Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

ANALYSIS OF FREQUENT NUCLEOTIDE PATTERNS IN COVID-19 GENOME SEQUENCES USING SPM ALGORITHMS

AQSA UMAR

Research Scholar. Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

SANIA BHATTI

Professor. Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

AREEJ FATEMAH MEGHJI

Assistant Professor. Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

NAEEM AHMED MAHOTO

Associate Professor. Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

SAPNA KUMARI

Research Scholar. Department of Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

Abstract

COVID-19 was discovered in Wuhan, China on 19th December 2021. It has been declared a pandemic and has now spread over the globe, impacting millions of people. The genome sequence of COVID-19 strains must be examined to comprehend the behavior and origin of this virus. For this purpose, in this research, we have applied Sequential Pattern Mining (SPM) techniques; fast vertical mining of sequential patterns using co-occurrence information CM-SPAM, vertical mining of maximal sequential patterns (VMSP), closed SPM using sparse and vertical id-lists CloFAST, and efficient mining of top-k sequential patterns (TKS) onto the six strains of COVID-19 genome sequences (CGS) of Pakistan, India, Spain, United Kingdom, China, and Brazil; to investigate genome sequence. First, from the CGS frequent patterns from genome sequence are extracted using CM-SPAM, VMSP, and CloFAST algorithms and after that the frequent extracted patterns are checked whether patterns encode codons of the amino acids. Second, another algorithm TKS is used with a user-defined value k=500 to extract the most frequent patterns from the six genome strains. Third, the obtained results have shown that the availability of frequent pattern as the most of the codons in the six strains. The obtained results are encouraging and show that our study has provided an efficient way for the analysis of CGS as well as given a future direction to CGS analysis.

Index Terms: COVID-19 Genome Sequences, Frequent Patterns, Nucleotide bases, Sequential Pattern Mining, Top-k Sequential Patterns, Vertical Mining of Maximal Sequential Patterns, and Nucleotides

1. INTRODUCTION

The coronavirus disease, caused by the novel Severe Acute Respiratory Syndrome Coronavirus 2 SARS-CoV-2 virus, was first identified in December 2019 in the China [1].

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

Also known as COVID-19 [2], the novel behavior of this disease led the World Health Organization to declare it a pandemic SARS-CoV-2 on March 11th, 2020 [1]. Till today (January 1st, 2023), this disease has infected more than 270 million people with more than 5.3 million deaths worldwide. SARS-CoV-2 is a pandemic beta-coronavirus with encapsulated mono RNA genomes, it has a size of 80 to 160 nm and shape ranging from spherical to pleomorphic. Spike S, Envelope E, Membrane M, and Nucleocapsid N are four structural proteins found in SARS-CoV-2 [3]. However, Spike and Envelope are the essential proteins during virus immune response [4]. A month after it emerged, a unique coronavirus genome was sequenced, and it is known as the CGS, which is made from a single-stranded sequence of nucleotides, and the nucleotide has four nucleotide bases, e.g., Adenine A, Guanine G, Cytosine C, and Thymine T [5]. Since the beginning, this virus has mutated very rapidly. Many COVID-19 genome sequences have been submitted to different health service platforms and still the number is rapidly increasing. COVID-19 Genome Sequencing is important to understand the virus's behavior, derivation, and mutation rate as well as the production of drugs/vaccines and successful preventive techniques. Therefore, the goal of this research is to further study the COVID-19 genome sequence so that we can better understand the behavior and controlling strategies of COVID-19. Various sequential pattern mining (SPM) algorithms are applied on the COVID-19 genome sequences. The analysis provides computerized techniques to bioinformatics to analyze large and complex genomic data.

The first step has to be the transformation of genomic data into a digital form [6]. Next, in order to use SPM to discover hidden knowledge in the genomic sequences and develop effective vaccines [7], [8], not only is a familiarity with the virus genomic sequence required, but also a detailed analysis of codon pair use characteristics [9], the ability to extract features from genome sequence [10], mathematical modeling along with advanced AI [4], [11], and the mechanism to evaluate alignment free methods for genomic sequences data [12]. Biological researchers invest labor, computational expenditures, and, most crucially, time to organize and interpret genomic data to produce precise hypotheses concerning SARS-CoV-2 mutations [13]. However, on sequential genomic data, SPM [14] can provide new information regarding viral mutations, pathogenicity, and clinical symptoms. The goal of this study is to explore the COVID-19 Genome Sequence with the help of SPM. More precisely three sub goals have been addressed:

1. To uncover frequent patterns of nucleotides in COVID-19 Genome Sequences, using SPM algorithms e.g., CM-SPAM algorithm of sequential patterns, VMSP algorithm of maximal sequential patterns, and CloFAST of closed sequential patterns [15]. Moreover, to check similar and dissimilar sequences in the two strains of CGS.
2. To evaluate more about CGS, Top-k's most frequent sequential patterns, the TKS algorithm was applied onto genomic sequences for the output. Moreover, the most frequent patterns encoding codon of 20 amino acids are checked in the six strains of different sub-region of the world.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

3. To analyze the availability of frequent codon of 20 amino acids in the six strains of CGS.

The rest of this paper is arranged in the following manner. Section 2 presents the related work done for the investigation of the COVID-19 genome sequence. Section 3 outlines the workflow of this paper. Section 4 discusses the implementation, section 5 presents the experimental results, followed by the conclusion presented in section 6.

