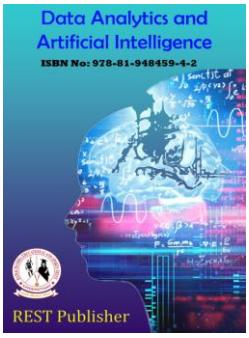
Data Analytics and Artificial Intelligence

Vol: 3(1), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/3/1/3



Streambit Dashboard For DNA Analysis

* **N.Sharmila Banu, K.Soundharya, C.Madula, Sarang S Babu, Y.Vineet Narayan, G.Rohit Kumar**

veltech hightech dr.rangarajan dr,sakunthala engineering college, Chennai,Tamil Nadu, India.

*Corresponding Author Email: jarmi.it@gmail.com

Abstract: *The humankind has faced severe effects on the virus. By analyzing the similarities in the characteristics and inheritable patterns, the analyst can get a better understand about the virus and this may helps to determine the cure or the drug. With an overall viral relationship with the accuracy rate of 96, the trial results are more encouraging.*

Keywords: *Humankind, inheritable, delicacy*

1. INTRODUCTION

The technique for deciding the sequence of the nucleotides in a certain or specified DNA molecule is known as DNA sequencing. The technique of extracting DNA from the genetic makeup of an organism and it is a usual practice in genetic engineering to join it to the deoxyribonucleic acid of another general thing. Analysts understand how to improve or change the characteristics pertaining to a biological system any creature, out of a pathogen to a cow, dog, or other animal, can be genetically engineered. This method can also be employed to investigate viral diseases that have a high level of mortality in animals and the environment. This can help them survive by assisting in the life of their human being. Traditional implant methods exploited in animals and plants are frequently comprehensive, leading to the increase of undesirable genes with desirable one. The Genetic Engineering mechanism involves the creation of tenacious DNA, which allows one to separate and implant only the desired genes into direct organisms. Protein synthesis is the primary route for amino acid disposal. These acids were activated by the binding of specific molecules involved in the transmission of gene instructions from DNA to nRNA in the organism's nucleus. This synthesis can be accomplished through two methods: transcription and translation. The process of converting a phase of DNA into RNA is called as Transcription. After the transcription of DNA into RNA in a living organism's cell nucleus, the ribosome or endoplasmic reticulum creates protein as part of the translation. This comprehensive method was known as Translation. In the case of transcription, the molecules use RNA as a guidebook and take up the DNA strand. Nucleotide molecules are the molecules found in both DNA and RNA. A single DNA strand is made up of a chain of nucleotides. These nucleotides are composed of phosphate groups, sugar groups, and four different types of nitrogen bases. These bases were Guanine, Adenine, Thymine, and Cytosine. The DNA structure is double helical, with two strands interlocked and nitrogenous bonds present. Although there are four nitrogenous bases, three of them are similar in DNA and RNA. SARS-CoV, Ebola virus, and Severe acute respiratory pattern were among the notable viruses. Middle East respiratory pattern (MERS) (EMC/2012) associated with the corona virus, Spanish flu (H1N1 Virus), and the recent trend COVID-19 (SARS-CoV-2) were some of the pandemic situations that had made life hard for the human race to survive. Recently, millions of people around the world are at risk of falling into extreme poverty. The majority of the viruses mentioned had similar characteristics and inheritable patterns. Researchers can gain a deeper understanding of the virus by analyzing similarities in characteristics and inheritable patterns, which may aid in the discovery of a cure or drug. To label these issues, we proposed an experimental based analysis for the specified virus data using correlation styles. The sequence considered for

this analysis includes some colorful amino acids, protein sequences, 3D modeling of the virus. This analysis were used by the scientists and consortiums such as WHO (World Health Organization), computational bio logistics, and inheritable masterminds to create a framework for studying DNA sequence distributions. The Biopython were used to illustrate the analysis of the noticeable virus and for the logical conclusions about the results using 3D modeling.

