Shivendra DUBEY, PhD Scholar (corresponding author)
shivendrashivay@gmail.com
Jaypee University of Engineering & Technology, India

Dinesh Kumar VERMA, PhD
dinesh.hpp@gmail.com
Jaypee University of Engineering & Technology, India

Mahesh KUMAR, PhD
mahesh.chahar@gmail.com
Jaypee University of Engineering & Technology, India

Identification of Unique Genomic Signatures in Viral Immunogenic Syndrome (VIS) Using FIMAR and FCSM Methods for Development of Effective Diagnostic and Therapeutic Strategies

Abstract. *The "Viral Immunogenic Syndrome" (VIS) incorporates the concepts of "viral" and "immunogenic" to emphasise the pathogenic character of the illness and the immunological response it generates, as well as the word "syndrome" to describe the broad set of symptoms and consequences. Our research focused on analyzing the COVID-19 genome sequence using a proposed framework to improve computation time and model efficiency. We also aimed to identify frequent patterns, missing indices, and variations in the genome sequence while comparing performance with varying minimum support using existing and proposed models. We used FCSM to classify COVID-19 genomic sequences and detect variations while calculating computation time. Additionally, we proposed a novel framework utilizing FIMAR to identify nucleotide sequences and compute consecutive sets, resulting in a more efficient and accurate approach to analyzing COVID-19 genome than existing methods. Our study shows that the proposed framework with FIMAR and FCSM algorithms is 94.34% more efficient than the existing system for computing the COVID-19 genome sequence. We also identified 0.2% and 1.61% genomic variation in the USA and China datasets, respectively, which the existing system failed to detect. Additionally, we conducted a comparative study using an Apriori and FIMAR methods to identify frequent patterns. In this work, we also present an analysis of the nucleotide substitution rate at each isolation step.*

Keywords*:* *Adenine, Guanine, Genome, SARS-COV-2, Nucleotides.*

JEL Classification: C80, C81, C82, C88, I18.

1. Introduction

SARS-CoV-2 is the cause of the COVID-19 epidemic; it still harms people's health and wellness despite infecting countless people worldwide. The new coronavirus infections, also known as COVID-19 or SARS-CoV-2, had spread

DOI: 10.24818/18423264/58.2.24.10

rapidly throughout the world, posing a threat to public health. The scientific reaction seen worldwide to the lethal virus that killed thousands of people worldwide is unparalleled. To prevent the spread of the fatal virus and its ability to breed, governments and health organisations adopt and advise quarantine and preventive measures. To create a vaccine or medicine that may effectively treat or prevent COVID-19, for that purpose scientists study the virus's DNA, functions, and behaviour (Dubey et al., 2024).

The initial genomic sequence for COVID-19 was made available on January 10, 2020, only nine days after the first probable case was detected. Since then, scientists all around the globe have sequenced tens of thousands of samples (Chaudhary et al., 2021; Makaremi et al., 2022). One of the most exciting areas of study for COVID-19 is genomic research, which examines an organism's genes and their roles. By analysing an organism's DNA, doctors can make better, individualised diagnose and treatments. Experts may believe that by deciphering the virus's genetic code and determining which people are most seriously affected, they would be able to more effectively influence public health policy, make well-informed decisions, and create clinically feasible treatments. Four distinct nucleic acid bases' adenine (A), guanine (G), cytosine (C), and thymine (T), are kept in an encoded sequence in an organism's genome. The unique genome sequence of COVID-19 is 30 kilobases (kb) long, positive-sense ribonucleic acid (RNA), and single-stranded from the large coronavirus family. Only two families of corona viruses are known to cause infections in mammals (Mousavizadeh et al., 2021). To properly analyse a genomic sequence, it is necessary to identify each nucleotide independently (Li et al., 2022; Mercatelliet al., 2021). Researchers and medical practitioners may gain insight into the genetic variants of SARS-CoV-2 or COVID-19 through cutting-edge artificial intelligence techniques and genomic sequence analysis methods.

Rest of manuscript: Section 2 presents the genome and genome sequence. Also, in Section 3, we present a variety of previously completed relevant studies and compare several methodologies, including the COVID-19 pandemic, with ML-based analysis. We explain the FIMAR based genomic sequence analysis in Section 4. Results and analysis of the system evaluation are presented in Section 5. Section 6 concludes the study and some guidelines for further research.

1.1 Genome Sequencing

There are available nucleotides (A, C, G, and T) for a DNA genome, but if it is available as a whole list of nucleotides called a genome sequence, those make up every one of the chromosomes of a species or an individual. Genome sequencing identifies the nucleotide sequence in a genome.

Genome sequencing improves healthcare by identifying, classifying, and interpreting disease-linked genetic variations. This method integrates phenotypic information and other multi-omics datasets with population-level sequencing and genomes data (Chaudhary et al., 2021). AI and allied technologies such as data analysis, machine, and deep learning (Ahmed et al., 2022; Li et al., 2022; Mishra et al., 2021) may speed up global responses. Growing need for computational tools to analyse complex, high-dimensional genetic data sets. AI may drive discoveries in genetic data sets without clear criteria and interests. Researchers have focused on analysis and perception when dealing with genetic data sets. Clinical genome analysis, mainly phenol typing in uncommon disorders and cancer, has benefitted from machine learning techniques (Bagabir et al., 2022; Arslan et al., 2021).

