View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Sunway Institutional Repository

Chaos, Solitons and Fractals 138 (2020) 110018

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model

Refat Khan Pathan[a], Munmun Biswas[a], Mayeen Uddin Khandaker[b,*]

[a] Department of Computer Science and Engineering, BGC Trust University Bangladesh, Chittagong-4381, Bangladesh

[b] Centre for Biomedical Physics, School of Healthcare and Medical Sciences, Sunway University, 47500 Bandar Sunway, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 22 May 2020
Accepted 12 June 2020
Available online 13 June 2020

Keywords:
SARS-Cov-2
Gene sequence
Mutation rate
Neural Network
LSTM model

ABSTRACT

SARS-CoV-2, a novel coronavirus mostly known as COVID-19 has created a global pandemic. The world is now immobilized by this infectious RNA virus. As of June 15, already more than 7.9 million people have been infected and 432k people died. This RNA virus has the ability to do the mutation in the human body. Accurate determination of mutation rates is essential to comprehend the evolution of this virus and to determine the risk of emergent infectious disease. This study explores the mutation rate of the whole genomic sequence gathered from the patient's dataset of different countries. The collected dataset is processed to determine the nucleotide mutation and codon mutation separately. Furthermore, based on the size of the dataset, the determined mutation rate is categorized for four different regions: China, Australia, the United States, and the rest of the World. It has been found that a huge amount of Thymine (T) and Adenine (A) are mutated to other nucleotides for all regions, but codons are not frequently mutating like nucleotides. A recurrent neural network-based Long Short Term Memory (LSTM) model has been applied to predict the future mutation rate of this virus. The LSTM model gives Root Mean Square Error (RMSE) of 0.06 in testing and 0.04 in training, which is an optimized value. Using this train and testing process, the nucleotide mutation rate of 400th patient in future time has been predicted. About 0.1% increment in mutation rate is found for mutating of nucleotides from T to C and G, C to G and G to T. While a decrement of 0.1% is seen for mutating of T to A, and A to C. It is found that this model can be used to predict day basis mutation rates if more patient data is available in updated time.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

The whole world is suffering by an ongoing pandemic due to Coronavirus disease brought by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. It was an outbreak from Wuhan, the capital of Hubei province in China during December 2019 [2]. The virus was identified on 7th January and observed that it is spread by human-to-human transmission via droplets or direct contact [3,4]. Its infection has been estimated to be a mean incubation period of 6.4 days and a basic reproduction number of 2.24-3.58. Since its identification, it has already been spread speedily over the whole globe, therefore the world health organization (WHO) had declared COVID-19 a global pandemic on 11th March 2020 [5].

The SARS-CoV-2 is a pathogenic human coronavirus under the Betacoronavirus genus. In the recent decade, the other two pathogenic species SARS-CoV and MERS-CoV were outbreaks in 2002 and 2012 in China and the Middle East, respectively [6–9]. The complete genomic sequence (Wuhan-HU1) of this large RNA virus (SARS-CoV-2) was first discovered in the laboratory of China on 10th January [10] and placed in the NCBI GenBank. The SARS-CoV-2 is a single positive-stranded RNA virus having non-segmented in nucleic acid sequence [11]. Although it is an RNA virus but for simplicity of understanding the gene sequence has been given as DNA type which means nucleobase Uracil (U) has been replaced by Thymine (T). The genomic sequence of SARS-CoV-2 virus shows about 79% and 50% similarity with the SAR-CoV and MARS-CoV, respectively [6].

SARS-CoV-2 performs mutation during replication of genomic information [12]. The mutation occurs due to some errors when copying RNA to a new cell. Mutations are mainly three kinds: Base substitution, Insertion, and Deletion. Further, in base substitutions, there are some more divisions: silent, nonsense, missense, and frameshift [13]. Micro-level alteration of mutation rate is also de-

* Corresponding author. Centre for Biomedical Physics, School of Healthcare and Medical Sciences, Sunway University, 47500 Bandar Sunway, Selangor, Malaysia.

E-mail addresses: mu_khandaker@yahoo.com, mayeenk@sunway.edu.my (M.U. Khandaker).

https://doi.org/10.1016/j.chaos.2020.110018
0960-0779/© 2020 Elsevier Ltd. All rights reserved.