2. RELATED WORK

One of the most difficult problems with outburst of biological data is determining how to algorithmically assess their functions and structures. Deep learning methods play a significant role. The 3 major processes of feature extraction, predictor development, and performance analysis are typical for predictors backed by machine learning techniques. Although numerous web applications and comprehensive devices have been developed to assist in analysis of genetic sequences, they only focus on individual steps. Numerous studies have used pathway analysis to assist the understanding of the findings of transcription throughout the genome identification analyses, and it have been effectively used to identify the relationships among the genetic research topic and the genetic basis. Over-representation analysis and sequence set analysis are the two most used approaches to pathway analysis. With the recent expansion of genome-wide transcription factor binding knowledge, pathway analysis can now be applied to the existing body of knowledge. In this work, we created the regulative enrichment during this study; an analytic model based on individuals' genomes is developed to forecast how they will react to COVID-19. Stopping the spread of COVID-19 infections requires pinpointing those who are vulnerable to infection. Previous studies have shown that individuals with comorbid illnesses are more susceptible to infection and the development of a wide range of significant COVID-19 disorders. Moreover, the severity of each patient's COVID-19 infection and hence the squared patterns were only identified through reciprocality studies comparing patient phenotypes. In this study, specific genes behind the observed symptom patterns were analyzed using machine learning of the COVID-19 knowledge from Genome - wide association studies. This can reveal physiological mechanisms underlying COVID-19 decrease that are crucial for the development of efficient and targeted medicine.In order to identify important single ester polymorphisms (SNPs) in the phenotypes of symptomatic versus symptomless, early versus late assortment time, recent versus young, and male versus female, this study uses both single-locus analysis and joint-SNPs analysis. The association between any two SNPs is also examined in this paper using linkage state of affairs analysis, as well as the patterns of SNP additive mutations across time. There are three results in total. First, it is found that the SNP at ester position 4321 is a freelance vital locus connected to all three major traits. Deoxyribonucleic acid has the capability to store innumerable times a lot of knowledge than best modern technologies however faces many discouraging challenges. The acid synthesis and accompanying sequencing automation are the key factors in creating storage medium with virtually infinite capacity. [6] They evaluated mathematical techniques for identifying genes that were considerably First State in two different kinds of biological data and found that the check statistics handled low-count genes quite differently depending on the sample type. First State outcomes, however, are awash in sequencing platform options like varying sequence lengths, base-calling activity approach (with and without letter of the alphabet X management lane), and flow-cell/library preparation impacts. We study how the browse count standardization technique affects First State outcomes and show that the quality approach of scaling by total lane counts will skew First State estimations. [7] It provides compact summary of data-analytic tasks related to microarray studies, tips that could chapters that may facilitate perform these tasks, and connections with designated data-analytic tools. [8] A group of 10 Research centres were formed to match information attained from 3 wide used platforms exploitation of the same ribonucleic acid samples. They have a tendency to accept applied mathematics analysis to show clearly that there are comparatively massive variations in data obtained from the centre's using identical platform, the most efficient results from the labs have been gathered.

3. EXISTING SYSTEM

In this study, deoxyribonucleic acid analysis is used to detect great single ester polymorphism (SNPs) within the phenotypes of symptomatic vs. well, first assortment time vs. the late assortment time the young, vs. the old and also the female vs. the male. In addition, linkage situation analysis is used to determine the relationship between any two SNPs, and it visualizes the model of additive variations of SNPs over the time. The outcomes are threefold. Firstly, the SNP at

ester location 4321 was discovered to be a freelance important site connected with all the 3 primary compositions. Furthermore, twelve substantial SNPs were discovered in the first two studies. Secondly, the cistron ORF1ab holding SNP-4321 is found to be significantly related to the first three compositions, and the three genes S, ORF3a, and N are found to be important in the first two compositions. Thirdly, a number of the noticed genes or SNPs have been linked to SARS-COV-2 in the literary study, indicating that the findings here could be useful for any research.

