Abstract:

The genome of the COVID-19 disease was sequenced for the first time in January 2020, in Wuhan, China's capital. COVID-19 genome sequencing is important for understanding the virus's origin and mutation rate, as well as the creation of vaccines and successful prevention strategies. The use of artificial intelligence techniques to learn details from COVID-19 genome sequences is investigated in this paper. SPM is first used on a computer understandable corpus of COVID-19 genome sequences to see whether any hidden patterns can be discovered, such as nucleotide base patterns and their relationships. Second, the corpus is subjected to sequence prediction models to see whether nucleotide base(s) can be predicted from previous ones. Third, an algorithm is designed to identify the positions in the genome sequences where the nucleotide bases are modified and determine the mutation rate for mutation analysis in genome sequences.

Introduction:

The SARS-CoV-2virus was discovered in a pneumonia patient in Wuhan, China's capital. According to the World Health Organization's most recent survey, the COVID-19 has affected more than 65 million people, with almost half of them dying. 1.5 million people have died as a result of the epidemic, which has now spread to over 200 countries. To establish effective vaccines that produce long-term immunity, it is important to understand the genome of the SARS-CoV-2and its functionalities . The genome of an organism is the accumulation of its entire genetic potential, encoded as a four-nucleotide sequence (Adenine A, Guanine-G, Cytosine-C, and Thymine-T) that makes up the nucleic acids The COVID-19 genome is made up of single-stranded DNA. Genome decoding is the process of determining the nucleotide sequence in a genome. Several groups have sequenced the SARS-CoV-2 genome worldwide, revealing several strains of the virus and revealing that its genome is 79 percent identical to SARS-CoV-1 and 50 percent similar to MERS-CoV . Many facets of disease behavior remain unclear, making actionable observations difficult to come by. The use of artificial intelligence methods including sequential pattern mining (SPM), has the ability to accelerate the process of finding actionable insights and eventually contribute to a better global response . . SPM [a form of hierarchical data mining], has been used in genomics to find patterns of unique elements in genes ,study gene expression, and mine the most contiguous frequent patterns. Second, see whether the next nucleotide bases can be predicted in COVID-19 genome sequences. Third, we suggest an algorithm to find mutations in genome sequences

Related Work

This segment discusses recent work on the use of AI-based approaches for the diagnosis, identification, forecasting, and prediction of COVID19. In COVID-19 research, a study presented a detailed description of the use of mathematical models and AI-based techniques. Machine learning and data processing are examples of AI. Medical imaging (such as X-ray and computed tomography (CT)) segmentation and diagnosis have largely relied on machine learning (and deep learning) techniques. Deep learning methods were used to diagnose and track COVID 19 from CT scans and X-ray images, using supervised learning methods such as support vector machine (SVM) in , logistic regression (LR) in , decision trees (DT), random forest (RF) , and ARIMA models .As a result of the COVID-19 epidemic, there have been a number of responses. SPM methods were often used to find common words/patterns in tweets, as well as their relationships. the mutation rate in genomic sequences gathered from COVID-19 patient data from GenBank was investigated. The rate of missense nucleotide mutation and the rate of codon mutation were first discovered in genomes. After that, a long short-term memory (LSTM) model based on recurrent neural networks was used.to forecast the virus's mutation rate in the future The authors of the research based on base substitution mutation rates and ignored insertion and deletion rates. also several methods are produced to map

SARSCoV-2 genomic variations. However, it is common knowledge that an algorithm's success on test data does not ensure that it will behave equally when implemented in the field. The primary explanation for this is that real world data is more susceptible to noise and other artefacts than training and test data. In the other hand .

Determine whether pattern mining can uncover fascinating patterns in COVID-19 genome sequences and whether gene prediction models can forecast nucleotide bases based on previous ones . Several pattern mining techniques have been designed and implemented on various types of databases .as well as gene sequences They are unable to identify patterns representing temporal associations between events or elements in such data. To over come this constraint, SPM techniques that can mine patterns in hierarchical sequential data have been developed. SPM entails finding essential subsequences (patterns) in a series of discrete sequences and calculating the value of each subsequence .sequence prediction models are used to see how the next nucleotide bases in the sequence can be determined from the previous ones in the sequence .A sequence of the genome Compact Prediction Tree (CPT), CPT+, Dependency Graph (DG), All-K-Order-Markov (AKOM), Transition Directed Acyclic Graph (TDAG), and LZ78