2. RELATED WORK

Using the VOSviewer software, a thematic analysis on COVID-19 tweets-related data is examined to observe public reaction towards pandemic outbreaks by Noor et al. [16]. Frequent words have been extracted using SPM techniques. However, this study may be used to categorize and evaluate the public's ideas and thoughts only on Twitter.

SPM on COVID-19 genome sequences has also been applied by Nawaz et al. [17] and Umar et al. [18]. Predication models are applied onto the sequences to predict nucleotide bases and an algorithm is also designed to calculate mutation change rate by Nawaz et al. [17]. The main objective achieved is to look at artificial intelligence approaches to evaluate the COVID-19 genome. The work done by Umar et al. [18] has focused on the extraction of frequent closed sequences and the performance of the used approach. The limitation of this study is that algorithms for predicting nucleotide base(s) in genomic sequences have low accuracy. Arslan [10] has used four machine learning algorithms for obtaining characteristics from genome sequences. To execute binary classification, the results obtained as a decision tree approach obtained 93% accuracy in the experiments. The dataset used in this study is gene sequences from the coronavirus resource database. In terms of mathematical modeling techniques, Susceptible Infected Recovered (SIR) and the Susceptible Exposed Infected Removed (SEIR) were used and for AI, CT - Images and convolutional neural network (CNN) on x-ray were used to compile datasets of COVID-19. For classification of syndrome coronavirus, an algorithm of deep learning has been proposed by Whata and Chimedza [4], which includes a bi-directional neural network (BNN) and a CNN. Furthermore, it also evaluates whether a genomic sequence possesses potential regulatory motifs. The results showed that CNN-Bi-LSTM models have 99.95 Alignment-Free methods [12]. These methods show similarity and dissimilarity between COVID-19 genome sequences, the obtained results using this approach showed that the alignment-free methods can be efficient for the analysis of the genomic study. However, Randhawa et al. [19] presents an alignment-free technique with machine learning to provide a prediction for real-time genomic sequences but this work implies that alignment-free comparative genomics techniques can be used to supplement alignment-based approaches. To evaluate the evolution of the virus, it is important to track SARS-CoV-2 mutation changes in the region in a time-specific way [13]. Wang and Jiang [14] have performed principal component analysis (PCA) of over 2000 genome sequences. Their study tracks the changes in sequence and also when the COVID-19 genome sequence is larger. This study offers a wonderful chance to research coronavirus mutations and can be used in tree-building methods for the pre-processing

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

step. Machine learning algorithms along with data mining techniques have been used to analyze biological sequences data by Yang et al. [21]. The challenge for the machine learning trained model in the study is that one data set may not be adequately adapted to other data sets due to variations in genomic data. Mathematical modeling and artificial intelligence are powerful tools in the battle against the disease, however, as pointed out by Mohamadou et al. [4] there is still a lot of work to be done in terms of diversifying the databases.

3. METHODOLOGY

In this section, the steps of SPM are discussed to evaluate interesting patterns available in COVID-19 genome sequences.

3.1 Data Collection

Genome sequences are taken from the National Center for Biotechnology Information (NCBI) database for SARS-COV-2 [22]. It maintains an online public database of nucleotides sequences named GenBank sequence databases [23]. It also offers the download of genomic sequences in the form of Nucleotides or Proteins. Table 1 shows the characteristics of COVID-19 genome sequences used in this paper, collected from NCBI in PASTA format containing genes. At the time of writing this paper, there are 2,590,870 COVID-19 nucleotides records and 2,461,904 Sequence Read Archive (SRA) runs available in the NCBI database (SARS-COV-2 Resources NCBI).

Table 1: Covid-19 Genome Sequences Data Taken From NCBJ

ID	Geo Location	Length	Collection Date (YY-MM-DD)	Release Date (YY-MM-DD)
MZ562707	Pakistan	29831	21-04-30	21-07-15
OK189654	India	29871	21-09-10	21-09-21
MZ824622	China	29903	20-02-03	21-08-18
MZ191508	Brazil	29870	20-12-09	21-07-08
MW715069	Spain	29875	21-02-06	21-03-09
OU892945	UK	29876	21-10-02	21-10-28

3.2 Data Transformation

The COVID-19 genome sequences collected in this paper are in the form of Nucleotides, there are four kinds of Nucleotides found in a DNA i.e: N = A, C, G, T; these are called Nucleotide bases [4]. Table 1 shows COVID-19 genome sequence characteristics, in which ID is genome sequence accession number. For applying SPM on CGS abstraction of nucleotides to integers is employed. All nucleotides are replaced by a positive integer i.e: N = 1,2,3,4. A is replaced with 1, C is replaced with 2, G is replaced with 3, and T is replaced 4.