4. PROPOSED SYSTEM

This model runs through the Biopython module. It includes the parsers for numerous bioinformatics train formats such as BLAST, FASTA, and Genbank and supports bioinformatics merchants such as NCBI and Expasy interfaces. This framework provides essential tools such as standard sequence-based orders and agglomeration strategies, as well as a number of information structures such as the KD tree. The dataset here used was in the FASTA format. The upper limit of the dataset should be approximately 200MB. After loading the deoxyribonucleic acid sequence in FASTA format, we select one of two primary stages: I order correlation-based analysis or (ii) order alignment-based analysis. The overall goal of this study is to discover a numerical and illustration correlation between all of the pathogens under consideration, as well as similarity between their order sequences, which may aid in medical cure, narcotic evaluation, and vaccinum development. The entire DNA/Genomic sequence of viruses such as COVID/SARS/MERS is obtained from the public ester sequence database: The National Centre for Biotechnology information (NCBI)." The deoxyribonucleic acid sequence information is stored in FASTA format. The machine learning and deep learning algorithms uses numerical rather than categorical information in preprocessing the data. The polymer dataset's genomic sequence is categorical. Numerous methods are available to transform categorical data into numerical data. The method of converting explicit verbal information into numerical form is known as cryptography Hamming Distance (HD) = $\sum z\ s=1\ |Ps - Qs\ |\ P = Q \Rightarrow HD = 0\ P \neq Q \Rightarrow HD$

5. ARCHITECTURE DIAGRAM

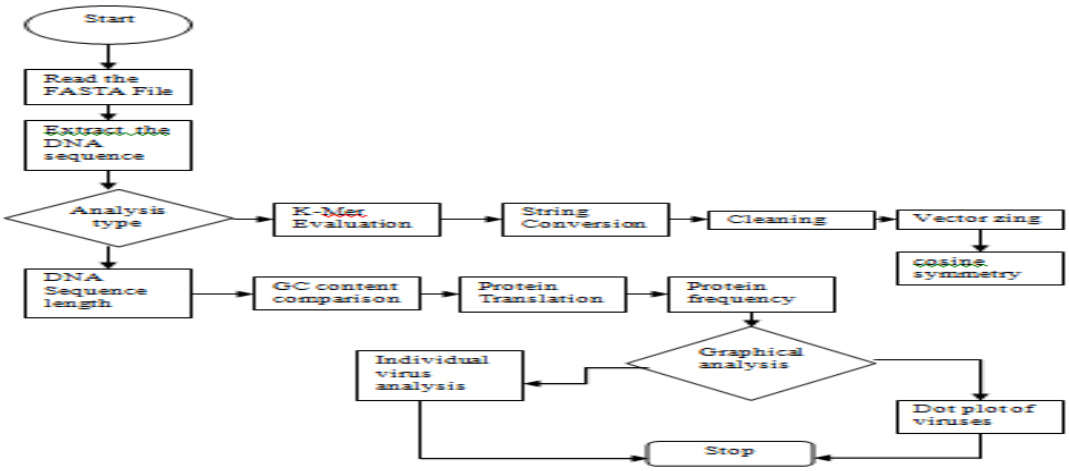


FIGURE 1: Architecture diagram for DNA sequence

6. RESULT

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 cordon table. The cordon is a trinucleotide sequence of DNA that

Corresponds to an amino acid. The three-stop cordon marks the end of the sequence of the protein. One start cordon, 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

FIGURE 2: DNA standard codon table

With a motivation to compare the trends of the COVID19 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