Methods/Analyzing :

This section outlines the proposed strategy for achieving the paper's first two sub-goals: determining whether pattern mining can uncover fascinating patterns in COVID-19 genome sequences and determining whether gene prediction models can predict nucleotide bases from previous ones. SPM techniques that can mine patterns in organized sequential data have been created. SPM entails finding essential subsequences (patterns) in a series of discrete sequences and calculating the value of each subsequence. Using various metrics such as a subsequence's occurrence frequency, profit, and duration. We choose SPM techniques to study genome sequences because they are a kind of discrete sequence.

The second sub-goal of this paper is to see how the next nucleotide bases can be predicted from previous ones using state-of-the-art sequence prediction models. a sequence of the genome Compact Prediction Tree (CPT), CPT+, Dependency Graph (DG), All-K-Order-Markov (AKOM), Transition Directed Acyclic Graph (TDAG), and LZ78 are the models considered.

-The overall method for studying COVID19 genome sequences using SPM and gene prediction models that has been suggested. It is divided into two sections:

1. Development of the corpus: Sequences of the COVID-19 genome

Each whole genome sequence is translated into a sequence of nucleotides, resulting in a corpus of discrete sequences.

2. Learning through SPM and Sequence Prediction Techniques: SPM algorithms are used to find commonly occurring nucleotides in the corpus .Relationships between nucleotides, and to predict the nucleotide base(s) of the next nucleotide in a series .The next two subsections go into these two pieces in greater depth. The findings are then presented .

using SPM :

Following the preparation of the corpus, various SPM techniques may be used to detect patterns (nucleotide sequences) in genome sequences. However, choosing fun items is a challenge .A suitable measure must be used to identify trends. The help metric (occurrence frequency) is the most common way to test patterns in pattern mining . This paper allows for simultaneous objects in a series. However, since nucleotides in genome sequences are often fully arranged, this situation is not discussed in this article.

Sequence prediction:

The next nucleotides in a genome sequence is predicted to see how predictable it is. For predicting the next nucleotides and their sequences, many models were compared.

a trend Each model is first trained on nucleotides and sequence patterns.

The next nucleotides and their patterns in a sequence are then predicted using a simulation model.

The patterns of the next nucleotides and their prediction are dependent on the nucleotide scores determined by the model For eg, CPT+ predicted T for the sequence A, C, and ACT, which is a common codon that encodes the am

Sequence prediction techniques:

Another learning task in this analysis was to build sequence prediction models using the COVID-19 genome sequences to see whether the nucleotide bases were arranged in the correct order.is well-known. To decide which model works well, many common models are used. CPT+, CPT, DG, AKOM, TDAG, and LZ78 are among the versions used. amino acid Threonine.

COVID-19 Genomes mutation analysis:

This section outlines the proposed method for identifying mutations in the COVID-19 genome sequences, which is the paper's third sub-goal. COVID19 causes a number of illnesses, ranging from asymptomatic to fatal respiratory failure, although the exact mechanism is unknown at this time. The SARS-CoV-2 virus, like many other species that split and propagate,

To best respond to new conditions, it is continually improving by modifying a few letters (nucleotides) at a time. The evolution mechanism remains unknown since it evolves slowly relative to most viruses, resulting in fewer mutations. Research The coronavirus' genome undergoes approximately two modifications every month on average. The majority of modifications to the COVID-19 genome structure are unlikely to impact the virus's behavior, however a few can. However, the analysis was criticized because the researchers could not show that the mutation was the source of its dominance; it may have benefited from other factors. Determinants or by chance Nonetheless, it is critical to comprehend the virus's mutation pattern as well as its mutation rate.

Results:

The COVID19 genome sequences collected from the NCBI GenBank were subjected to the techniques described in the previous section, and the findings are discussed in this section. All of the tests were carried out on an HP laptop with an Intel processor ,8 GB RAM and a fifth-generation Core i5 cpu. the order of things Each sequence in the NCBI GenBank can be downloaded as a nucleotide, coding region, or protein. The genome sequences were downloaded in nucleotide format In JAVA, the SPMF data mining library was developed. is a tool for analyzing gene sequences. SPMF is a pattern mining application that is opensource and cross-platform . It has over 180 data mining algorithms implemented in it.