUG", signifies the beginning of a protein and corresponds to the amino acid methionine.

DNA sequence lengths of considered viruses are shown in Table 2. The length of DNA segment is calculated by finding the number of base pairs and multiplying it by the distance between adjoining base pairs. Table 3 shows the GC content of all the above-mentioned viruses. From Table 3, it is clearly shown that the GC base pairs have 3 hydrogen bonds, while AT base pairs have two. Therefore, double-stranded DNA with a higher number of GC base pairs will be more strongly bonded together, more stable, and will have a higher melting temperature. It is clearly seen that all the viruses are thermally more stable than COVID-19 virus, proving that the annealing temperature will be lower for this type of virus. It is also found that the only exception is the HIV virus that has lower GC content than COVID-19; thus, it is also having less heat stability.

	T	C	A	G	
T	TTT F	TCT S	TAT Y	TGT C	T
T	TTC F	TCC S	TAC Y	TGC C	C
T	TTA L	TCA S	TAA Stop	TGA Stop	A
T	TTG L(s)	TCG S	TAG Stop	TGG W	G
C	CTT L	CCT P	CAT H	CGT R	T
C	CTC L	CCC P	CAC H	CGC R	C
C	CTA L	CCA P	CAA Q	CGA R	A
C	CTG L(s)	CCG P	CAG Q	CGG R	G
A	ATT I	ACT T	AAT N	AGT S	T
A	ATC I	ACC T	AAC N	AGC S	C
A	ATA I	ACA T	AAA K	AGA R	A
A	ATG M(s)	ACG T	AAG K	AGG R	G
G	GTT V	GCT A	GAT D	GGT G	T
G	GTC V	GCC A	GAC D	GGC G	C
G	GTA V	GCA A	GAA E	GGA G	A
G	GTG V	GCG A	GAG E	GGG G	G

Fig.2 DNA standard codon table

Table 2 Finding the length of each considered DNA sequence

Virus name	Length of protein sequence
COVID-19 virus	29,903
SARS virus	29,751
MERS virus	30,119
HIV virus	999
Ebola virus	18,959
Dengue virus	15,256
Rotavirus	213
Hanta virus	3653
Spanish flu virus	930
Swine flu virus	982

Table 4 exhibits the pairwise sequence alignment which is used in determining the regions of commonness that may be functional evolutionary or structural occurring within the two biological sequences that may be a nucleic acid or a protein. For the global alignment, end to end alignment of the entire sequence is done to find the region with the most similarity among the two sequences, it contains alphabets from both the inquiry and quarry sequences. From Table 4, it can be easily seen that the SARS has the highest pairwise global alignment weightage with the COVID19. The graphical analysis shown in Fig. 3 will describe the content of amino acids in each of the considered proteins. The

Table 3 GC content of each of the virus

Virus name	GC content percentage
COVID-19 virus	37.97277865097148
SARS virus	40.7616550704178
MERS virus	41.23642883229855
HIV virus	33.233233233233236
Ebola virus	41.07284139458832
Dengue virus	49.0954378605139
Rotavirus	42.857142857142854
Hanta virus	39.41965507801807
Spanish flu virus	47.41935483870968
Swine flu virus	47.04684317718941

Table 4 Pairwise global alignment between COVID-19 DNA and other viruses DNA sequence

Virus name	Pairwise global alignment percentage (%)
SARS virus	89.0
MERS virus	72
HIV virus	57
Ebola virus	58
Dengue virus	62
Rotavirus	21
Hanta virus	62
Spanish flu virus	64
Swine flu virus	62

following analysis now represents the comparative analysis of DNA sequences of the COVID-19 with all other considered viruses. The dot plot is used here for visualization in-order to assist the virologist, researchers and medicinal experts to find similarities between the Wuhan sample and other viruses that have affected mankind.