3.3 SPM Algorithms

SPM is the most important data mining task to extract hidden knowledge [14]. In the set of sequences, interesting subsequences are discovered using SPM.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

Algorithm for Mining Sequential Patterns: CM-SPAM and CM-SPADE have the best performance among all algorithms for mining sequential patterns [14]. For the results of our paper, we choose the CM-SPAM algorithm of the sequential pattern because CM-SPAM performs better for bioinformatics data, CM-SPAM is also very efficient and the most popular algorithm to discover sequential patterns for long databases [24].

Algorithms for Mining Maximal Sequential Patterns: VMSP is an efficient Vertical Mining of Maximal Sequential Patterns algorithm, it is based on SPAM search, but fewer Frequent Patterns are extracted than spam algorithms [25]. MaxSP and VMSP are the best algorithms for a maximal sequential pattern. However, according to Fournier-Viger et al. [26], VMSP is 100 times faster than MaxSP, thus, in this study, we have chosen the VMSP algorithm.

Algorithms for Mining Closed Sequential Patterns: CloFAST is a pattern mining algorithm for closed sequential patterns. When densities of datasets increase then CloFAST outperforms other closed sequential patterns algorithms i-e BIDE, CloSpan, and ClaSP [27]. For the most complex task of mining from long sequences, its behavior is also evident. CloFAST mines closed frequent patterns and enumerate the search space with sparse id-lists, while it generates closed sequence patterns with vertical id-lists.

Algorithms for Mining the Top-K Sequential Patterns: With wide applications of top-k sequential pattern algorithms, TKS and TSP are Top-K Sequential are well-studied algorithms. However, extensive study in has shown that TKS outperforms TSP [28]. With the help of the TSK algorithm user can choose the desired number of frequent patterns from very big datasets without the consumption of more time.

3.4 Frequent Sequential Patterns

All the SPM Algorithms output the set of frequent sequential patterns. Hence, each algorithm sequential finds sequential patterns. It depends upon the algorithm strategy to search efficient patterns, as some algorithms are more efficient than others and most of the algorithms return the same set of sequences [14].

3.5 Implementation

To implement SPM algorithms, Sequential Pattern Mining Framework (SPMF) data mining library is used [27]. The SPMF is developed in java, which is open-source and cross-platform. It has about 180 data mining algorithms implemented in it. All the experiments were performed on released version v2.49 of SPM, which provides a Graphical User Interface GUI and a Command Line Interface CLI.

Applying Algorithms: four algorithms of SPM are shown: CM-SPAM, VMSP, CloFAST and TKS. CM-SPAM, VMSP, and CloFAST are then applied on the strain of Pakistan and India. However, TKS is applied to six strains which can be found at GitHub [29]. Three algorithms, CM-SPAM, VMSP, and CloFAST require a minimum support threshold. Whereas the TKS algorithm requires user input parameter (K) value. To apply SPM algorithms on the SPMF framework, FASTA format sequence; which is a DNA-based

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

sequence, is converted into SPMF format [14]. In the Data transformation step of Fig 1. b), FASTA database shows nucleotides sequences which are then converted into a sequence database i-e., nucleotides into integers. Moreover, the separator characters -1, -2 are added in SPMF format where the value of -1 indicates the end of an item-set and -2 the end of a sequence, as shown in Fig 1.c). After applying algorithms, a lot of frequent nucleotide bases were discovered.

Data visualization: Sequences base(s) that were not multiple of three nucleotides (codon) were filtered out. More- over, the codons encoding the Amino Acids are further visualized.

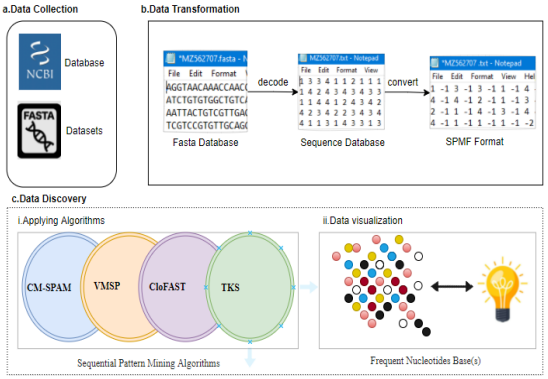


Figure 1: The Proposed Approach for Analyzing COVID-19 Genome Sequence

4. EXPERIMENTAL RESULTS

The results of the frequent nucleotides extracted by applying SPM algorithms on COVID-19 genome sequences have been presented in this section; the genome sequences can be found at GitHub [29]. Table 2. shows COVID-19 genome sequences nucleotides percentages in all strains used in this research, the number of A's and G's nucleotides in the strain of India is the same as the strain of China, and the number of T's nucleotides in the strain of India is the same as in the strain of Spain. The strain of India has three more C nucleotide than as in the strain of Pakistan. From all six strains, Thymine T from Brazil has maximum number of nucleotides percentage (32.58%) from all other strains whereas Cytosine C from Brazil has minimum number of nucleotides percentage (18.20%) from all other strains, as shown in Table 2.