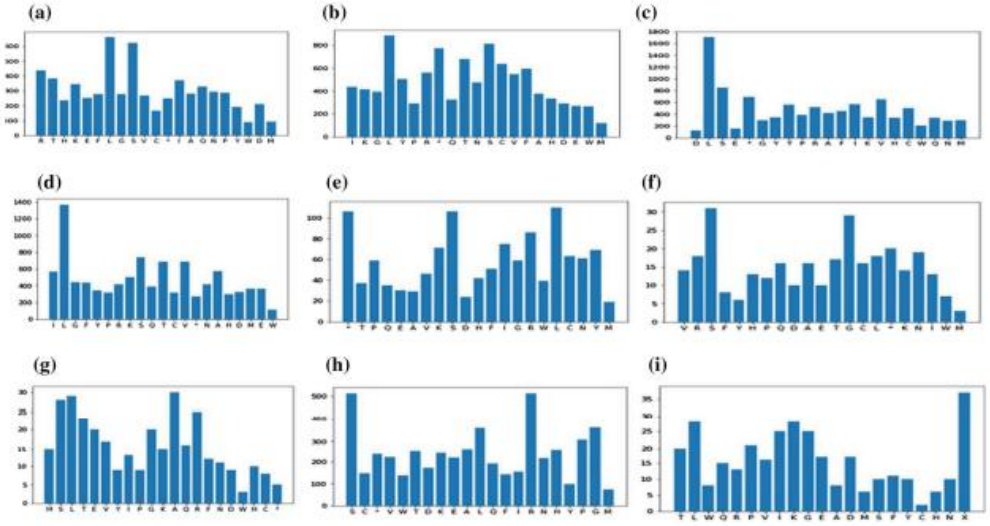


FIGURE 3: Protein sequence distribution of various virus such as (i) COVID-19, (ii) Ebola virus, (iii) MERS, (iv) SARS virus, (v) Hanta virus, (vi) Spanish fu virus, (vii) Dengue virus, (viii) Swine flu virus, (ix) Rota

