Pakistan Journal of Engineering and Technology, PakJET
Multidisciplinary & Peer Reviewed
Volume: 4, Number: 4, Pages: 41- 45, Year: 2021

Analysis of Covid-19 Genome Sequences based on Geo-Locations

Aqsa Umar*, Naeem Ahmed Mahoto, Sania Bhatti, Sapna Rathi

Department of Software Engineering, Mehran University of Engineering Technology, Jamshoro, Pakistan
*Corresponding author's email: Aqsa Umar (aqsa.umar3505@gmail.com)

***Abstract*- COVID-19 pandemic has become a major worldwide serious health risk of the current 21st century. One month after emerging of the deadly virus, the unique genomic sequence was found. It is necessary to examine the genomic sequences of COVID-19 strains to fully understand the virus's behavior, origin, and how rapidly it mutates. In this research, we have looked at the usage of sequential pattern mining *SPM*, a closed sequential pattern technique to discover valuable information from COVID-19 genome sequences *CGS*. The analysis is performed on the three strains of China, Pakistan, and India. Three countries' CGS strains are selected based on the most populated geo-location and the strains are taken from the national center for biotechnology information *NCBI*. Furthermore, in this research first, the sequences data files of genome sequences are being transformed to the computer-readable corpus of CGS and then the SPM technique is applied to discover the frequent patterns of nucleotides. Second, from the medical guidelines, the frequent codons of amino acids are extracted from three strains of genome sequences. Third, we have also evaluated the performance of the proposed approach in terms of time execution, the number of frequent patterns, and memory consumption. Obtained results suggested that the pattern ACA that encodes *Threonine* amino acid with support 1576 in Pakistan is the most frequent pattern from the other two strains. Moreover, the closed sequential pattern mining using sparse and vertical id-lists CloFAST algorithm performance evaluated that when the user minimum threshold value is low, the high number of frequent patterns are extracted with more time and memory consumption.**

Keywords: **COVID-19, Sequential pattern mining, Genome sequences, Closed sequential patterns, Nucleotide bases, Amino acid codons.**

I. INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus strain causes Coronaviruses disease, in late December 2019, the origin of the COVID-19 occurred in the Wuhan City of Hubei Province of China [1]. Later, the World Health Organization declared this virus on March 11, 2020 [2]. While writing this study [2-5], more than 200 countries have been infected with this virus, with more than 300 million infections in people and more than 5 million deaths globally [3].

SARS-CoV-2 is an RNA virus, with large RNA genomes ranging in length from 27 to 32 kilobases [4]. On January 12, 2020, Chinese researchers revealed the genetic sequence of COVID-19, which is said to be a COVID-19 genome sequence [4], Understanding the genomic sequence of SARS-CoV-2 is necessary to fully understand the behavior of this unique pandemic virus. COVID-19 genome sequence has been demonstrated to be exceedingly heterogeneous, challenging to grasp its many clinical and epidemiological characteristics. In our study, based on the high population rate, Asian countries with geolocation China, Pakistan, and, India COVID-19 strains are taken to learn more about the characteristics of the COVID-19 genome sequence.

According to World Health Organization last update [3], in China, Pakistan, and India there have been 134,711, 1,309,248, and 36,070,510 confirmed cases with 484,655 28,987, and 1,309,248 deaths of COVID-19. Still, many actionable insights of the genome sequence of covid-19 are unknown and more research is needed to investigate the behavior of this epidemic virus.

To learn more about the CGS, the predictive models with the help of artificial intelligence are built-in studies [5-6]. Alignment-Free [7-10], approaches are used to examine and comprehend the patterns and intrinsic features of biological sequences, whereas COVID-19 virus genomes are classified with the help of machine learning-based alignment-free methods [6] and deep learning for analysis of COVID-19 disease [7].

Pattern analysis allows humans, particularly bioinformaticians, to examine complex and vast genetic and genomic data [11-24]. Artificial technologies such as deep learning [25, 26] can be used to assist speed up the process by predicting which existing medications or new drug-like compounds would be effective against COVID-19. Moreover, in studies [26-29] machine learning, artificial intelligence and deep learning are used for the early detection of SARS-CoV-2. In study [27] 553 complete datasets of genome non-repeated sequences are used with a convolutional neural network which generated features automatically based on the virus's genome sequences with an average accuracy of 98.75%, further in the study [28] deep learning is again used for COVID-19 detection from the X-ray imaging.