The experiment depicts that after considering a TF-IDF-based matrix, the outcome of the cosine-based matrix is 10*4107 dimensions, where each of the 10 viruses is represented as a column vector of 4107 features. TF-IDF-based matrix of the virus genomes is being passed into the cosine similarity function, and thus, each virus gets treated as a vector-based representation on a plane with their frequency values as the coordinates of weighted subsequences on that plane. This is how using the vector space the angle between the viruses is being evaluated, thereby proving that the mathematical analysis methods are meant to be more effective in the gene analysis. Figure 4 represents a dot plot-based analysis between all the viruses and the COVID-19. A sharp demarcating straight line for the matching amino acid–base pairs can be observed for SARS versus COVID-19 graph, indicating the genetic similarity between these viruses. Also, Fig. 5 represents the PDP parser output of DNA gene sample data collected from the NCBI web database. We have used the algorithm mentioned in Sect. 4 to depict the 3-D structure of these viruses.

Table 6 represents the analysis of the cosine similarity. From the experimentation, it is found that the COVID-19 is similar to the SARS virus that appeared in 2002 and is least similar to the rotavirus. The cosine value 0.92229834 states that when we consider COVID genome-based vector as the starting ray and the SARS genome vector as the ending ray, the angle between them is given as theta = Cosine-1(value). As cosine is a periodic function, so lesser the value of theta, the closer will be the inclination between two vectors. Thus, it can be said that the more the angle, the lesser the similarity between the gene structures of the two considered viruses because genome mutation has not occurred and constant DNA sequence data have been taken into account. Therefore, one with least angle that is SARS is most similar and Rotavirus being the one with least similarity.

A comparative analysis has been carried out recently for the newly observed mutations of COVID-19 in various parts of the world, namely in Australia, Bahrain, France, Germany, Greece, India, Italy, Japan, Netherlands, Saudi Arabia, South Korea, Spain, Thailand, Tunisia, USA (California, Minnesota, New Mexico, Wisconsin), UK (the latest variant which has affected many people), Vietnam and Zambia. So a new sample of 20 more variants, recently obtained from various parts of the world, are used directly from the NCBI database [24]. The algorithm and methodologies suggested earlier in this paper were carried out on these new variants, and some tremendous results were being found out. Firstly, the natural language processing-based text analysis on DNA sample of COVID-19 mutants works remarkably well similar to our previous analysis. Secondly, the technique suggested for finding similarity between the COVID-19 variants can be used for any DNA virus genome in the future. Hence, it can be stated that the same techniques, regardless of mutation or appearance of new strands in the genome of COVID-19 or any virus, can be used in all the cases. It also implies that our findings as given in Table 7 of the following samples the USA (New Mexico variant) and the USA (Wisconsin variant) are least similar to the Wuhan variant obtained from China, as the angle between there DNA vectors is significantly larger.

With a motivation to compare the trends of the COVID-19 Wuhan sample, with other Asian and global variants to provide a holistic outcome, accounting all mutations and trends of the viruses is the next target of our research. The analysis aims to contribute to mankind a step towards the vaccination and drug preparation for the COVID-19 pandemic, which has taken more than 500 K lives as of June 2020. An approach to using unsupervised machine learning