1.1.1 Nucleotide Sequences

The most basic understanding of a genome or gene is its nucleotide sequence. Without having it, it would be impossible to comprehend how genes function or how organisms evolve, because it is the blueprint that holds the instructions for developing an organism (Lopez-Rincon et al., 2021).

1.1.2 Gene

The basic functional and physical heredity unit is called the gene, represented in Fig. 1. DNA helps to make genes. So many genes work like an instruction that helps make molecules known as proteins.

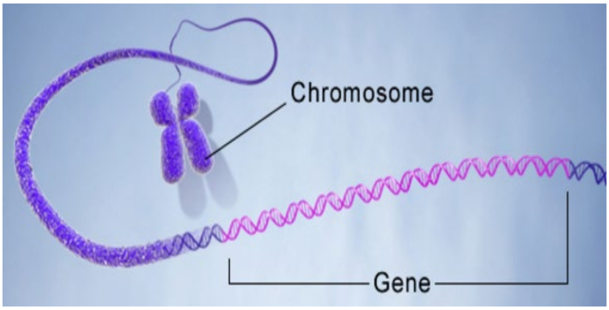

Figure 1. Representation of Gene
Source: U.S. National Library of Medicine.

1.1.3 DNA

Deoxyribonucleic Acid or DNA (Fig. 2) is the genetic substance in human beings and approximately all around the world organisms. Almost every cell in a human has the same DNA.

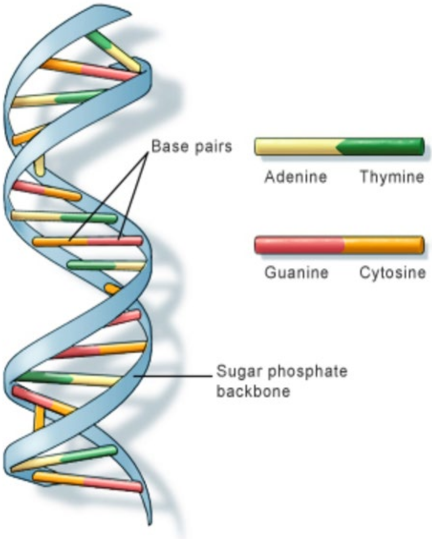

Figure 2. DNA's structure
Source: U.S. National Library of Medicine.

1.1.4 Cell

Cells are the basic building blocks of all living things. The human body contains several billion cells. They give the body structure, take in food's nutrients, convert them into energy, and carry out certain functions. Cells can duplicate the genetic material of the body.

2. Related Work

Forecasting and risk analysis of numerous illnesses, were among the first use of technologies like statistics, data mining, and machine learning. The machine learning-based methodologies examined to analyse and forecast the pandemic (Shujaet al., 2022). A list of the many healthcare therapies compiled machine learning-based methodologies and big data technologies (Khanraet al., 2020). Researchers built up a predictive system for evaluating the likelihood of an individual's early death using a machine learning approach. A machine learning technique has been used to calculate the probability that an infected person will spread COVID-19 or have ARDS using clinical database (acute respiratory distress syndrome). Examining the data from mobile devices created a technique to identify COVID-19-affected individuals (Barnett et al., 2023).

A comprehensive overview compiled of machine learning algorithms which are used in crisis and pandemic response (Chamola et al., 2021). The COVID-19 pandemic was analysed (Li et al., 2021) using dynamic modelling. Another one created a regression model based on the number of cases reported from nations other than China to predict that COVID-19 spread rapidly (Fan et al., 2020). A framework demonstrated for health tracking uses a textual clinical data set to analyse the COVID-19 pandemic using IoT and big data analytics (Himeur, Y. et al., 2023).

Taxonomy (Hassan et al., 2021) provided existing techniques into four broad categories. Then, they informed the experts in machine learning of the difficulties they faced and suggested ways to improve the methods already in use to foresee situations of COVID-19. Suggested employing deep and machine learning technologies to distinguish exponential reactions and estimate the projected spread of COVID-19 throughout the countries. The machine learning technique has been used to estimate the number of COVID-19 patients who contract the disease (Alali et al., 2022). Monitoring social distance and stopping the spread of COVID-19 was made possible by using deep learning models trained on a surveillance data set (Himeur et al., 2023). However, other investigators have also analysed the COVID-19 genome using artificial intelligence-based methods, such as sequential pattern mining, which involves analysing large datasets of genomic sequences (Mercatelli et al., 2021; Chaudhary et al., 2021). The deep learning method (Lopez-Rincon et al., 2021) can distinguish between RNA and SARS-CoV-2 viruses that often co-circulate. The CNN model (Rohaim et al., 2020) proposed, which is successfully identified and classified SARS-CoV-2. There are numerous Artificial Intelligence-based monitoring techniques for COVID-19 was investigated (Lopez-Rinconet al., 2021; Madhav et al., 2022).