Table 2: Covid-19 Genome Sequences Nucleotides Percentage

ID	A (%)	B (%)	C (%)	D (%)
Pakistan	8915(29.89)	5474(18.35)	5852(19.62)	9590(32.15)
India	8924(29.88)	5477 (18.34)	5862(19.62)	9608(32.16)
Spain	8954(29.94)	5493(18.37)	5862(19.60)	9594(32.08)
UK	8924(29.88)	5489(18.38)	5856(19.60)	9601(32.14)
China	8923(29.87)	5481(18.35)	5863(19.63)	9608(32.16)
Brazil	8891(29.76)	5438(18.20)	5812(19.45)	9735(32.58)

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

4.1 Frequent Nucleotide Patterns Analysis

From the COVID-19 genome sequence, frequent nucleotides patterns are extracted to learn more about the genome sequence of COVID-19. SPM algorithms are applied to sequence databases of genome sequences to extract interesting patterns. In sub-sections (4.1.1, 4.1.2 and 4.1.3) of 4.1, Pakistan and India strains are taken for extraction of the pattern. Similar Tables (Table 3, 5, and 7) indicate that the sequences are present in the strains of Pakistan and India. However, the dissimilar tables (Table 4, 6, and 8) indicate that the sequences are present in one country Pakistan or India. It can also be observed from Table 1 that the strain of Pakistan has forty fewer nucleotides than the strain of India.

Frequent nucleotides extracted by CM-SPAM: CM- SPAM requires setting the minSup threshold to execute the frequent patterns [30]. Tables 3 and 4 show the patterns extracted by CM-SPAM with varying minSup thresholds; Table 3 shows similar sequences in the strains of Pakistan and India extracted by CM-SPAM. In Table 3, the frequent AAAAAAAAAGAA and AAAAAAAAAGCC with supports 405 and 383; encode the first three codon AAA of the Lysine amino acid, with one GAA codon of Glutamic Amino Acid and one GCC codon of Alanine Amino Acid. Furthermore, there are three codons AGT of Serine Amino Acid and three codons TGT condon of Cysteine Amino Acid with the supports 450. Most of the data in Table 3 has pattern ATA, as the first codon encoding the Isoleucine Amino Acid. Except for the pattern ACACAA, which encodes a codon ACA of Threonine Amino Acid and a codon CAA of Glutamine Amino Acid. Table 4, shows dissimilar patterns in Pakistan and India extracted by CM-SPAM. Looking at the table, there are patterns TTCGGA and TTTGGC in Pakistan, and India with support 475. However, the first patterns ITC and TIT encode Phenylalanine Amino Acids, and other patterns CCA and CCC encode two codons of Glycine Amino Acid. In Pakistan the pattern AAAAAAAAAGGT and in India the pattern AAAAAAAAACTA encodes three AAA codons of Lysine Amino Acid with a codon CCT of Glycine Amino Acid and a codon CTA of Valine Amino Acid.

Frequent nucleotides extracted by VMSP: VMSP requires setting a minSup threshold to execute the frequent patterns [25]. Tables 5 and 6 show the patterns extracted by VMSP with varying minSup thresholds. Most of the frequent patterns using VMSP came out from length seven with most of the G and T nucleotides. In Table 5, there are the patterns CTCTCTTTT, TCTTTCTTC, CCT11111 C with support of 462 and five patterns CCCTTATTT, TATTTTCCC, TTCTTCCTA, TTCTCGTTA, TTTTCCTCT with support of 457. Moreover, the pattern TTGTTGGTA encodes all three codons of Leucine Amino Acid. In table 5 there's also a unique codon TGG Tryptophan Amino Acid with two TCCT and TGT codons of Cysteine Amino Acid. In Table 6, the pattern CCCCCC in Pakistan with support of 462 encodes two codons of Proline Amino Acid. The pattern TATCTCTAT in India with support 468 has two codons of Tyrosine Amino Acid with one codon CTC of Valine Amino Acid.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

Table 3: Frequent Nucleotides Similar In Pakistan and India Extracted By Cm-Spam

Patterns	Min. Sup	Sup. In Pakistan	Patterns	Min. Sup	Sup. In India
AACAACAAC	20%	443	ATAGTG	30%	494
CCTCCTCCT	20%	403	ATACCA	30%	496
AGTAGTAGT	20%	450	ATACCT	30%	496
AGGAGGAGG	20%	350	ATACGA	30%	495
TGTTGTTGTT	20%	450	ATACGG	30%	492
AAAGAGGAG	25%	431	ATACGT	30%	495
AAAGAACAAG	25%	415	ACACAA	30%	494
AAAAAAAATTAT	25%	473	ATAAGT	30%	498
AAAAAAAAAGAA	25%	405	ATAATA	30%	498
AAAAAAAAAGCC	25%	383	ATAATC	30%	496