SPM algorithms for mining closed sequential patterns are used to discover sequential frequent patterns from the sequence databases [11-12]. Moreover, in our study, we are using SPM closed sequential pattern mining algorithm for discovering hidden information present in the COVID-19 genome sequences, for that reason we have taken the strains of three countries based on the most populated geolocation i-e China, Pakistan, and India. This study can provide new insights about the virus genome sequence

and that can be helpful for biological researchers to understand more about the behavior and characteristics of COVID-19.

The goal of this whole study is to analyze more about CGS based on geolocations. More specifically, the SPM algorithm, for the mining closed sequential patterns is used on CGS:

- To extract frequent patterns of nucleotides in CGS.
- To analyze more about extracted frequent nucleotides by encoding codon(s) of amino acids.
- To show the availability of frequent codon(s) of amino acids available in the strains of China, Pakistan, and India. Moreover, similar & dissimilar frequent patterns in the three strains.
- To evaluate performances of the CloFAST algorithm on genome sequence data based on time execution, the number of frequent patterns, and memory consumption.

The rest of this paper is organized as follows. Section II presents the workflow of the proposed research study where SPM approach is used to discover nucleotides encoding codon(s) and their frequent patterns in genome sequences, obtained results are discussed in Section III. Finally, the paper is concluded with some remarks in Section IV.

II. METHODOLOGY

CGS are analyzed to identify hidden patterns using SPM. The proposed methodology contains four main steps: (a) data collection, (b) data transformation, (c) data discovery (d) data evaluation, shown in Fig. 1.

(a). DATA COLLECTION

The strains of Covid-19 for China, Pakistan, and India are collected from the SARS-CoV-2 Data Hub available in the NCBI database [19]. The genome sequences contain sequences of nucleotides, NCBI SARS-CoV-2 data hub offers to download each sequence in the form of proteins or nucleotides [19-20]. For our study, we have used FASTA format genome sequences of China, Pakistan, and India in the form of nucleotides. Table 1 shows the statistics about the collected genome sequence and ID.

Table 1 shows the accession number of genome sequences with the collected information about the sequence geolocation, length, collected, and released date.

TABLE 1: CGS based on geo-location

ID	Geo Location	Length	Collection Date	Release Date
MZ824622	China	29903	2020-02-03	2021-08-18
MZ562707	Pakistan	29831	2021-04-30	2021-07-15
OK189654	India	29871	2021-09-10	2021-09-21

(b) DATA TRANSFORMATION

Each line in Covid-19 strains shows genome sequence, Table 2 contains three lines (genome sequences) with IDs 1,2, and 3. The four raw nucleotides (A = Adenine, T = Thymine, C = Cytosine and G = Guanine) are then transformed into digital format (1, 2, 3 and 4). In Fig. 1. b) Nucleotides are first transformed to sequences database, and then into SPMF format whereas -1 indicates the end of item set and -2 indicates the end of the line in SPMF format.

TABLE 2: A Sample of CGS

ID	Sequences
1	(…AGGTAACAAACCAACCAACTTTCGATCTCTTGTA…)
2	(…ATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGT…)
3	(…AATTACTGTCGTTGACAGGACACGAGTAACTCG…)

(c) DATA DISCOVERY

In Fig. 1. c) Data discovery shows that a SPM algorithm is used for extracting frequent patterns from the covid-19 genome sequence. In this research, we have used the CloFAST algorithm for discovering closed sequential patterns in sequence databases.

(d) DATA EVALUATION

The frequent patterns are then extracted, shown in step d) of Fig. 1. The extracted frequent patterns are based on multiple of three nucleotide base(s) encoding codon(s). Furthermore, from the frequent codon(s) the amino acids are encoded during the data evaluation step.