Table 1. Summary of COVID-19 studies with various mathematical models

S No.	Model	COVID 19 learning	References
1	MCMC	Evaluations of self-defence effectiveness	(Hernandez-Matamoros et. al., 2020)
2	SEIR	Management studies, dynamics, forecasting, the impact of humidity, and the impact of temperature	(Jalaber et. al., 2020)
3	ODE meta population	Economics and COVID-19	(Madhav et. al. 2022)
4	SIQR	Strategies for Management and Quarantine	(Lopez-Rincon et al., 2021)
5	SPSS Modeller	Temperature level effects	[52]
6	SIR	Track transfer, data fitting, and rate recovery where necessary	(Lopez-Rincon et al., 2021)
7	SESARSM	Prediction and Transmissibility with age dependency	(Chaudhary et al., 2021)
8	BHRP	imitate the spread of bats among people	(Chamseddine et al., 2022)
9	SEIQR	Management studies, and Prediction	(Chaudhary et al, 2021)

Source: authors' own creation.

A thorough examination of the methodologies used to examine COVID-19 utilising machine learning-based techniques has been conducted. Segmentation and diagnosis in medical imaging have been the primary applications of the machine learning approach (Hu et al., 2022). Some studies have used deep learning methods (Apostolopoulos et al., 2020), while others have used supervised learning methods, including support vector machine (Marques et al., 2020), logistic regression (Chaudhary et al., 2021), decision trees (DT), random forest (Basiriet al., 2021; Chamseddine et al., 2022), and ARIMA models (Hernandez-Matamoros et al., 2020). Using the VOS viewer software, the research authors (Jalaber et al., 2020) thematically analysed tweets on COVID-19 to gauge public opinion on the spread of the disease. As a bonus, SPM methods were used to analyse tweets for relationships between commonly used terms and patterns. Researchers analysed the mutation rate in genomic sequences obtained from GenBank data on COVID-19 patients (Chaudhary et al., 2021). Table 1 displays numerous mathematical models that have been utilised in studies to analyse COVID-19, such as Markov Chain Monte Carlo known as the MCMC model, Susceptible-Exposed-Infected-Removed, Susceptible Infectious Quarantined Recovered known as the SIQR model, Susceptible Infected Recovered known as SIR model, Susceptible Exposed Symptomatic Asymptomatic Recovered Seafood Market, Bats Hosts Reservoir People Transmission, Susceptible Exposed Infectious Quarantined Recovered, SPSS Modeler, and ODE meta population model.

3. Materials and Method

The two main sub-objectives of this manuscript are to test the effectiveness of the pattern mining methods and the ability of sequence prediction approaches to predict nucleotide bases from previous nucleotide bases in the COVID-19 genome sequences.

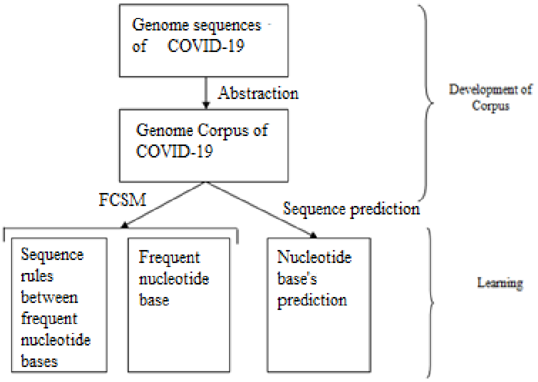
Genome sequences of COVID-19
Abstraction
Genome Corpus of COVID-19
FCSM
Sequence prediction
Sequence rules between frequent nucleotide bases
Frequent nucleotide base
Nucleotide base's prediction
Development of Corpus
Learning

Figure 3. The genome sequence analysis of COVID-19 is being done using the sequence prediction and FCSM technique
Source: authors' own creation.

This article aims to establish whether pattern mining can uncover intriguing COVID-19 genome patterns that can forecast nucleotide bases based on prior ones. This section discusses the technique suggested to address these two sub-goals. Several pattern mining algorithms have been developed and used on various datasets, including financial transactions, graphs, texts, and sequences (Eikenberry et al., 2020), all to discover interesting patterns in the data. These methods have found use in a wide variety of contexts. However, sequential or time-based organised data, such as genetic sequences, poses particular difficulties for traditional pattern mining techniques. They cannot find patterns that would characterise the sequence in which events occurred or added elements for this information. In response to this shortcoming, pattern-mining approaches for SPM have been developed (Chaudhary et al., 2021) to sift through structured, sequential data for discernible trends. In a collection of discrete sequences, FCSM can be used to identify meaningful patterns. The significance of a subsequence is influenced by several factors, including its frequency of occurrence, profitability, and duration. Sequences genome is a discrete sequence. Hence, we have decided to analyse them using FCSM techniques. The proposed sequence prediction and FCSM approach for analysing genome sequence analysis of COVID-19 is represented in Fig. 3. These models have two parts:

(1). Learning using Sequence Prediction Techniques and FCSM: The FCSM techniques are used in the corpus to identify sequential relationships between nucleotides, identify frequently occurring nucleotides, and determine the bases of a given sequence's subsequent nucleotides.

(2). Development of corpus: In this part, genome sequences of COVID-19 are converted into a discrete sequences corpus, where entire genome sequences are transferred into a nucleotide sequence.

The succession data for SARS-CoV-2 strains is obtained using the genomic sequence database GenBank (Akamatsu et al., 2021). According to the genome sequence corpora of COVID-19, which is downloaded from GenBank, every genome sequence has been represented in a file structure with a particular FASTA format that provides the names of the genes, which is follow the letters A, C, G, and T (nucleotides sequence). These show that when the gene field is taken out, the entire genome sequence is just a nucleotide series (represented as NS). Table 2 displays these genomic sequence corpora for COVID-19. Combining all these nucleotide sequences could result in a discrete sequence corpus. The public as- describes this corpus.