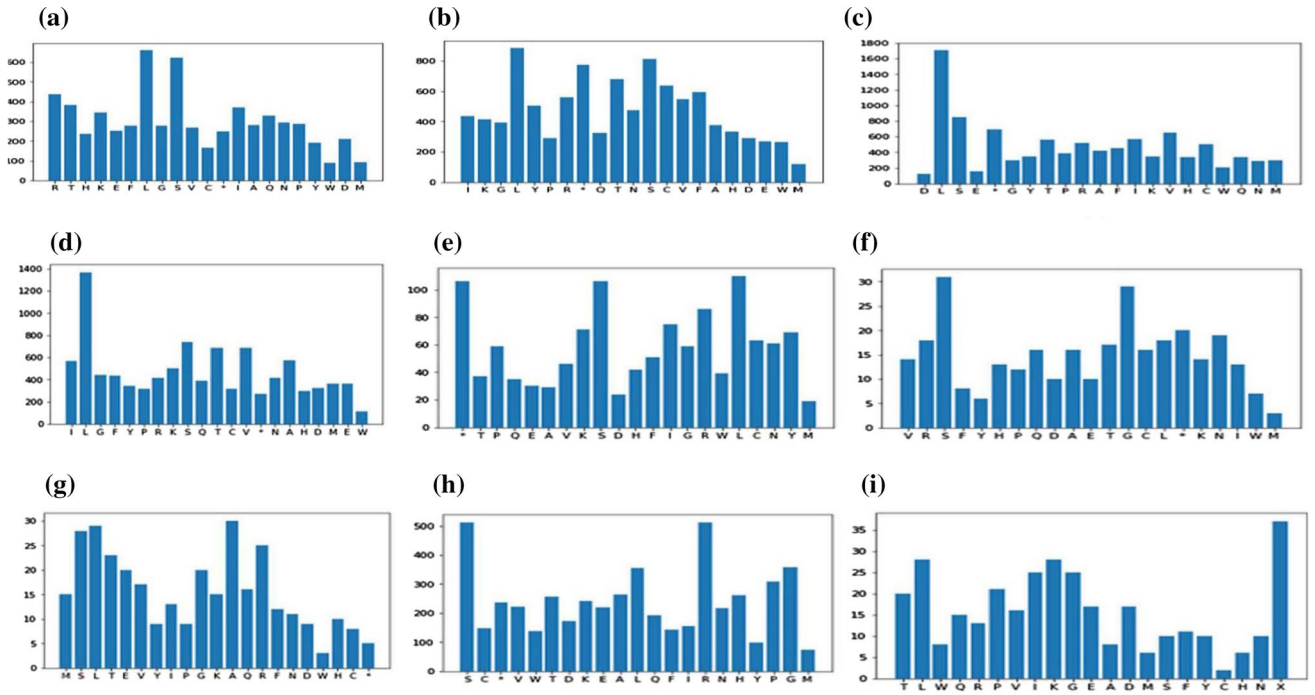

Fig.3 Protein sequence distribution of various virus such as (i) COVID-19, (ii) Ebola virus, (iii) MERS, (iv) SARS virus, (v) Hanta virus, (vi) Spanish flu virus, (vii) Dengue virus, (viii) Swine flu virus, (ix) Rota