III. RESULTS AND DISCUSSION

This section presents and discusses the results obtained by applying the mining closed sequential patterns algorithm to the strains of China *MZ824622*, Pakistan *MZ562707*, and India *OK189654*, strains, which can be found in [22]. All experiments were performed on the SPMF, the open-source data mining library developed in JAVA. The release version of SPMF provides a command-line interface and a graphical user interface [9]. Table 3 shows the total number of nucleotides present in the strains of CGS of China, Pakistan, and India. Subsection a) describes frequent nucleotide patterns extraction. However, subsection b) evaluates the performance of the approach used in this study.

TABLE 3: CGS nucleotides

ID	Geo Location	A	C	G	T
MZ824622	China	8923	5481	5863	9608
MZ562707	Pakistan	8915	5474	5852	9590
OK189654	India	8924	5477	5862	9608

a) FREQUENT NUCLEOTIDES EXTRACTED BY CLOFAST

Frequent nucleotide bases have been extracted from a sequence database of Covid-19 strains with the help of the CloFAST algorithm. The CloFAST requires setting the minSup threshold as input to execute the frequent patterns [13]. With the given minimum threshold values of 20%, 50%, 90% and 100%. the CloFAST algorithm discovered the closed frequent patterns from Pakistan, India, and China genome sequences. The following subsections c) and d) describe more details about the frequent patterns extracted from CGS.

