Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, **2021**, *22*, 583-595

RESEARCH ARTICLE

Exploring the Lethality of Human-Adapted Coronavirus Through Alignment-Free Machine Learning Approaches Using Genomic Sequences

Rui Yin[1,2,*], Zihan Luo[3] and Chee Keong Kwoh[1]

[1]*School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore;* [2]*Department of Biomedical Informatics, Harvard University, Boston, MA 02138, USA;* [3]*School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, 430074, China*

Abstract: ***Background*:** A newly emerging novel coronavirus appeared and rapidly spread worldwide and World Health Organization declared a pandemic on March 11, 2020. The roles and characteristics of coronavirus have captured much attention due to its power of causing a wide variety of infectious diseases, from mild to severe, on humans. The detection of the lethality of human coronavirus is key to estimate the viral toxicity and provide perspectives for treatment.

***Methods*:** We developed an alignment-free framework that utilizes machine learning approaches for an ultra-fast and highly accurate prediction of the lethality of human-adapted coronavirus using genomic sequences. We performed extensive experiments through six different feature transformation and machine learning algorithms combining digital signal processing to identify the lethality of possible future novel coronaviruses using existing strains.

***Results*:** The results tested on SARS-CoV, MERS-CoV and SARS-CoV-2 datasets show an average 96.7% prediction accuracy. We also provide preliminary analysis validating the effectiveness of our models through other human coronaviruses. Our framework achieves high levels of prediction performance that is alignment-free and based on RNA sequences alone without genome annotations and specialized biological knowledge.

***Conclusion*:** The results demonstrate that, for any novel human coronavirus strains, this study can offer a reliable real-time estimation for its viral lethality.

ARTICLE HISTORY

Received: August 26, 2021
Revised: December 02, 2021
Accepted: December 14, 2021

DOI:
10.2174/1389202923666211221110857

Keywords: Coronavirus, lethality inference, alignment-free, machine learning, genomic nucleotide, SARS-CoV.

1. INTRODUCTION

Coronaviruses are positive, single-stranded RNA virus and have been identified in humans and animals. They are categorized into four genera: α, β, γ and θ [1]. Previous phylogenetic analysis revealed a complex evolutionary history of coronavirus, suggesting ancient origins and crossover events that can lead to cross-species infections [2, 3]. Bats and birds are a natural reservoir for coronavirus gene pool [4]. The mutation and recombination play critical roles that may enable cross-species transmission into other mammals and humans [5]. Human coronavirus (HCoV) was first identified in the mid-1960s [6], and up to now, seven types of coronavirus can infect people. Four of them, *i.e.,* HCoV--229E, HCoV-NL63, HCoV-OC43 and HCoV-HKU1, usually cause mild to moderate upper-respiratory tract illnesses like common cold when infecting humans [7]. The other three members include severe acute respiratory syndrome coronavirus (SARS-CoV) and middle east respiratory syndrome coronavirus (MERS-CoV) and the most lately severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). They all belong to Betacoronavirus that led to the epidemics or pandemics [8-10].

Emerging in November 2002 in Guangdong province, China, SARS caused 8,096 human infections with 774 deaths by July 2003 [11]. MERS was first reported in Saudi Arabia in September 2012 and finally resulted in 2,494 human infections by November 2019 [12] Recently, a novel coronavirus named SARS-CoV-2 is emerging and spreading to 215 countries or territories on June 12, 2020, leading to 7,390,702 confirmed cases with 417,731 deaths according to the World Health Organization [13]. Though precautions such as lockdown of cities and social distance have been taken to curb the transmission of COVID-19, it spreads far more quickly than the SARS-CoV and MERS-CoV diseases [4, 14]. To make matters worse, the number of infected cases still increases rapidly and the global inflection point about COVID-19 is unknown. Like other RNA viruses, *e.g.* influenza virus [15], coronaviruses possess high mutation and gene recombination rates [16], which makes constant evolution of this virus with the emergence of new variants. From

*Address correspondence to this author at the School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore; E-mail: yinr0002@e.ntu.edu.sg

1389-2029/21 $65.00+.00

SARS in 2002 to COVID-19 in 2019, coronaviruses have caused high morbidity and mortality, and unfortunately, the fast and untraceable virus mutations take the lives of people before the immune system can produce the inhibitory antibody [17] Currently, no miracle drug or vaccines are available to treat or prevent the humans infected by coronaviruses [18, 19]. Therefore, there is a desperate need for developing approaches to detect the lethality of coronaviruses not only for SARS-CoV-2 but also the potential new variants and species. This would facilitate the diagnosis of coronavirus clinical severity and provide decision-making support.