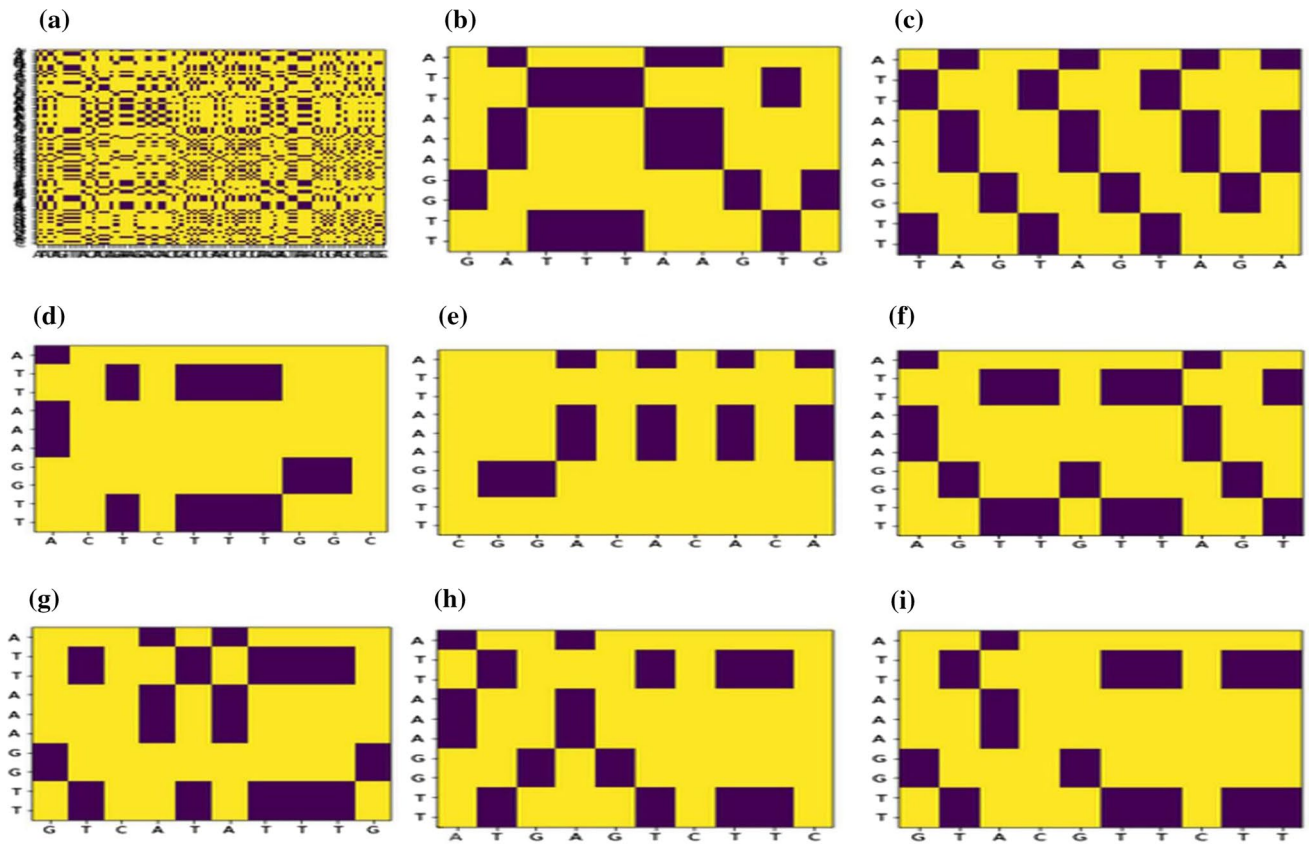

Fig.4 Dot plot-based comparison of the COVID-19 versus the various virus such as (i) SARS (ii) Ebola virus, (iii) MERS, (iv) Hanta (v) Spanish flu (vi) Dengue virus, (vii) Swine flu virus, (viii) Rota