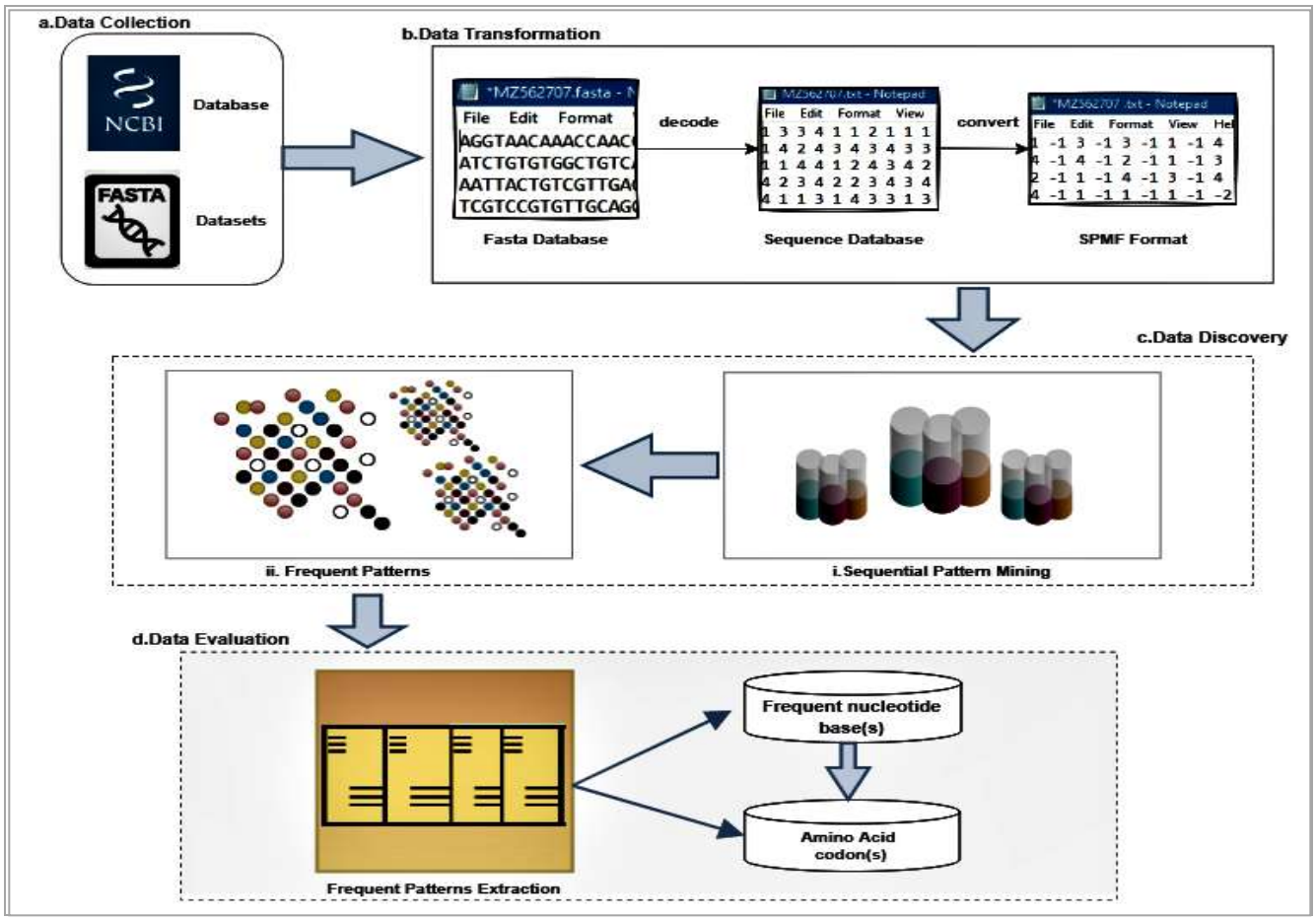

FIGURE 1: Main Phases of the Proposed Approach

b) FREQUENT NUCLEOTIDE EXTRACTION

With the maximum minSup i-e., 100% no pattern came out, but patterns count came out as 4, 3, and 48 from the covid-19 strains of China, Pakistan, and India with minSup 90% using CloFAST Algorithm. Four nucleotides of China, two nucleotides of Pakistan, and four nucleotides of India with support percentage values are shown in Tab. 4. The maximum support percentage of *6.4%* of frequent nucleotide *(A)* came out from Pakistan whereas the minimum support percentage of *4.6%* of frequent nucleotide *(C)* came out from China.

TABLE 4: Frequent nucleotide support percentage

Frequent Nucleotide	Support (%) in China	Support (%) in Pakistan	Support (%) in India	Min.Sup
A	1440(4.8%)	1921(6.4%)	1492(5.0%)	90%
C	1378(4.6%)	1867(6.3%)	1474(4.9%)	90%
G	1381(4.6%)	–	1470(4.9%)	90%
T	1458(4.9%)	–	1491(5.0%)	90%

c) FREQUENT PATTERNS ENCODING AMINO ACIDS CODON

An amino acid is encoded by a three-nucleotide sequence [13]. From the literature [13-14], expect from the Tryptophan and the Methionine amino acids, most of the amino acids have more than one codon, there are $4^3 = 64$ unique codons, 61 of which encode the 20 amino acids and the remaining three are stopping codons. Table 5 shows the most frequent patterns extracted from the genome sequences of China, Pakistan, and India. The codons of 20 amino acids can be seen in studies [16-18], out of 20 Amino acids 7 frequent amino acids are extracted from the genome sequences of China, Pakistan, and India given in Tab. 5.

TABLE 5: Frequent codon(s) availability

Frequent Codon(s)	Full name of Amino Acid	China	Pakistan	India
AAA	Lysine	✓	✓	✓
AAG		✓	✓	✓
ATG	Methionine	✓	✓	✓
CAT	Histidine	✓	✓	✓
CAC		✓	✓	✓
GTT	Valine	✓	✓	✓
GTC		✓	✓	✓
GTA		✓	✓	✓
GTG		✓	✓	✓
TTC	Phenylalanine	✓	✓	✓
TTT		✓	✓	✓
TAT	Tyrosine	✓	✓	✓
TAC		✓	✓	✓
TGG	Tryptophan	✓	✓	✓

Moreover, Tab. 6 shows the extracted similar and dissimilar patterns from the genome sequences of China, Pakistan, and India using the CloFAST algorithm with varying minSup 50% and 20%. Table 6 also shows the support percentages of patterns whereas the pattern *ACA* from Pakistan has the maximum support value of 1576 that means the pattern *ACA* is the most frequent pattern with 5.6%. However, the pattern CGC from Pakistan has a minimum support value of 961 with 3.2%.