Table 2. CGSC sample

Sequences	Ids
{…..AATAACTCTATTGCCATACCCACAAATT….}	1
{…….TGCAGCAATCTTTTGTTGCAATATGGC…..}	2
{…….CAGGTGCTGCATTACAAATACCATTTG….}	3
{…….CCCTAATGTGTAAAATTAATTTTAGTA…..}	4

Source: https://zenodo.org/records/10682056.

Description 1 NBS is the combination of nucleotide bases A, C, G, and T. NBS stands for nucleotide base set. The cardinality of the set is indicated by the symbol |NBS|; in this case, depict |NBS| = 4 if four different nucleotides exist.

The genome sequence corpus of COVID-19 is represented as follows using the base set of nucleotide specification.

Description 2 (Genome Sequence of COVID-19) The Genome Sequence of COVID-19 is an ordered enumeration of the nucleotides and bases. GSC = NBS_1, NBS_2, ...,NBS_n, such that $NBS_i \subseteq NBS$ $(1 \leq i \leq n)$.

Description 3 (Corpus of Genome Sequence of COVID-19) A Corpus of Genome Sequence of COVID-19 is represented by a genome sequences list and it is denoted by CGSC19. CGSC19 = GSC_1, GSC_2, ..., GSC_p, where all genome sequences have their ID (Unique Identifier). For example, CGSC holds genome sequences with four lines with unique identifiers 1, 2, 3 and 4, represented in Table 2.

A genome is where the sequence of each row's nucleotides originates. The obtained sequence has been added back using positive integers for every nucleotide. As an illustration, Nucleotide 1 can be used in place of nucleotide A. The outcome will be 2, 3, and 4, respectively, which is replacement of the nucleotides C, G, and T.

3.1 Learning using FCSM

After the corpus preparation, the FCSM method is applied to discover nucleotide subsequence's (patterns) that become visible in genome sequences. However, a suitable measure must used to choose an exciting nucleotide subsequence. In consecutive set mining, we have to evaluate consecutive sets using a generally ordinary measure, and the occurrence frequency (support measure) (Eikenberry et al., 2020; Chaudhary et al., 2021).

3.1.1 Fuzzy Consecutive Set Mining (FCSM)

This measurement is appropriate for learning because it makes it possible to identify the subsequence bases of nucleotides that appear within various genome sequences, revealing their similarities. It frequently contains a thorough explanation of each regular subsequence in a collection of discrete sequences (Chaudhary et al., 2021). FCSM is defined as follows for the framework of analysis of COVID-19's genomic sequences:

Description 4 (Containment of Genome Sequence): Let the sequence of genome $S_x = x_1, x_2, ...,x_n$ is contained (or present) in a different genome sequence $S_y = y_1, y_2,..., y_m$, iff surrounding be present numbers $1 \leq j_1 < j_2 < ... < j_n \leq m$, like $x_1 \subseteq y_{i1}$, $x_2 \subseteq \beta_{i2}$, ..., $\alpha_n \subseteq \beta_{im}$ (represented as S_x S_y). In the event that S_x is held in S_y, S_x is assumed to be a S_y subsequence.

Description 5 (Support): The genomic subsequence supports S_x in the specified corpuses. CGSC refers to the entire collection of sequences that surround S_x. It is shown as sup(S_x) and is explained as follows: sup(S_x) is equal to sup(S_x) = $|\{S|S_x\ \ S \wedge S \in \text{CGSC19}\}|$.

Description 6 (Corpus of Genome sequence with FCSM) Let CGSC19 represent a dataset of genomic sequences, and min_sup represent a user-label least amount support threshold with a value greater than zero. The full frequent sub sequences of the genome are specified in a CGSC19's by FCSM job. Everywhere a subsequence of the genome S is found, it should be higher than or equal to min_sup and would be a sup(S) value.

Searching approaches like depth-first search and breadth-first search concept comprise in FCSM algorithms. The most popular scanning is used to look data set to discover the sequential patterns in a systematic manner, that hold a 1-sequence (particular item), known as the breadth-first search algorithm. After that, we extended 1-sequence in both i- and s-direction to produce 2-sequences. Likewise, using 2-sequences produces 3-sequences, and so on. These pattern production processes are maintained until they can produce nothing.

On the other hand, a depth-first search algorithm finds various patterns through different methods. Besides FCSM, it is one of the moments also exciting to locate nucleotide sets that commonly come out in genome sequences, not including any sequential order. For this intention, the FIM task (Frequent Itemset Mining) is measured, which could be observed as an FCSM particular case. In the background of this manuscript, frequent item set mining is defined as follows.

Definition 7 (FIM) Let CGSC19 is a genome sequence corpus of COVID-19 and minsup is a minimum support threshold which is user-defined like minsup> 0. Let nucleotide bases set represents as NBS, like NBS ⊆ NB'. In a corpus CGSC19, NBC's support is the sequence numbers with NBC's nucleotides. It is described like: sup(NBS) = $|\{S'|\exists a \in S'\ \forall a \in \text{NBS}\}|$ and symbolized like sup(NBS). The FIM's task in CGSC19 is to specify every frequent NBS. If sup(NBS) ≥ minsup, then represented as a nucleotide bases set is assumed to be frequent.