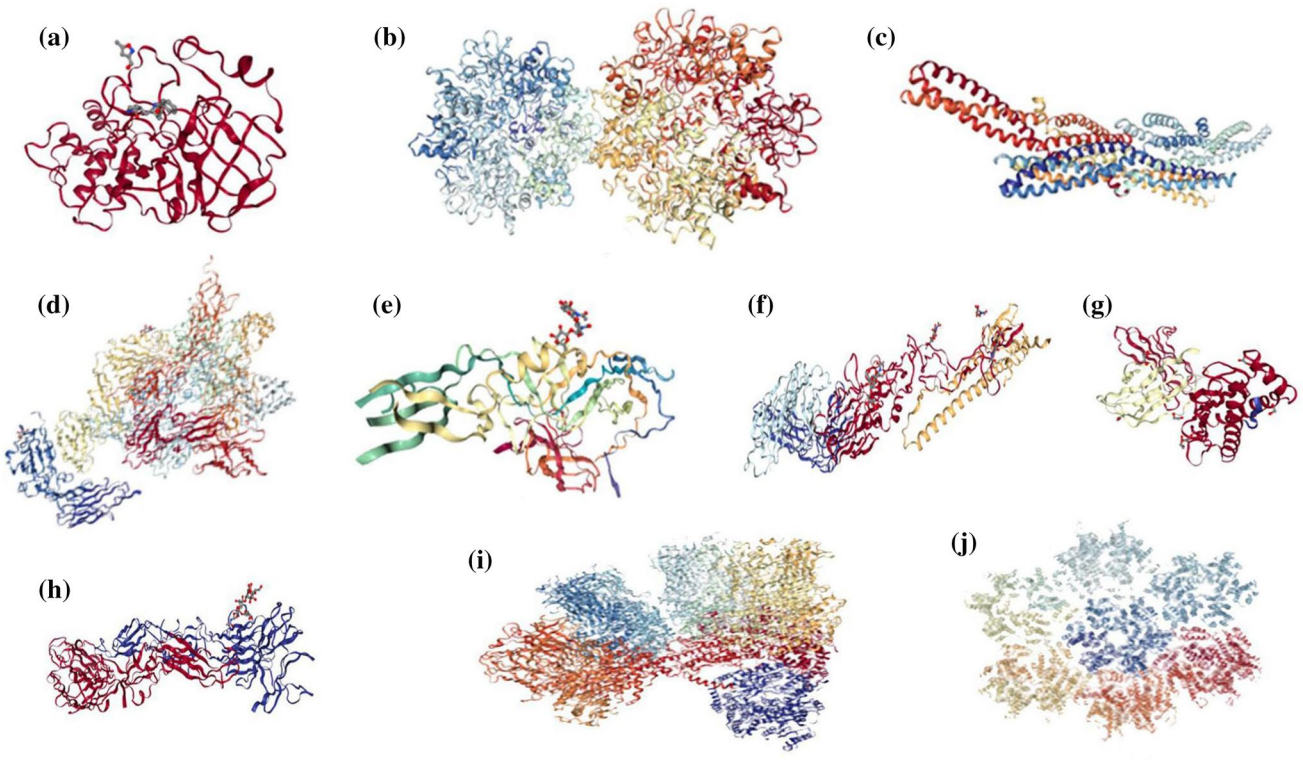

Fig.5 Visualization (3D model) of various virus such as (i) COVID-19, (ii) Ebola virus, (iii) MERS, (iv) SARS virus, (v) Hanta virus, (vi) Spanish flu virus, (vii) Dengue virus, (viii) Swine flu virus, (ix) Rota

Table 6 Cosine similarity-based analysis and angle between COVID-19 DNA sequence and other sequences

Virus name	Cosine of angle between COVID-19 virus vector and the considered virus, the (value)	Angle between COVID-19 & the virus cos−1 (value)
COVID-19 virus	1	0°
SARS virus	0.92229834	22.7355694°
MERS virus	0.89049951	27.06391766°
HIV virus	0.06197418	86.44686409°
Ebola virus	0.787751	38.02416713°
Dengue virus	0.65690051	48.93608574°
Rotavirus	0.05758748	86.69865401°
Hanta virus	0.64873133	49.55398225°
Spanish flu	0.31512345	71.63173348°
Swine flu	0.32418355	71.08388204°

(ML)-based clustering algorithms for the sequential analysis is under progress and the future scope for this research.

6 Conclusion

In this paper, a comparative analysis of the most prominent disease-causing viruses has been made. The objective is to show the differences between their gene structures and to help the researchers, medicine company and bioinformatics practitioners. This paper serves as the intermediate analysis of the COVID-19 against several viruses. Algorithms for plotting the DNA sequence data and finding conclusions from them have been made. Also, we have come at a consensus that overall research and trends can be profoundly studied when the viruses are considered as vectors in a plane. It has been proven in the analysis that the GC content—the heat stability of the genome—is least in case of the HIV and a slightly less than 39% for the COVID-19 virus. From this analysis, it is stated that the

Table 7 Cosine similarity-based analysis and angle between COVID-19 DNA sequence (countrywise)