TABLE 6: Similar & dissimilar frequent patterns support percentages extracted by CloFAST algorithm

Frequent patterns	Support (%) in China	Support (%) in Pakistan	Support (%) in India	Min.Sup
ATT	1150(3.8%)	1562(5.2%)	1372(4.6%)	50%
ACT	1352(4.5%)	1086(3.6%)	1352(4.5%)	50%
ACC	1303(4.4%)	1017(3.4%)	1229(4.1%)	50%
AGG	1034(3.5%)	1339(4.5%)	1283(4.3%)	50%
CAT	1089(3.6%)	1339(4.5%)	1496(5.0%)	50%
CTT	1078(3.6%)	1364(4.6%)	1360(4.6%)	50%
ACA	1096(3.7%)	1576(5.3%)	1366(4.6%)	50%
CGC	–	972(3.3%)	1112(3.7%)	50%
CGC	1035(3.5%)	961(3.2%)	1088(3.6%)	50%
TTG	1114(3.7%)	1420(4.8%)	1383(4.6%)	50%
ACCGGG	–	1328(4.5%)	–	20%
AAACCG	1164(3.9%)	1443(4.8%)	1437(4.8%)	20%
AATTAG	1136(3.8%)	–	1498(5.0%)	20%
AATTAT	–	1315(4.4%)	1330(4.5%)	20%
AATTGA	1350(4.5%)	1245(4.2%)	1331(4.5%)	20%
AGGCGG	–	1316(4.4%)	–	20%

d) PERFORMANCE EVALUATION

To illustrate the practicality of our study, we examined the performance of the closed sequence extraction using the CloFAST algorithm onto the COVID-19 genome sequence with varying minimum support thresholds. The China genome sequence is used here for the performance, because of the reason that it has a larger length than the other two strains, given in table 1. The performance of the algorithm is evaluated under the three categories i-e., execution time, the number of frequent patterns, and memory consumption have been evaluated in Fig. 2-4. All the experiments were performed on an HP laptop with a seventh-generation Core i5 processor and 4GB RAM.

e) EXECUTION TIME

In Fig. 2, results are reported in terms of running time (seconds). The CloFAST execution time ranges between a few milliseconds from support threshold 40% to 90% and about 7.30 seconds for support threshold 20%.

f) NUMBER OF FREQUENT PATTERNS

In Fig. 3, the number of frequent patterns extracted with varying minimum support threshold values are shown. We observed that with the maximum threshold value of 90% minimum number of frequent patterns 4 were extracted and with the minimum threshold value of 20% a maximum number of frequent patterns 55554 were extracted.

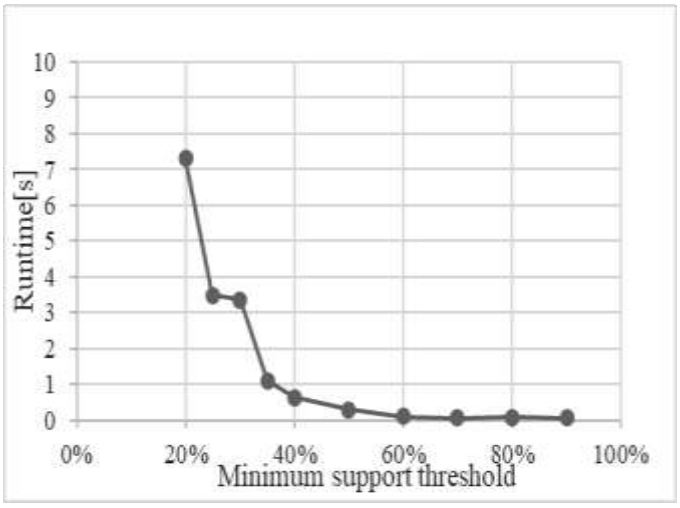

FIGURE 2. Execution time with varying minimum support threshold.

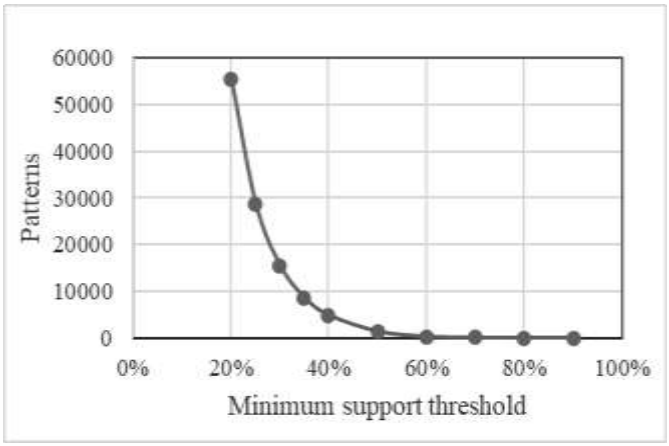

FIGURE 3. Number of frequent patterns with varying minimum support threshold value.

g) MEMORY CONSUMPTION

In Fig. 4 results are reported in terms of memory consumption(mb). We observed that the maximum memory consumption came out 871.38 mb with a minimum support threshold of 20% and minimum memory consumption came out 56.22 mb with a minimum support threshold of 80%.