Table 4: Frequent Nucleotides Dissimilar In Pakistan and India Extracted By Cm-Spam

Patterns	Min. Sup	Sup. in Pak	Patterns	Min. Sup	Sup. in India
AGTGTAGTA	20%	375	AAAGAAACA	20%	461
TTGAGGAAT	20%	471	TTGAGGACA	20%	462
AGGACT	30%	459	TTCGGA	30%	475
AAAAAAAAAGGT	25%	385	TTTGGC	30%	475
AAAAAAACAACA	25%	401	AAAAAAAAAATT	25%	374
TTTTTTTTTTTT	25%	421	AAAAAAAAAGTA	25%	406
AAAGGTCCC	30%	474	-	-	-

Table 5: Frequent Nucleotides Similar In Pakistan and India Extracted By VMSP

Patterns	Min. Sup	Sup. In Pakistan	Patterns	Min. Sup	Sup.in India
GTATTTTTG	20%	466	TTGTTGGTA	25%	457
TTGTGTTTG	20%	467	GGTTGTTAT	25%	462
GTGTGTTTT	20%	462	CTTTTAGTT	25%	468
TTTTGGTGT	20%	457	TGTTTGTTG	30%	462
TTTTGTTTG	20%	472	GTTTCTTTT	30%	460
GTGTTTTTG	25%	466	TTTTGGGTT	30%	461
TGCTGTTGG	25%	472	GGTTTTTTG	30%	462
GGGTTATTT	25%	457	TTGAGTTTG	30%	471
GGTTTTGTT	25%	465	TTATTCGTT	30%	460
TATTTTGGG	25%	457	TTGTGGTTA	30%	457

Table 6: Frequent Nucleotides Dissimilar In Pakistan and India Extracted By VMSP

Patterns	Min. Sup	Sup. in Pakistan	Patterns	Min. Sup	Sup. In India
CCGCGG	20%	462	TATGTGTAT	20%	468
CCCGGG	20%	458	GTTCTTTTT	20%	379
TTTTGGTTT	25%	410	TTTGTATTT	25%	345
TTGTTGTTT	25%	386	TGTTGGCTT	25%	369
GTTTGTTTG	25%	398	TTTGGGTTT	25%	380
TTTTTGGTT	25%	410	GTTTTTTGG	30%	499
TTTTTCGTT	30%	499	TCTTTGTTT	30%	497
TTTTTGGGT	30%	496	GTGGTTTTT	30%	493

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

Frequent nucleotides extracted by CloFAST: With varying minSup using CloFAST, the patterns of Pakistan and India are shown in Table 7 and Table 8. Most of the patterns in Tables 7 and 8 are of the nucleotides A and C and with maximum support of 1818 and minimum support of 1248. In Table 7, patterns ACT and ATA with minSup 20% encode one codon of Threonine Amino Acid. There are also nine patterns of AAA in Table 7, encoding a codon AAA of Lysine amino acid. Two patterns AAAATA and AAAATC with supports 1818 and 1678 encodes two codons of Lysine Amino Acids with two codons ATA and ATC of Isoleucine Amino Acid. Moreover, a unique codon ATC of Methionine Amino Acid is encoded from pattern AAAATC is discovered by CloFAST. In Table 7, there are four patterns of CTC with CCT, CCC, CCA, and CCC in Pakistan encoding four codons of the Leucine Amino Acid and four codons of Glycine Amino Acid and CTC with CCT, CCC, CCA, and CCC. In Table 7, there're three patterns of CAC with CCC, CCC, and CCT in India. Encoding three codons of Histidine Amino Acid and two codons of Proline Amino Acid.

Table 7: Frequent Nucleotides Similar In Pakistan and India Extracted By Clofast

Patterns	Min. Sup	Support	Patterns	Min. Sup	Support
ACT	20%	1355	AAAAAG	25%	1598
ATA	20%	1426	AAAACG	25%	1536
ATT	20%	1440	AAAACT	25%	1738
CAA	20%	1344	AAAAGA	25%	1725
CAC	20%	1252	AAAAGC	25%	1596
CAG	20%	1248	AACCAG	30%	1494
AAAACC	20%	1586	AACAGT	30%	1613
AAAATA	20%	1818	AACCCT	30%	1488
AAAATC	20%	1678	AACCGA	30%	1443
AAAATG	20%	1770	AACCGC	30%	1353

Table 8: Frequent Nucleotides Dissimilar In Pakistan and India Extracted By Clofast