The detection of viral lethality has already been explored in influenza viruses [20, 21]. Through a meta-analysis of predicting the virulence and antigenicity of influenza viruses, we can infer the lethality of the virus timely to improve the current influenza surveillance system. Regarding the risk of novel emerging coronavirus strains, much attention has been captured to investigate the lethality or clinical severity of new emerging coronavirus. Typically, epidemiological models are certainly built to estimate the lethality and the extent of undetected infections associated with the new coronaviruses. Bastolla suggested an orthogonal approach based on a minimum number of parameters robustly fitted from the cumulative data easily accessible for all countries at the John Hopkins University database to extrapolate the death rate [22]. Bello-Chavolla *et al.* proposed a clinical score to evaluate the risk for complications and lethality attributable to COVID-19 regarding the effect of obesity and diabetes in Mexico [23]. The results provided a tool for quick determination of susceptibility patients in a first contact scenario. Wang *et al.* leveraged patient data in real-time and devise a patient information based algorithm to estimate and predict the death rate caused by COVID-19 for the near future [24]. Aiewsakun *et al.* performed a genome-wide association study on the genomes of COVID-19 to identify genetic variations that might be associated with the COVID-19 severity [25]. Moreover, Jiang *et al.* established an artificial intelligence framework for data-driven prediction of coronavirus clinical severity [26]. Saha *et al.* proposed a deep learning framework to identify an unknown viral sequence, using Long Short Term Memory (LSTM) [27]. Lopez-Rincon *et al.* designed a classification and specific primer for accurate detection of SARS-CoV-2 with convolutional neural networks [28]. The development of computational and physics-based approaches has relieved the labors of experiments by utilizing epidemiological and biological data to construct the model. However, direct evaluation of potential novel coronavirus strains for their lethality is crucial when clinicians are forced to make difficult decisions without past specific experience to guide clinical acumen. Inferring the lethality of novel coronavirus is possible by identifying the patterns from a large number of coronavirus sequences.

In this paper, we propose an alignment-free framework that can leverage machine learning approaches to infer the lethality of human-adapted coronavirus. The main contribution is that we formulate the problem of estimating the lethality of human-adapted coronavirus through machine learning approaches based on genomic nucleotide, which could assist biologists or virologists for the investigation of coronavirus with new insights. By appropriate feature transformation, we can encode genomic nucleotides into numbers that allow us to convert it into a prediction task. The experimental results suggest our models deliver accurate prediction of lethality without prior biological knowledge. We also provide preliminary analysis validating the effectiveness of our models through other human coronaviruses.

2. METHODS

2.1. Problem Formulation

The pandemic of novel coronaviruses has caused thousands of fatalities, making tremendous treats to public health worldwide. The society is deeply concerned about its spread and evolution with the emergence of any potential new variants that would increase the lethality. Typically, lethality refers to the capability of causing death. It is usually estimated as the cumulative number of deaths divided by the total number of confirmed cases. Among all the human-adapted coronaviruses, MERS-CoV caused the highest fatality rate of 34.5% [29], followed by the SARS-CoV with 9.2% fatality rate [30] In comparison, COVID-19 indicates a lower mortality rate of 5.5% [13]. The lethality rate of COVID-19 is likely to decrease with better treatment and precautions. In this paper, we mainly focus on these three types of human-adapted coronaviruses and define the degree of viral in terms of historical fatality rates. As a result, MERS-CoV strains are high lethal, while SARS-CoV and SARS-CoV-2 strains are middle and low lethal, respectively.