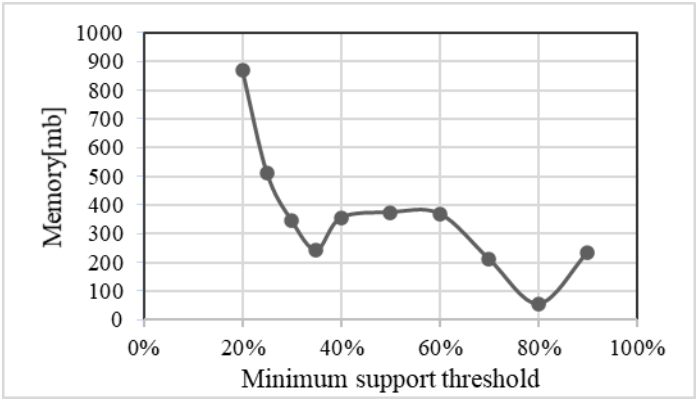

FIGURE 4. Memory consumption with varying minimum support threshold value.

IV. CONCLUSION

In this paper, we presented a SPM approach to analyze the COVID-19 genome. The approach is based on the extraction of frequent closed sequences. The strains of CGS of most populated countries China, Pakistan, and India are taken from NCBI's GenBank and then analyzed for: First, frequent nucleotide bases in the sequences are extracted with pattern mining techniques. then, frequent amino acids codons are extracted and finally, the performance feasibility of the used approach is evaluated. Our obtained results show that out of 20 amino acids 7 amino acids codons are available in the genome sequences of China, Pakistan, and India whereas Lysine, Methionine, Histidine, Valine, phenylalanine, Tyrosine, and Tryptophan amino acids are discovered with all nucleotide bases encoding codons. This study has given directions for some future work, some of which are:

The proposed approach of SPM be applicable for other human viruses' analysis too.

The proposed approach can be extended with analysis of amino acids codons making spike protein of COVID-19.

SPM along with machine learning can be used to train models for predictions of codon families and pairs based on different regions of the same country or different countries of the world.

Contrast pattern mining techniques could be applied to genome sequences of covid-19 for finding out the contrast between two classes.

REFERENCES

[1] He, F., Deng, Y. and Li, W., "Coronavirus disease 2019: What we know?" *Journal of medical virology*, vol. *92, no.* 7, pp.719-725, 2020.

[2] Cucinotta, D. and Vanelli, M., "WHO declares COVID-19 a pandemic," *Acta Bio Medica: Atenei Parmensis*, vol.,91, no. 1, p.157, 2020.

[3] Covid19.who.int. 2021. *WHO Coronavirus (COVID-19) Dashboard*. [online] Available at: <https://covid19.who.int/> [Accessed 13 January 2021].

[4] Raskin, S., "Genetics of COVID-19," *Jornal de pediatria*, vol. *97*, pp.378-386, 2021.

[5] Nawaz, M.S., Fournier-Viger, P., Shojaee, A. and Fujita, H., "Using artificial intelligence techniques for COVID-19 genome analysis," Applied Intelligence, vol. 51, no. 5, pp.3086-3103, 2021.

[6] Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study.

[7] Heidari, A., Navimipour, N.J., Unal, M. and Toumaj, S., "The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions," Computers in biology and medicine, 2021, p.105141.

[8] Fournier-Viger, P., Lin, J.C.W., Kiran, R.U., Koh, Y.S. and Thomas, R., "A survey of sequential pattern mining," *Data Science and Pattern Recognition*, vol. *1, no.* 1, pp.54-77, 2017.