Patterns	Min. Sup	Sup. in Pakistan	Patterns	Min. Sup	Sup. in India
ATTTTTAT	25%	1394	AGGCGG	20%	1316
ATTTTTTT	25%	1400	ATCGCG	20%	1309
CTACGG	30%	1386	CACCCC	20%	1302
CTCGCA	30%	1387	CACCGC	20%	1304
CTCGGA	30%	1386	CACCCT	20%	1303
CTCGGT	30%	1400	AATTTTTT	25%	1396
CTCGCA	30%	1392	ATTTTAAA	25%	1374
CTCGGA	30%	1382	ATTTTTAA	25%	1394

4.2 Analyzing of CCS with the Top-k SPM Algorithm

In this section, the TKS algorithm is applied onto the strains of Pakistan, India, Spain, UK China, and Brazil. The TKS algorithm is used for Top-k most frequent sequential patterns as output. In the TKS algorithm, k is a user-specified parameter that is used instead of the minimum support threshold [28].

From the literature [31-33]:

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

1. A set of three nucleotides encode one amino acid; i-e., the codons GGC, GGA and GGG encode the amino acid known as Glycine.
2. There are 4^3 different i-e., 64 codons, from which 61 possible codons make up 20 amino acids, and the remaining 3 codons encode stopping codons.
3. Expect Tryptophan and Methionine; most of the amino acids have more than one codon r327 r34l.

Table 9, shows extracted patterns by the TKS algorithm with k=500 from the six strains of COVID-19. The Table 9, shows the frequent patterns as the codon or codons of amino acid, indicate the availability of codons as a frequent pattern in the strains of Pakistan, India, Spain, UK, China, and Brazil.

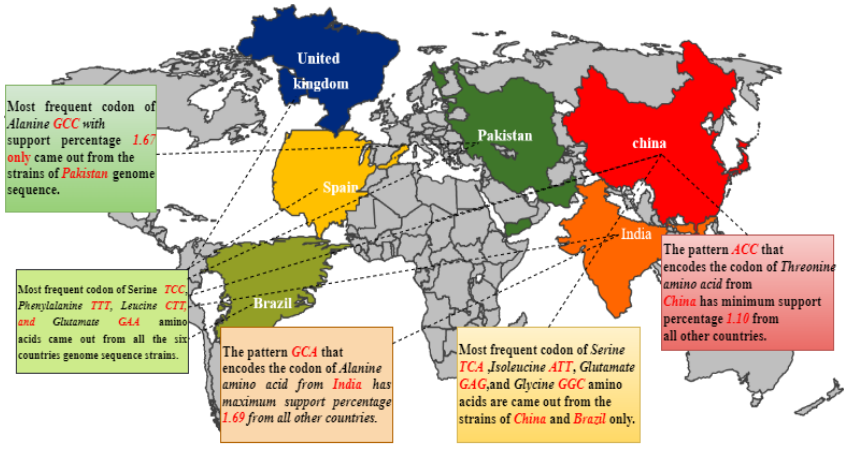
Table 9: Frequent Codons Availability Extracted By the TKS Algorithm

Patterns (Codon)	Full name of Amino Acid	PAKISTAN	India	Spain	UK	China	Brazil
TCT	Serine	+		+	+	+	+
TCC		+		+	+	+	+
TCA						+	+
AGT			+	+	+	+	+
AGC		+	+		+	+	+
AGT			+	+		+	+
TTT	phenylalanine	+	+	+	+	+	+
TTC		+	+	+	+		+
TAT	Tyrosine		+	+	+	+	+
TAC		+	+	+	+		
TGT	Cysteine		+		+	+	+
TGC			+		+	+	+
TGG	Tryptophan		+	+	+		+
TTG	Leucine			+	+	+	+
CTT		+	+	+	+	+	+
CTC				+	+	+	+
CTA				+	+	+	+
CTG			+		+	+	+
TTA		+	+	+	+		+
CCT	Proline		+	+	+	+	+
CCC			+	+		+	+
CCA		+	+	+		+	+
CCG						+	+
CAT	Histidine		+	+	+	+	+
CAC						+	+
CAA	Glutamine		+	+	+	+	+
CAG			+			+	+
CGT	Arginine	+			+	+	+
CGC							
CGA			+	+	+	+	+
CGG							
AGA				+	+	+	+
AGG					+	+	+
ATT	Isoleucine					+	+
ATC				+	+	+	+
ATA				+	+	+	+

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

ATG	Methionine			+	+	+	+
ACT	Threonine			+	+	+	+
ACC				+	+	+	+
ACA				+	+	+	+
ACG		+	+		+		
AAT	Asparagine	+		+	+	+	+
AAC					+	+	+
AAA	Lysine			+	+	+	+
AAG				+	+	+	+
GTT	Valine	+		+	+	+	+
GTC		+		+	+	+	+
GTA				+	+	+	+
GTG		+		+	+	+	+
GCT	Alanine	+			+	+	+
GCC		+					
GCA		+	+		+	+	+
GCG			+			+	+
GAT	Aspartate	+		+	+	+	+
GAC		+	+			+	+
GAA	Glutamic acid	+	+	+	+	+	+
GAG						+	+
CAA	Glutamine	+	+			+	+
CAG			+			+	+
GGT	Glycine	+			+	+	+
GGC						+	+
GGA		+			+	+	+
GGG							+

Fig 2, shows further analysis of Table 9. It is observed that the maximum support percentage came out from pattern of India, and minimum support percentage came out from pattern of China. Most of the frequent codons came out from the strains of China and Brazil and the most frequent codon TCC, TTT, CTT and CAA found to be available in the all six strains. Moreover, the pattern CCS only came from strain of Pakistan.



Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

5. CONCLUSION AND FUTURE WORK

The research done in this paper is based on the genome sequences for COVID-19 strains of Pakistan, India, Spain, United Kingdom, China, and Brazil taken from NCBI's GenBank. The genome sequences have been analyzed with four SPM algorithms: CM-SPAM, VMSP, CloFAST, and TKS. First, Frequent Nucleotides are extracted using pattern mining algorithms and then, the TKS algorithm is used for further analyzing the COVID-19 genome sequence. Top-k sequential patterns are extracted from six strains of COVID-19 genomic sequences and then the patterns encoding codons of amino acids are checked in each strain. The obtained results suggest that most of the codons of amino acids came from Spain, UK, China and Brazil. The pattern GCA that encodes the codon of Alanine amino acid from India has maximum support percentage 1.69 and the pattern ACC that encodes the codon of Threonine amino acid from China has minimum support percentage 1.10 from all other countries. These proposed approaches must lead to future work directions, some of which are as follow:

1. To analyze strains of COS for protein sequences using SPM techniques as done by Cascella et al. [2]. The strains of COVID-19 genome sequence in the form of protein are available (Castro-Chavez 2011).
2. To apply contrast set mining techniques [34] on to COVID-19 genome sequence to discover similarities and dissimilarities amino acids or proteins in the strains [35].
3. To evaluate mutation change rate by applying the Mutation analysis technique to different strains of the same city of the same country.
4. Tools like CAFE [36] and AF [12] can be used for analyzing COVID-19 genome sequence. Furthermore, a study to compare the tools can also be conducted.

References

1) D. Cucinotta and M. Vanelli, “WHO declares COVID-19 a pandemic,” Acta Bio Medica: Atenei Parmensis, vol. 91, no. 1, pp. 157, 2020.

2) M. Cascella, M. Rajnik, A. Aleem, S.C. Dulebohn, and R. Di Napoli, “Features, evaluation, and treatment of coronavirus (COVID-19),” Statpearls, https://www.ncbi.nlm.nih.gov/books/NBK554776/. 2022.

3) M. Pal, G. Berhanu, C. Desalegn, and V. Kandi, “Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-): an update,” Cureus, vol. 12, no. 3, 2020.

4) A. Whata and C. Chimedza, “Deep learning for sars cov-2 genome sequences,” IEEE Access, vol. 9, pp. 59597-59611, 2021, doi: 10.1109/ACCESS.2021.3073728.

5) Y. Mohamadou, A. Halidou, and P.T. Kapen, “A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19,” Applied Intelligence, vol. 50, no. 11, pp. 3913-3925, 2020.

6) J. Chaki and N. Dey, “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.

ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

7) M. Naeem, H.F. Alkhodairy, I. Ashraf, and A.B. Khalil, “CRISPR/Cas System toward the Development of Next-Generation Recombinant Vaccines: Current Scenario and Future Prospects,” Arabian Journal for Science and Engineering, pp. 1-11, 2022, doi: 10.1007/s13369-022-07266-7.

8) A.S. Yadav, P. BM, N. Ahlawat, and M. Abid, “Mathematical Modeling of Corona Virus Vaccine Supply Chain Inventory System with Retailer using Machine Learning and Bacterial Foraging Optimization,” Xi'an Shiyou Daxue Xuebao, vol. 65, no. 7, pp. 22-45, 2022.

9) J. Kames, D.D. Holcomb, O. Kimchi, M. DiCuccio, N. Hamasaki-Katagiri, T. Wang, A.A. Komar, A. Alexaki, and C. Kimchi-Sarfaty, “Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design,” Scientific reports, vol. 10, no. 1, pp.1, 2020.

10) H. Arslan, “Machine learning methods for covid-19 prediction using human genomic data,” Multidisciplinary digital publishing institute proceedings, vol. 74, no. 1, pp. 20, 2021.

11) R. Vaishya, M. Javaid, I.H. Khan, and A. Haleem, “Artificial Intelligence (AI) applications for COVID-19 pandemic,” Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 4, pp. 337-339, 2020.

12) M.S. Nawaz, P. Fournier-Viger, X. Niu, Y. Wu, and J.C. Lin, “COVID-19 Genome Analysis Using Alignment-Free Methods,” International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 316-328, and 2021.