3.1.2 FIMAR Algorithm

The Apriori algorithm is the primary and generally renowned frequent item set mining method. In large databases, it is commonly used to locate frequent item sets. It helps to identify standard item sets that could improve to provide well-built item sets that functioned properly. NBS in this work (Item sets) taken out by the Fuzzy Item Set Mining with Association Rule algorithm is used to discover relationships (association rules) between different items. One more pattern type measured in this learning to analyse a genome sequence corpus is sequential rules. The FIMAR algorithm finds patterns by considering not only their confidence, but also their support. FIMAR algorithm is representing as a sequential rule mining job which is defined like this, whenever we deal with genome sequences.

Description 8 (SR) A sequential rule denoted as SR, and it is a relationship (A → B) among two NBS; A, B ⊆ NBS, such that A ∩ B = ϕ as well as A, B ≠ ϕ. A sequential rule SR: A → B is represented as if A's items happen in a sequence, B's items will take place afterward within a similar sequence.

Description 9 (A sequential rule's confidence and Support) An A is enclosed in S_x (write as A $\subseteq S_x$) if and only if A $\subseteq \bigcup_{j=1}^{m}\{xj\}$. A sequential rule SR: A → B is enclosed in S_x (SR $\subseteq S_x$) if and only if there be present an integer l like as $1 \leq l < m$, A $\subseteq \bigcup_{j=1}^{l}\{xi\}$ and B $\subseteq \bigcup_{j=l+1}^{m}\{xj\}$. The support and confidence of a sequence rule SR in a CGSC19 corpus are described as follows:

$$\mathrm{suf}_{CGSC19}(SR) = \frac{|\{S|SR \subseteq S \wedge S \in CGSC19\}|}{|CGSC19|}$$

$$\mathrm{Con}_{CGSC19}(SR) = \frac{|\{S|SR \subseteq S \wedge S \in CGSC19\}|}{|\{S|A \subseteq S \wedge S \in CGSC19\}|}$$

Description 10 (SRM) Let CGSC19 represent as a corpus of genome sequence of COVID-19 with threshold min_conf∈ [0, 1], and min_sup> 0. That is user-defined minimum confidence and minimum support. If and only if supCGSC19(SR) ≥ min_sup then sequence rule SR represents a frequent sequential rule and SR is a legal sequential rule if and only if that is frequent along with min_conf ≤ confCGSC19(SR). In a corpus, mining sequential rules is used to discover every suitable sequential rule.

The proposed algorithm implemented here to analyse genome sequence from the sample collected from various locations. First, the genome sequences are compared and the patterns in which the variation is available using the Text Matching Algorithm (TMA). The genome sequence is passed as input to the FIMAR algorithm, which finds the dependency patterns from the sequence by analysing frequent sets.

Proposed Algorithm (FCSM)
Input: *min_sup; the minimum support threshold and S; a sequence database*
Output: *The entire collection of sequential patterns as an output.*
Method: *Call Text Matching Algorithm (a=sequence1, b=sequence2)*
Subroutine: *Columns= [line_no, sequence_ratios]*
Dataset: *DS1_genome, DS2_genome*
Step-1: Find Sequence from data frame in columns with two input values i.e., line _no and sequence ratios.

df_S←DF(c(i,j));

// where df_S= *data frame sequence, DF = data frame, C = column with two parameter line _no and sequence ratios*

Step-2: Open both input files and then enumerate the sequence matcher to compare both strings with their parameter values and count the value one by one from the data frame in columns with two input values, i.e. line _no and sequence ratios.

```
                    α← input_file1
                    β← input_file2
Step -3:for i from s1 to input_file1 do
Step-4: for j from s2to input_file2 do i[(a,b),1]
            s = TMA(a,b);
            df_ls [len(df_is)] = [i,j];
            df_s [i] = df_s [i].astype(int)count_lines+1;
            df_sm = df_s[df_s[i]<1];
            G_lc= len(df_sm);
            P_mv = (G_lc /count_lines)*100;
```

where α *= data from file1, β = data from file2, s1 = sequence1, s2= sequence2, s = sequence,* df_ls= *data frame local sequence,* df_is= *data frame lindex sequence,* df_sm= *data frame sequence missing,* G_lc= *length_changed_genome,* P_mv= *perc_missing_values,* P_GSc= *Percentage of changed Genome Sequence*

Step-5: Combining the collection of often occurring items with newly generated items yields the candidate set.

Step-6: FIMAR(X, min_support, min_conf)

Step-7: If the candidate set contains an uncommon group of items, you should conduct subset tests and remove them.

Step-8: while(i):

Fk = {c | c ∈ Ck ∧ σ(c) ≥ N × minsup}. {The often occurring k-itemsets}

Step-9: By obtaining those that meet the minimum requirements, determine the final frequent itemset.

Step-10*: end while*

Step-11*:* Stop

3.2 Data Set

GenBank is home to viral genome sequences in .fasta and .gb format. We employed 200 unique coding sequences from across all four viral families. Among the many well-known, freely accessible databases of nucleotide sequences, GenBank is especially noteworthy for its support of biological annotation. It has been a roughly annual increase in the number of sequence files in GenBank during the last several decades.