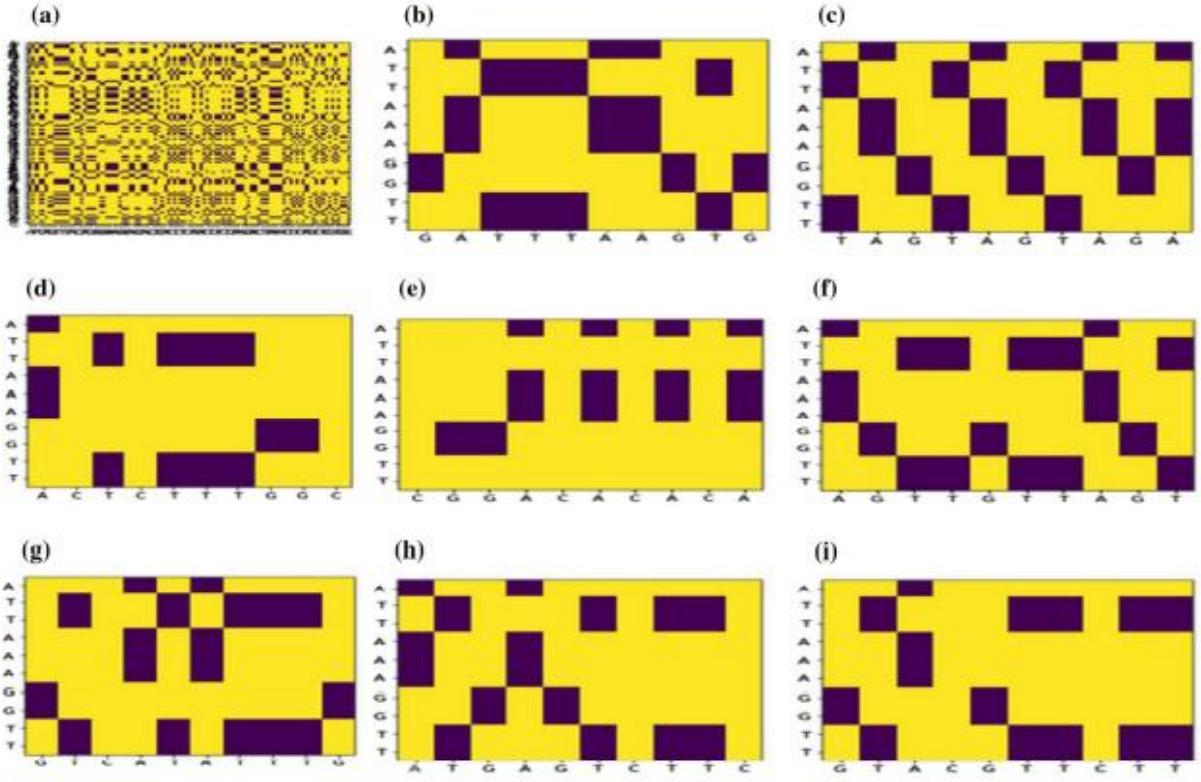


FIGURE 4: Dot plot-based comparison of the COVID-19 versus the various virus such as (i) SARS (ii) Ebola virus, (iii) MERS, (iv) Hanta (v) Spanish fu (vi) Dengue virus, (vii) Swine flu virus, (viii) Rota

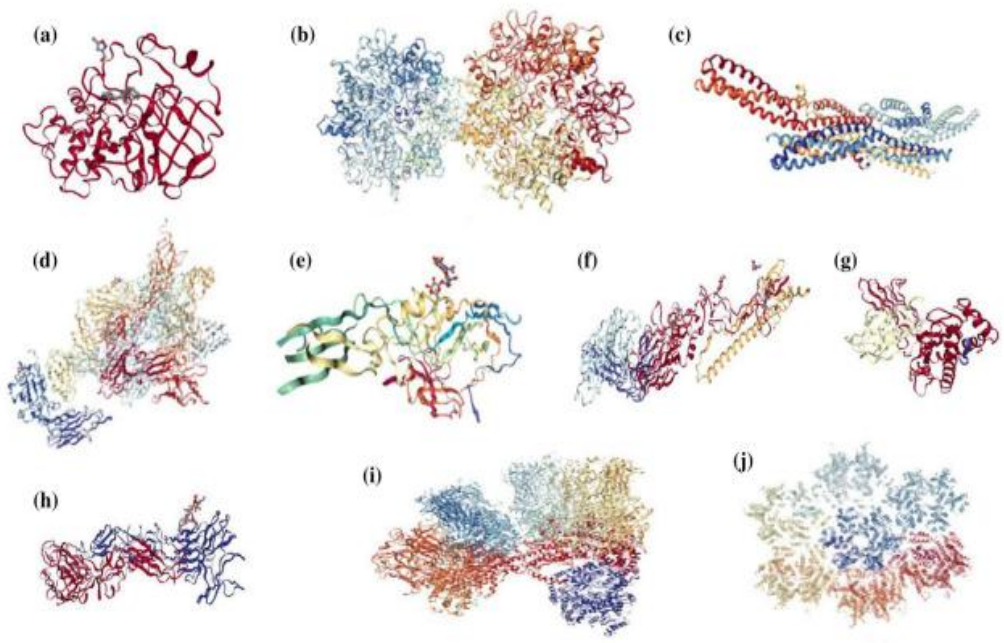


FIGURE 5: Visualization (3D model) of various virus such as (i) COVID-19, (ii) Ebola virus, (iii) MERS, (iv) SARS virus, (v) Hanta virus, (vi) Spanish fu virus, (vii) Dengue virus, (viii) Swine flu virus, (ix) Rota