2.2. Data Collection and Preprocessing

Genomic nucleotide sequences of three different coronaviruses with the human host are downloaded from the National Center for Biotechnology Information on April 30, 2020 [31]. Duplicate sequences and incomplete genomes with a length smaller than 20000 are removed from the collection to address the possible issues raised from sequence length bias. Some SARS-CoV strains from the laboratory are included that are cultivated in Vero cell cultures to enrich the training samples. Finally, we end up with 321, 351, 1638 samples for MERS-CoV, SARS-CoV and SARS-CoV--2, respectively. In addition, we also collect the genomic data of other four human coronaviruses with 27, 64, 32 and 142 distinct strains for HCoV-HKU1, HCoV-NL63, HCoV--229E and HCoV-OC43, respectively. Apart from the four symbolic bases (A, C, T, G) of each strain, we have degenerated base symbols that are an IUPAC representation [32] for a position on genomic sequences, which could denote multiple possible alternatives. These degenerate base symbols contain W (A, T), S (C, G), M (A, C), K (G, T), R (A, G), Y (C, T), B (C, G, T), D (A, G, T), H (A, C, T), V (A, C, G), N(A, C, T, G), where the letters in the bracket are alternative nucleotide representing degenerate bases. We randomly substitute these degenerate bases so that genomic sequences can be mapped into discrete numerical representation through feature transformation.

2.3. Feature Transformation

Numerical representations have been successfully employed in the field of bioinformatics [33, 34], mapping biological sequences into real-value vector space where the information or pattern characteristic of the sequence is kept in order. This is important as the existing machine learning approaches can only deal with vectors but not sequence samples. Several methods are proposed that convert genomic sequences into numerical vectors, *e.g.,* the fixed mapping between nucleotides and real numbers without biological significance [35], based on physio-chemical properties [36], deduction from doublets or codons [37], and chaos game representation [38]. To accommodate comprehensive analysis and comparison, we adapt different types of numerical representations for biological RNA sequences. Randhawa *et al.* [39] showed that "Real", "Just-A" and "Purine/Pyrimidine (PP)" numerical representation yield better performance over other methods for DNA classification, which are included for analyzing genomic data. The electro-ion interaction potential (EIIP) and nearest-neighbor based doublet are incorporated that are based on physio-chemical properties and nearest-neighbor values, respectively. Apart from the aforementioned one-dimensional representation, we have introduced 2D Chaos Game Representation (CGR) into feature transformation of original sequences.

The real number representation is a fixed transformation technique through which we obtain values of four bases as: adenine (A) = -1.5, thymine (T) = 1.5, cytosine (C) = 0.5, and guanine (G) = -0.5 [40]. It is efficient in finding a complementary strand of DNA/RNA sequence and can endure complementary property. "Just-A" method maps the four bases into binary classification as the presence of adenine is labeled 1, while others are 0 [41]. PP representation is a DNA-Walk model that shows nucleotides sequences in which a step is taken upwards if the nucleotide is pyrimidine with T/C = 1, or downward if it is purine with A/G = -1 [42]. EIIP describes the distribution of the energy of free elections along with nucleotide sequences that a single EIIP indicator sequence is formed through replacing its nucleotides, where A=0.1260, C=0.1340, G=0.0806, and T=0.1335 [43]. The sequence-to-signal mapping for nearest-neighbor based doublet representation is illustrated in another study [37], where the last position is followed by the first in the sequence. Lastly, CGR is a method proposed by Jeffrey [44] that has been successfully used for a visual representation of genome sequence patterns and taxonomic classification [45, 46]. The CGR images of RNA/DNA sequences are drawn in a unit square. The four vertices of the square are labeled by four nucleotides. The first nucleotide of the sequence is plotted halfway between the center of the square and the vertex representing this nucleotide. The next base is mapped into the image that the coordinate is assigned halfway between the previous point and the vertex corresponding to the previous nucleotide. The mathematical formulation of the successive points that calculates the coordinates in the CGR of the sequences is described below:

$$\begin{aligned} X_i &= 0.5 * (X_{i-1} + C_{ix}) \\ Y_i &= 0.5 * (Y_{i-1} + C_{iy}) \end{aligned} \quad (1)$$

where C_{ix} and C_{iy} denote the X and Y coordinates of the vertices matching the nucleotide at position i of the sequence, respectively.

