Genomics Paper

COVID-19 Genome Analysis Using SPM

Ahmed Hussein Ismail

21-4-2021

Abstract

The genome of the novel coronavirus (COVID-19) disease was first sequenced in January 2020, approximately a month after its emergence in Wuhan, capital of Hubei province, China. COVID-19 genome sequencing is critical to understanding the virus behavior, its origin, how fast it mutates, and for the development of drugs/vaccines and effective preventive strategies. This paper investigates the use of artificial intelligence techniques to learn interesting information from COVID-19 genome sequences. Sequential pattern mining (SPM) is first applied on a computer-understandable corpus of COVID-19 genome sequences to see if interesting hidden patterns can be found, which reveal frequent patterns of nucleotide bases and their relationships with each other. Second, sequence prediction models are applied to the corpus to evaluate if nucleotide base(s) can be predicted from previous ones. Third, for mutation analysis in genome sequences, an algorithm is designed to find the locations in the genome sequences where the nucleotide bases are changed and to calculate the mutation rate. Obtained results suggest that SPM and mutation analysis techniques can reveal interesting information and patterns in COVID-19 genome sequences to examine the evolution and variations in COVID-19 strains respectively.

Introduction

After declaring that Coronavirus 2 (SARS-CoV-2) virus, also known as COVID-19 as a global epidemic on March 11, 2020, no effective therapeutic or vaccine has not emerged yet, due to the novelty of the virus and its behavior.The COVID-19 genome sequence is made from a single-stranded sequence of nucleotides called RNA and is approximately 30 Kb long.Genome of SARS-CoV-2 has been sequenced by different groups around the world which revealed multiple strains of the virus.Identification of genome characteristics helps biomedical experts to produce hypotheses about the effect of these characteristics on the disease manifestations in the population.The use of artificial intelligence methods including sequential pattern mining (SPM), has the potential to accelerate the process of finding actionable insights contributing to a better global response. The pattern analysis field provides efficient computer based techniques that enable humans, particularly bioinformaticians, to analyze complex and large genetic and genomic data . SPM, a special case of structured data mining, has been applied in genomics to find patterns of specific elements in genes, to analyze gene expression, to mine maximal contiguous frequent patterns from DNA sequence datasets, to discover motifs in DNA sequences, to predict protein function and diseases, to discover gene interactions and their characterizations, to interpret patterns extracted from DNA microarrays, to mine k-mers and to construct the phylogenetic tree. Using SPM on sequential genome data can provide new insight about virus mutations, virulence and the various disease manifestations. Moreover, discovering important hidden information in genomes by using SPM can help speed up the process of biological research and is of great significance to the biological world. The general goal of this paper is to explore the use of artificial intelligence techniques for COVID-19 genome analysis.

Related Works

Recent work done on the use of AI-based techniques for the diagnosis, detection, forecasting and prediction of COVID-19 is discussed in this section.A review provided a comprehensive overview on the use of mathematical models and AI-based techniques in COVID-19 studies. AI (machine learning, data mining and deep learning) techniques have been used mostly for medical imaging (such as X-ray and computed tomography(CT)) segmentation and diagnosis. For example, COVID-19 diagnosis and detection from CT scans andX-ray images were done using deep learning techniques, using supervised learning techniques such as support vector machine (SVM), using logistic regression (LR) and using decision trees (DT), random forest (RF) in and ARIMA models.

For text based COVID-19 related data, the study conducted a thematic analysis of COVID-19 related tweets with the VOSviewer software to examine general public reactions related to the COVID-19 outbreak. Moreover, SPM techniques were used to find frequent words/patterns and their relationship in tweets. The mutation rate was studied in genomic sequences gathered from COVID-19 patients data from GenBank. The missense nucleotide mutation rate and codon mutation rate were first found in genomes. After that, a recurrent neural network based long short-term memory (LSTM) model was used to predict the future mutation rate of this virus. In the study, authors focused on the base substitution mutation rates and did not consider the insertion and deletion rates.

Methodology

Results