7. CONCLUSION

A comparison of the common well-known disease caused viruses had conducted. The goal is to highlight the similarities in their sequence structures in order to assist analyzers, drug companies, and other well known sources. Our paper is a comparative analysis of COVID-19 versus a variety of viruses. Deoxyribonucleic acid sequence data is plotted using algorithms that draw inferences from the data. Additionally, we have come to the conclusion that broad studies and trends are frequently thoroughly examined once viruses are thought of as vectors in a very plane. The Gc content has been well- tested within the analysis, with the HIV virus having the lowest and the COVID-19 virus having the highest. The COVID-19 samples have collected from the cities from the overall. In numerous medical forums, reovirus and HIV were mentioned as possible points for studies related to medical genetics and vaccinum development. Based on the fin Ddings, the genetic composition of Rota-associated HIV is only twenty one and fifty seven percentage similar to COVID-19, respectively. The cos based similarity analysis is used for our study that revealed the two pathogens should be inclined at an angle greater than 86 degree and may not for the targeting of genomes for vaccinum preparation. Based on the mathematical cosine symmetry, the angle between viruses is in the order SARS MERS Dengue Ebola Swine Flu Hanta = Rotavirus Spanish Flu HIV. As a result, the order of uniformity was reversed, with COVID-19 for being the most liked to SARS and Rota for being the least similar to the COVID-19. As we concluded on our study, the findings were consistent with fact that the technique of the cos similarity-based analysis is frequently used to determine virus similarity. Our method had determined more than 96% accuracy. It is common to determine the weights of k-mers that distributed throughout the entire deoxyribonucleic acid structure. Those assign weight that had aided North American countries in determining the optimal angle of whatever pathogens involved. The variation nature of the genome of COVID-19, the methods that were used here could be adapted to consider the DNA structure that provide actively as input to our method. We had thought of numerous such mutations of the specimens acquired once the severe event occurred in March 2020 to December 2020 in world-wide, in twenty one major geo-positions. As a result, regardless of the presence of recent strands or mutations, approach that we had used here has got some valuable result with the usecase of COVID-19.

REFERENCES

[1]. Liu, Bin. "BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches." *Briefings in bioinformatics* 20.4 (2019): 1280-1294.

[2]. Patra, P.; Izawa, T.; Peña-Castillo, L.: REPA: applying pathwayanalysis 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-associationsIn: 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 mutationsof COVID-19 via the whole genomic data.IEEE/ACM Trans.Comput. Biol. Bioinform.

[5]. Campbell, M.: DNA data storage: automated DNA synthesis andsequencing 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 diferential expressionin 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 GeneExpression Data: Methods and Software, pp. 102–119. SpringerVerlag, 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.: Classifcation of COVID-19 and other pathogenic sequences: a dinucleotide frequency and machine learningapproach. IEEE Access 8, 195263–195273 (2020)

[10]. Srikanth, M.: Application of Cosine Similarity in Bioinformatics;2018, MS Thesis, University of Nebraska, Lincoln. [accessed onJune 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.: Dynamicload balancing based on constrained K-D tree decomposition forparallel particle tracing. IEEE Trans. Vis. Comput. Graph. 24(1)

[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 detectionbased on a refned TF-IDF algorithm. IEEE Access 7, 26996–27007 (2019)

[17]. Pan, T.; Flick, P.; Jain, C.; Liu, Y.; Aluru, S.: Kmerind: A flexibleparallel library for K-mer indexing of biological sequences ondistributed memory systems. IEEE/ACM Trans. Comput. Biol.Bioinform. 16(4), 1117–1131 (2019)

[18]. Sadik, A.Z., & Hussain, Z.M.: Short word-length LMS fltering. (2007) https://doi.org/10.1109/isspa.2007.4555427

[19]. Staden, R.: A strategy of DNA sequencing employing computerprograms. 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 In: Proceedings of theThirtieth Annual ACM Symposium on Theory of Computing, p604613

[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; [Cited 2020 June 10]. Availablefrom: https://www.ncbi.nlm.nih.gov/

[25]. Robson, B.: Preliminary bioinformatics studies on the design ofsynthetic vaccines and preventative peptidomimetic antagonist against the wuhan seafood market coronavirus. Possible importance of the KRSFIEDLLFNKV Motif (2020)