13) A.T. Chen, K. Altschuler, S.H. Zhan, Y.A. Chan, and B.E. Deverman, “COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest,” Elife, vol. 10, pp. e63409, 2021.

14) P. Fournier-Viger, J.C. Lin, R.U. Kiran, Y.S. Koh, and R. Thomas, “A survey of sequential pattern mining,” Data Science and Pattern Recognition, vol. 1, no. 1, pp. 54-77, 2017.

15) A. Borah and B. Nath, “Comparative evaluation of pattern mining techniques: an empirical study,” Complex & Intelligent Systems, vol. 7, no. 2, pp. 589-619, 2021.

16) S. Noor, Y. Guo, S.H. Shah, P. Fournier-Viger, and M.S. Nawaz, “Analysis of public reactions to the novel Coronavirus (COVID-19) outbreak on Twitter,” Kybernetes, 2020, doi:10.1108/K-05-2020-0258.

17) M.S. Nawaz MS, P. Fournier-Viger, A. Shojaee, H. Fujita, “Using artificial intelligence techniques for COVID-19 genome analysis,” Applied Intelligence, vol. 51, no. 5, pp. 3086-4103, 2021.

18) A. Umar, N.A. Mahoto, S. Bhatti, and S. Rathi, “Analysis of Covid-19 Genome Sequences based on Geo-Locations,” Pakistan Journal of Engineering and Technology, vol. 4, no. 4, pp. 41-45, 2021.

19) G.S. Randhawa, M.P. Soltysiak, H. El Roz, C.P. de Souza, K.A. Hill, and L. Kari, “Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study,” Plos one, vol. 15, no.4, pp. e0232391, 2020.

20) B. Wang and L. Jiang, “Principal Component Analysis Applications in COVID-19 Genome Sequence Studies,” Cognitive computation, vol. 13, pp. 1-2, 2021.

21) A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han, and L. Zhang, “Review on the application of machine learning algorithms in the sequence data mining of DNA,” Frontiers in Bioengineering and Biotechnology, vol. 8, pp.1032, 2020.

22) S. Sharma, S. Ciufo, E. Starchenko, D. Darji, L. Chlumsky, I. Karsch-Mizrachi, and C.L. Schoch, “The NCBI biocollections database,” Database. 2018, doi:10.1093/database/baz057.

23) D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and E.W. Sayers, “GenBank,” Nucleic acids research, vol. 41, no. D1, pp. D36-42, 2012.

ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 66 Issue 02 | 2023
DOI 10.17605/OSF.IO/D7GA3

24) J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, “Sequential pattern mining using a bitmap representation,” In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 429-435, 2002.

25) P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, “Fast vertical mining of sequential patterns using co-occurrence information,” Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 40-52, 2014.

26) P. Fournier-Viger, C.W. Wu, A. Gomariz, and V.S. Tseng, “VMSP: Efficient vertical mining of maximal sequential patterns,” In Canadian conference on artificial intelligence, pp. 83-94, 2014.

27) F. Fumarola, P.F. Lanotte, M. Ceci, and D. Malerba, “CloFAST: closed sequential pattern mining using sparse and vertical id-lists,” Knowledge and Information Systems, vol. 48, no. 2, pp. 429-63, 2016.

28) P. Fournier-Viger, A. Gomariz, T. Gueniche, E. Mwamikazi, and R. Thomas, “TKS: efficient mining of top-k sequential patterns,” In International Conference on Advanced Data Mining and Applications, pp. 109-120, 2013.

29) Dataset: https://github.com/Aqsa48/COVID-19-Strains. 2022.

30) P. Fournier-Viger, J.C. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, and H.T. Lam, “The SPMF open-source data mining library version 2,” In Joint European conference on machine learning and knowledge discovery in databases, pp. 36-40. 2016.

31) P.D. Cristea, “Conversion of nucleotides sequences into genomic signals,” Journal of cellular and molecular medicine, vol.6, no. 2, pp. 279-303, 2002.

32) J.J. Shu, “A new integrated symmetrical table for genetic codes,” Biosystems, vol. 151, pp. 21-26, 2017.

33) J. Athey, A. Alexaki, E. Osipova, A. Rostovtsev, L.V. Santana-Quintero, U. Katneni, V. Simonyan, and C. Kimchi-Sarfaty, “A new and updated resource for codon usage tables,” BMC bioinformatics, vol. 18, no. 1, pp.391, 2017.

34) S. Ventura and J.M. Luna, Supervised Descriptive Pattern Mining. Cham: Springer International Publishing, 2018.

35) F. Castro-Chavez, “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, 2011.

36) Y.Y. Lu, K. Tang, J. Ren, J.A. Fuhrman, M.S. Waterman, and F. Sun, “CAFE: a C celerated A lignment-F r E e sequence analysis,” Nucleic acids research, vol. 45, no. W1, pp. W554-559, 2017.