4. Result Discussion

This segment represents outcomes achieved by applying the methods represented in the earlier part of the genome sequences of COVID-19 achieved from the NCBI dataset, GenBank. This experimentation is executed on a fifth-generation HP laptop with processor Core i5 and 8 GB RAM.

The findings of testing the FIMAR-based system used to categorise genomic sequences are represented in this section. Although it is often used, more than classification accuracy has been required to algorithm performance evaluation. In this study, we used various assessment measures to determine the system's effectiveness, as mentioned above. Genome analysis has been performed using the python-based FCSM algorithm. This section gives detail the outcomes of running algorithms on the corpus.

Table 3 presents some statistical information on the gathered genome sequences. The accession number of each genome sequence has been shown by the column headed "ID." The NCBI GenBank makes each sequence available for download in three different formats: protein, coding region, and nucleotide. The genome's nucleotide sequences were recovered and are shown in Table 3.

Table 3. COVID-19 genome's characteristics occupied from NCBI Dataset

ID	Location	Length	Release Date	Collection Date
MT291827	China: Wuhan	29858	06-04-2020	30-12-2019
MT291828	China: Wuhan	29858	06-04-2020	30-12-2019
MT750058	USA: Wisconsin	29782	13-07-2020	09-06-2020
MT750057	USA:Illinois	29782	13-07-2020	17-06-2020
MT745584	Bahrain	29860	13-07-2020	22-06-2020

Source: https://zenodo.org/records/10682056.

Table 4 lists how frequently each nucleotide was found in four more genomic sequences of COVID-19. The quantity of nucleotides containing the bases adenine and guanine in their sequences varies between the two distinct strains (MT745584 and MT750058). One less Adenine and one more Guanine are present in the sequence of MT750058 compared to MT745584.

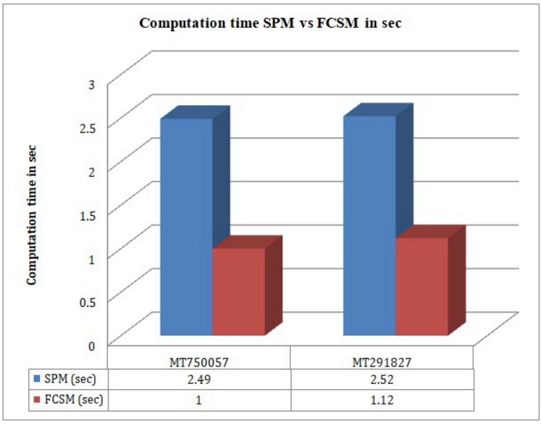

Figure 4. Computation time SPM vs FCSM (in sec)
Source: https://zenodo.org/records/10682056.

Table 4. Proportion of nucleotides in various COVID-19 genomes for datasets

Base	% GC Length	Content China Dataset Percentage of genome	% GC Length	Content US Dataset Percentage of genome
Adenine	8891	29.85	8932	29.91
Guanine	5849	19.64	5859	19.62
Thynime	9572	32.14	9585	32.1
Cytosine	5470	18.37	5482	18.36

Source: https://zenodo.org/records/10682056.

Figure 4 shows the SPM's computation time and the FCSM's computation time for particular genome ID MT750057, MT291827, and Figure 5 shows the percentage variation of two different datasets i.e., USA and China. Also, shows that how COVID-19's alignment resembles that of the other genome sequences.

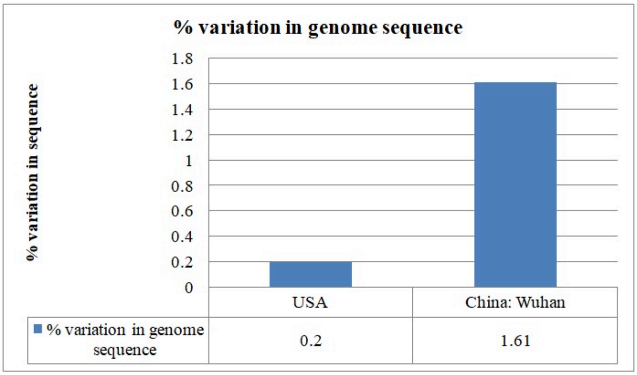

Figure 5. Percentage variation change in genome Sequence for Datasets
Source: authors' own creation.

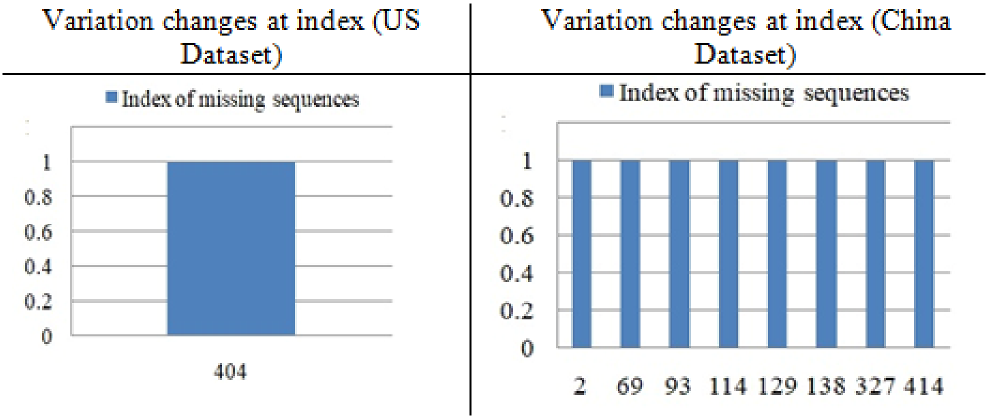

Figure 6. Total changes at index level in genome Sequence for Datasets
Source: authors' own creation.