Mutant virus country specific/NCBI ID	Angle between Wuhan COVID-19 DNA and the considered virus variant. cos−1 (value). (In Degrees)	Cosine of angle between Wuhan COVID-19 virus DNA vector and the considered virus variant
Australia/MT007544.1	0.68362699	0.99992882
Bahrain/MW332535.1	2.94291781	0.99868118
France/MT470127.1	0.91760131	0.99987176
Germany/MT270108.1	2.78379925	0.99881991
Greece/MT459897.1	2.75083162	0.99884769
India/MW243003.1	2.85687987	0.99875715
Italy/MW423686.1	3.20000429	0.99844076
Japan/LC529905.1	0.54542658	0.99995469
Netherlands/MT705205.1	0.83872064	0.99989286
Saudi-Arabia/MT755890.1	1.05335376	0.99983101
South-Korea/MT039890.1	0.95329293	0.99986159
Spain/MW375727.1	2.77313779	0.99882893
Thailand/MT447155.1	2.66007165	0.99892246
Tunisia/MW426406.1	1.55209695	0.99963311
USA (California)/MW306388.1	2.96895029	0.99865775
USA (Minnesota)/MW349056.1	2.95714573	0.9986684
USA (New Mexico)/MW269901.1	89.7553073	0.00427068
USA (Wisconsin)/MW342048.1	46.1985380	0.69216159
UK (England)/MW059036.1	1.18159272	0.99978736
Vietnam/ MT192773.1	2.83321587	0.99877765
Zambia/MT790522.1	2.68441672	0.99890265

annealing temperature of COVID-19 virus is low. This is the reason behind COVID19 varying forms (different and multiple forms) in different continents and countries.

The COVID-19 sample from Wuhan (November 2019) may not resemble the samples collected from other parts of the world. Rotavirus and HIV were being discussed in various medical lobbies as the potential targets for the genetic-based medical analysis and vaccine preparation. Our study also states that the genetic composition of the Rota and HIV resembles only 21% and 57%, respectively, with COVID-19. The cosine similarity-based analysis used in this paper has shown that the two viruses are inclined at an angle greater than 86º may not be the targeting genomes for the vaccine preparation [25]. Further, the mathematical cosine symmetry states that the angle between viruses is in the order SARS < MERS < Ebola < Dengue < Hanta < Swine Flu = Spanish Flu < HIV < Rotavirus. Therefore, the opposite is the order of similarity with the COVID-19 vs. SARS which is being the most similar and Rota being least similar to the COVID-19.

As discussed in this research, the results are in compliance with the fact that the cosine similarity-based analysis can be employed to find the similarity between the viruses. Our approach has observed over 96% accuracy. The TF-IDF-based approach is faster, more reliant and consistent. It has been used to find the weights of k-mers distributed in the whole DNA sequence. These assigned weights have helped us in finding the perfect angle between the any two concerned viruses. Given the mutating nature of COVID-19 genome, it implies that the methodologies employed here can also be devised to consider the DNA sequences provided dynamically as an input to our algorithm. We have considered several such mutating samples obtained after the severe outbreak in March 2020 to December 2020 all around the world, in 21 major geo-locations. Thus, no matter the occurrence of new strands or mutation, the given approach can thus be used for every possible use case of COVID-19 or in general any virus/DNA genome variant.

References

1. Blot, A.: Fuzzy edit sequences in genetic improvement. In: 2019 IEEE/ACM International Workshop on Genetic Improvement (GI), Montreal, pp. 30–31. QC, Canada (2019)
2. Patra, P.; Izawa, T.; Peña-Castillo, L.: REPA: applying pathway analysis to genome-wide transcription factor binding data. IEEE/ACM Trans. Comput. Biol. Bioinform **15**(4), 1270–1283 (2018)
3. Wang, R.Y., Guo, T.Q., Li, L.G., Jiao, J.Y., Wang, L.Y.: Predictions of COVID-19 infection severity based on co-associations