[9] Fournier-Viger, P., Lin, J.C.W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z. and Lam, H.T., "The SPMF open-source data mining library version 2," In *Joint European conference on machine learning and knowledge discovery in databases*, 2016, (pp. 36-40). Springer, Cham.

[10] Nawaz, M.S., Fournier-Viger, P., Niu, X., Wu, Y. and Lin, J.C.W., "COVID-19 Genome Analysis Using Alignment-Free Methods" In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2016, (pp. 316-328). Springer, Cham.

[11] Mabroukeh, N.R. and Ezeife, C.I., "A taxonomy of sequential pattern mining algorithms," *ACM Computing Surveys (CSUR)*, vol. 43, no. 1, pp.1-41, 2010.

[12] Fumarola, F., Lanotte, P.F., Ceci, M. and Malerba, D., "CloFAST: closed sequential pattern mining using sparse and vertical id-lists," *Knowledge and Information Systems*, vol. *48, no.* 2, pp.429-463, 2016.

[13] Cristea, P.D., "Conversion of nucleotides sequences into genomic signals," *Journal of cellular and molecular medicine*, vol. 6, no. 2, pp.279-303, 2002.

[14] Castro-Chavez, F., "Most used codons per amino acid and per genome in the code of man compared to other organisms according to the rotating circular genetic code," *NeuroQuantology: an interdisciplinary journal of neuroscience and quantum physics*, vol. 9, no. 4, 2002.

[15] Biro, J.C., Benyo, B., Sansom, C., Szlavecz, A., Fördös, G., Micsik, T. and Benyo, Z., "A common periodic table of codons and amino acids," *Biochemical and biophysical research communications*, vol. 306, no. 2, pp.408-415, 2006.

[16] Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L.V., Katneni, U., Simonyan, V. and Kimchi-Sarfaty, C., "A new and updated resource for codon usage tables," *BMC bioinformatics*, vol. 18, no. 1, pp.1-10, 2017.

[17] Komar, A.A., "The "periodic table" of the genetic code: A new way to look at the code and the decoding process. *Translation*, vol. 4, no. 2, pp.e1234431, 2016.

[18] Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I. and Schoch, C.L., "The NCBI bioCollections database," *Database*, *2018*.

[19] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L., "GenBank," *Nucleic acids research*, vol. *28, no.* 1, pp.15-18, 2000.

[20] Wang, L., Zhang, Y., Dongguang, W., Tong, X., Liu, T., Zhang, S., Huang, J., Zhang, L., Fan, H. and Clarke, M., "Artificial intelligence for COVID-19: a systematic review" *Frontiers in medicine*, p.1457, 2021.

[21] Chaki, J., & Dey, N., "Pattern analysis of genetics and genomics: a survey of the state-of-art," *Multimedia Tools and Applications*, vol. 79, no. 15, pp. 11163-11194, 2020.

[22] M. R. Nemati, J. Ansary, and N. Nemati, Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical Data, 2020,https://www.sciencedirect.com/science/article/pii/S2666389920300945.

[23] Arslan, H., "Machine learning methods for covid-19 prediction using human genomic data," In *Multidisciplinary Digital Publishing Institute Proceedings* (vol. 74, no. 1, p. 20), 2021.

[24] Salman, F. M., Abu-Naser, S. S., Alajrami, E., Abu-Nasser, B. S., & Alashqar, B. A., "Covid-19 detection using artificial intelligence, 2020.

[25] Mohamadou, Y., Halidou, A., & Kapen, P. T., "A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19," *Applied Intelligence*, vol. 50no. 11, pp. 3913-3925, 2020.

[26] Whata A, Chimedza C. Deep Learning for SARS COV-2 Genome Sequences. *IEEE ACCESS*, vol. 9, pp. 59597-59611. 2021, doi:10.1109/ACCESS.2021.3073728.

[27] Nguyen, D. C., Ding, M., Pathirana, P. N., & Seneviratne, A., "Blockchain and AI-based solutions to combat coronavirus (COVID-19)-like epidemics: A survey," *IEEE ACCESS*, vol. 9, pp. 95730-95753, 2021.

[28] Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., & Cheng, X., "Artificial intelligence and machine learning to fight COVID-19" *Physiological genomics*, vol. 52, no. 4, pp. 200-202, 2021.