When determining the degree of similarity between four distinct kinds of genome sequences, we relied on the sequence alignment approach. The sequence alignment places many sequences in a predetermined order, which makes it easier to spot areas of overlap between them. Table 5 analyses frequent sets generated by Apriori vs. FIMAR, that FIMAR recovered from the A genome sequence for various min_support values. Only four distinct common patterns were produced by FIMAR when the min_support values were between 55% and 100%. The Apriori method produced 18 patterns once it lowered the minimum support threshold to 1%. Fig. 6 represents the total change at the index level in genome sequence for different datasets (US and China datasets).

Table 5. Analysis of frequent sets generation by Apriori vs FIMAR

Patterns	Pattern	Minsupport	Support	Support (Apriori)	Support (FIMAR)	Minsupport
A	AGT	100%	8915	52	55	10%
C	ACT	100%	5487	48	50	5%
G	CGT	100%	5859	32	36	5%
T	ACG	100%	9599	12	15	1%

Source: https://zenodo.org/records/10682056.

Table 6. Performance Comparison on varying minsup

	Apriori		FIMAR	
Min. Sup (%)	Time (Sec)	Patterns	Time (Sec)	Patterns
33%	2	3549	3.2	3645
25%	42	194361	32	194378
20%	231	1372868	218	1372881

Source: https://zenodo.org/records/10682056.

The performance of the pattern mining procedure was satisfactory, it was quick. The performance of Apriori and FIMAR is shown in Table 6, which compares many distinct min_support threshold settings. It has been shown that when min_support is decreased, FCSM can uncover more common patterns, even though this increases both runtime and memory utilisation. It is clear from this figure that the distribution of nucleotides differs throughout the various genome sequences. After that, it used FCSM methods to uncover previously unknown sequential connections between nucleotides. The FCSM method was carried out, which entails establishing a min_sup threshold.

5. Conclusions

This research provided different strategies for investigating and analysing SARS-CoV-2 genomic sequences. The primary method involves using fuzzy itemset mining with association rules mining (FIMAR) algorithms to find shared sequences with nucleotide bases in addition to their sequential and regular patterns. Tested several sequence prediction methods on genomic sequences, and the findings indicated that FIMAR performed much better than other approaches that are considered to be state-of-the-art. Other viruses besides the SARS-CoV-2 virus can be treated using the methods described in this article. They could be applied to research several other human viruses. In the not too distant future, we may broaden the scope of our effort to include the examination and categorisation of additional genome sequences. Researchers are highly encouraged to explore alternative approaches based on artificial intelligence to study and classify genomic sequences. Based on this research, we have worked on two data sets, one of the US data sets and the second one of China data sets collected from the NCBI repository. Firstly, we identify the itemsets and then calculate the computation time using FCSM and compare it with SPM, which is measured in seconds; here, we extract the itemsets for two genomes, MT750057 and MT291827, and it is computed within 1 second and in 1.12 seconds, respectively. After that, we observe the percentage variation change in genome sequence for different datasets; also examined the total change at index level in genome sequence for different datasets, and analysis of frequent itemsets generation by Apriori vs FIMAR is also done for different patterns (A, C, G, and T). In this work, we also presented an analysis of the nucleotide substitution rate at each isolation step.

References

[1] Ahmed, I., Jeon, G. (2022), *Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. Interdisciplinary sciences: computational life sciences*, 14(2), 504-519.

[2] Alali, Y., Harrou, F., Sun, Y. (2022), *A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. Scientific Reports*, 12(1), 2467.

[3] Apostolopoulos, I.D., Mpesiana, T.A. (2020), *Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and engineering sciences in medicine*, 43, 635-640.

[4] Arslan, H., Arslan, H. (2021), *A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. Engineering Science and Technology. An International Journal*, 24(4), 839-847.

[5] Barnett, K.C., Li, S., Liang, K., Ting, J.P.Y. (2023), *A 360 view of the inflammasome: Mechanisms of activation, cell death, and diseases. Cell*, 186(11), 2288-2312.

[6] Bagabir, S.A., Ibrahim, N.K., Bagabir, H.A., Ateeq, R.H. (2022). *Covid-19 and Artificial Intelligence: Genome sequencing, drug development and vaccine discovery. Journal of Infection and Public Health*, 15(2), 289-296.

[7] Basiri, A., Mansouri, F., Azari, A., Ranjbarvan, P., Zarein, F., Heidari, A., Golchin, A. (2021), *Stem cell therapy potency in personalizing severe COVID-19 treatment. Stem Cell Reviews and Reports*, 17, 193-213.

[8] Chamseddine, E., Mansouri, N., Soui, M., Abed, M. (2022), *Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss. Applied Soft Computing*, 129, 109588.

[9] Chamola, V., Hassija, V., Gupta, S., Goyal, A., Guizani, M., Sikdar, B. (2020), *Disaster and pandemic management using machine learning: a survey. IEEE Internet of Things Journal*, 8(21), 16047-16071.