between the SNPs of co-morbid diseases and COVID-19 through machine learning of genetic data. In: 2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT), pp. 92–96. Dalian, China (2020)
4. Lv, J., Tu, S., Xu, L.: Detection of phenotype-related mutations of COVID-19 via the whole genomic data. IEEE/ACM Trans. Comput. Biol. Bioinform.
5. Campbell, M.: DNA data storage: automated DNA synthesis and sequencing are key to unlocking virtually unlimited data storage. Computer **53**(4), 63–67 (2020)
6. Bullard, J.H.; Purdom, E.; Hansen, K.D., et al.: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinform. **11**, 94 (2010)
7. Irizarry, R.A., Gautier, L., Cope, L.M.: In: Parmigiani, G., Garrett, E.S., Irizarry, R.A., Zeger, S.I. (eds.) The Analysis of Gene Expression Data: Methods and Software, pp. 102–119. Springer-Verlag, New York (2003)
8. Irizarry, R.A., et al.: Multiple-laboratory comparison of microarray platforms. Nat. Methods **2**, 345–350 (2005)
9. Dlamini, G.S., et al.: Classification of COVID-19 and other pathogenic sequences: a dinucleotide frequency and machine learning approach. IEEE Access **8**, 195263–195273 (2020)
10. Srikanth, M.: Application of Cosine Similarity in Bioinformatics; 2018, MS Thesis, University of Nebraska, Lincoln. [accessed on June 29 2020]
11. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 40410 (1990)
12. Liu, S., et al.: Efficient cryo-electron tomogram simulation of macromolecular crowding with application to SARS-CoV-2. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), , pp. 80–87. Seoul, Korea (South) (2020)
13. Zhang, J.; Guo, H.; Hong, F.; Yuan, X.; Peterka, T.: Dynamic load balancing based on constrained K-D tree decomposition for parallel particle tracing. IEEE Trans. Vis. Comput. Graph. **24**(1), 954–963 (2018)
14. Newberg, L. A.: Error statistics of hidden Markov model and hidden Boltzmann model results. BMC Bioinform. **10**(1) (2009)
15. Xiao, M., et al.: 2019nCoVAS: developing the web service for epidemic transmission prediction, genome analysis, and psychological stress assessment for 2019-nCoV. IEEE/ACM Trans. Comput. Biol. Bioinform.
16. Zhu, Z.; Liang, J.; Li, D.; Yu, H.; Liu, G.: Hot topic detection based on a refined TF-IDF algorithm. IEEE Access **7**, 26996–27007 (2019)
17. Pan, T.; Flick, P.; Jain, C.; Liu, Y.; Aluru, S.: Kmerind: A flexible parallel library for K-mer indexing of biological sequences on distributed memory systems. IEEE/ACM Trans. Comput. Biol. Bioinform. **16**(4), 1117–1131 (2019)
18. Sadik, A.Z., & Hussain, Z.M.: Short word-length LMS filtering. In: 9th International Symposium on Signal Processing and Its Applications. (2007) https://doi.org/10.1109/isspa.2007.4555427
19. Staden, R.: A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. **6**(7), 2601–2610 (1979)
20. De Bruijn; N.G.: A Combinatorial Problem. In: Koninklijke Nederlandse Akademie V. Wetenschappen **49**, 758–764 (1946)
21. Piotr, I., Rajeev, M.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, p 604613 (1998)
22. Zhang, J., et al.: Navigating the pandemic response life cycle: molecular diagnostics and immunoassays in the context of COVID-19 management. IEEE Rev. Biomed. Eng.
23. Robson, B.: Computers and viral diseases preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Comput. Biol. Med. **119**, 103670 (2020)
24. National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1988 [Cited 2020 June 10]. Available from: https://www.ncbi.nlm.nih.gov/
25. Robson, B.: Preliminary bioinformatics studies on the design of synthetic vaccines and preventative peptidomimetic antagonists against the wuhan seafood market coronavirus. Possible importance of the KRSFIEDLLFNKV Motif (2020)