[10] Chaudhary, J.K., Yadav, R., Chaudhary, P.K., Maurya, A., Kant, N., Rugaie, O. A., ... , Dhamija, P. (2021), *Insights into COVID-19 vaccine development based on immunogenic structural proteins of SARS-CoV-2, host immune responses, and herd immunity. Cells*, 10(11), 2949.

[11] Fan, C., Cai, T., Gai, Z., Wu, Y. (2020), *The relationship between the migrant population's migration network and the risk of COVID-19 transmission in China—empirical analysis and prediction in prefecture-level cities. International Journal of Environmental Research and Public Health*, 17(8), 2630.

[12] Hassan, B., Izquierdo, E., Piatrik, T. (2021), *Soft biometrics: a survey: Benchmark analysis, open challenges and recommendations. Multimedia Tools and Applications*, 1-44.

[13] Himeur, Y., Al-Maadeed, S., Varlamis, I., Al-Maadeed, N., Abualsaud, K., Mohamed, A. (2023), *Face mask detection in smart cities using deep and transfer learning: lessons learned from the COVID-19 pandemic. Systems*, 11(2), 107.

[14] Hu, Q., Gois, F.N.B., Costa, R., Zhang, L., Yin, L., Magaia, N., de Albuquerque, V.H.C. (2022), *Explainable artificial intelligence-based edge fuzzy images for COVID-19 detection and identification. Applied Soft Computing*, 123, 108966.

[15] Hernandez-Matamoros, A., Fujita, H., Hayashi, T., Perez-Meana, H. (2020), *Forecasting of COVID19 per regions using ARIMA models and polynomial functions. Applied soft computing*, 96, 106610.

[16] Jalaber, C., Lapotre, T., Morcet-Delattre, T., Ribet, F., Jouneau, S., Lederlin, M. (2020), *Chest CT in COVID-19 pneumonia: A review of current knowledge. Diagnostic and interventional Imaging*, 101(7-8), 431-437.

[17] Khanra, S., Dhir, A., Islam, A.N., Mäntymäki, M. (2020), *Big data analytics in healthcare: a systematic literature review. Enterprise Information Systems*, 14(7), 878-912.

[18] Li, J., Liu, H.H., Yin, X.D., Li, C.C., Wang, J. (2021), *COVID-19 illness and autoimmune diseases: recent insights. Inflammation Research*, 70, 407-428.

[19] Li, C.X., Noreen, S., Zhang, L.X., Saeed, M., Wu, P.F., Ijaz, M., ..., Li, J.H. (2022), *A critical analysis of SARS-CoV-2 (COVID-19) complexities, emerging variants, and therapeutic interventions and vaccination strategies. Biomedicine & Pharmacotherapy*, 146, 112550.

[20] Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Mulders, D.G., Molenkamp, R., Perez-Romero, C.A., ..., Kraneveld, A.D. (2021), *Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Scientific reports*, 11(1), 947.

[21] Mercatelli, D., Holding, A.N., Giorgi, F.M. (2021), *Web tools to fight pandemics: the COVID-19 experience. Briefings in bioinformatics*, 22(2), 690-700.

[22] Makaremi, S., Asgarzadeh, A., Kianfar, H., Mohammadnia, A., Asghariazar, V., Safarzadeh, E. (2022), *The role of IL-1 family of cytokines and receptors in pathogenesis of COVID-19. Inflammation Research*, 71(7-8), 923-947.

[23] Mousavizadeh, L., Ghasemi, S. (2021), *Genotype and phenotype of COVID-19: Their roles in pathogenesis. Journal of Microbiology, Immunology and Infection*, 54(2), 159-163.

[24] Mahum, R., Munir, H., Mughal, Z.U.N., Awais, M., Sher Khan, F., Saqlain, M., ..., Tlili, I. (2023), *A novel framework for potato leaf disease detection using an efficient deep learning model. Human and Ecological Risk Assessment: An International Journal*, 29(2), 303-326.

[25] Mishra, S.K., Tripathi, T. (2021), *One-year update on the COVID-19 pandemic: Where are we now? Acta tropica*, 214, 105778.

[26] Madhav, A.S., Tyagi, A.K. (2022), *The world with future technologies (Post-COVID-19): open issues, challenges, and the road ahead. Intelligent Interactive Multimedia Systems for e-Healthcare Applications*, 411-452.

[27] Marques, G., Agarwal, D., de la Torre Díez, I. (2020), *Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. Applied soft computing*, 96, 106691.

[28] Dubey, S., Verma, D.K., Kumar, M. (2024), *Severe acute respiratory syndrome Coronavirus-2 GenoAnalyzer and mutagenic anomaly detector using FCMFI and NSCE. International Journal of Biological Macromolecules*, 258, 129051.

[29] Rohaim, M.A., Clayton, E., Sahin, I., Vilela, J., Khalifa, M.E., Al-Natour, M. Q., ..., Munir, M. (2020), *Artificial intelligence-assisted loop mediated isothermal amplification (AI-LAMP) for rapid detection of SARS-CoV-2. Viruses*, 12(9), 972.

[30] Shuja, J., Alanazi, E., Alasmary, W., Alashaikh, A. (2021), *COVID-19 open source data sets: a comprehensive survey. Applied Intelligence*, 51, 1296-1325.