2.4. Model Construction

Machine learning has been utilized in many aspects of viral genomic analysis, *e.g.,* antigenicity prediction of viruses [20], genome classification of novel pathogens [46], reassortment detection [47], receptor binding analysis [48] and vaccine recommendation [49], *etc.* With increasingly available genomic sequences, it will play more critical roles in helping biologists analyze large, complex biological data for prediction and discovery. In this work, we provide a comprehensive analysis for the lethality prediction of potential new human-adapted coronaviruses *via* alignment-free machine learning approaches. We follow a similar protocol in the studies [39] and [50], adapting it to fit our setting as follows. We utilize the retrospective way to train and test the model since the isolation time of viral strains are available. For each type of coronavirus, the samples isolated from earlier times are used to train the model, while those generated in recent times are for testing. The time threshold will be determined based on the condition that divides training and testing set in a rough 0.8:0.2 ratio. The retrospective test enables our models to infer the lethality of coronavirus for any new strains that could emerge in the near future. Six different types of numerical representations are implemented in comparison with the predictive performance of machine learning models. The proposed methods not only contain traditional machine learning models but also deep learning techniques in combination with Discrete Fourier Transform for genome analysis. Traditional machine learning models consist of logistic regression (LR), random forest (RF), K-nearest neighbor (KNN) and neural network (NN) [51], while three variants of convolutional neural networks (CNN) are leveraged. The CNN models contain AlexNet [52], VGG [53] and ResNet [54].

Logistic regression is a supervised machine learning algorithm that is a linear regression but for classification problems. Random forest is an ensemble learning method that operates by constructing a multitude of decision trees for classification and regression tasks. k-nearest neighbors algorithm is a non-parametric classification method where the function is only approximated locally that the object being assigned to the class most common among its k nearest neighbors. Neural networks are computing systems with interconnected nodes that work much like neurons in the human brain, which can recognize hidden patterns and correlations in raw data, cluster and classify it, and over time continuously learn and improve. Convolutional neural networks are very similar to ordinary neural network made up of neurons that have learnable weights and biases. The major difference is that CNN architectures allow us to encode certain properties into the architecture, which makes the forward

function more efficient to implement and significantly reduce the amount of parameters in the network.

Following the choices of five one-dimensional numerical representation for viral sequences, digital signal processing is introduced through Discrete Fourier Transform (DFT) techniques. We assume that the number of input sequence is n and all the sequences have the same length l sequence S_i = (S_i (0), S_i (2),..., S_i (l - 1)), where $1 \leq i \leq n$, S_i (k) ϵ {A, C, T, G} and $0 \leq k \leq l$ - 1, the corresponding discrete numerical representation is formulated as:

$$N_i = (f(S_i(0)), f(S_i(2)), \ldots, f(S_i(l-1))) \quad (2)$$

where f (S_i (k)) denotes the numerical value after mapping by function f(.) at the position k of nucleotide sequence S_i. The signal N_i computed after DFT is represented as vector F_i. The formulation of F_i is presented below. We define that the magnitude vector that corresponds to the signal N_i as M_i, M_i is the absolute value of F_i.

$$F_i(k) = \sum_{j=0}^{l-1} f\left(S_i(j)\right) \cdot e^{\left(-\frac{2\pi i k j}{l}\right)} \quad (3)$$

Typically, the length of numerical digital signal N_i is equal to the magnitude spectrum M_i that is originated from the length of the genomic sequence. However, the input genome sequences are in different lengths; thus they need to be length-normalized after DFT. Median length-normalization is leveraged for the input digital signals using zero padding. We employ anti-symmetric padding that begins from the last position if the input sequences are shorter than the median length, these short signals are extended to the median length with zero-padding, while the longer sequences are truncated after the median length.

As for the two-dimensional numerical representation, *i.e.,* CGR, a point that corresponds to a sequence of length$l$ will be contained within a square with a side of length 2^{-l}. We assume a square CGR image is generated with a size of 2^k x 2^k matrix, where k is the parameter that determines the size of the image. The frequency of occurrence of any oligomer in a sequence can be obtained by partitioning the CGR space into small squares. Therefore, the number of CGR points in each unit square of 2^k x 2^k grid is equal to the number of occurrences of all possible k-mers in the sequence. By counting the frequency of CGR points, it is possible to calculate oligonucleotide frequencies at various grid resolutions. We define the element a^j as the number of points that are located in the corresponding sub-square j, where $1 \leq j \leq 2^{2k}$. Each sequence will be mapped into a 2^k x 2^k dimensional vector space based on CGR.

2.5. Implementation and Evaluation

We implement all the models by Scikit-learn [55] and PyTorch [56]. We utilize the retrospective method to train and test the model since the isolation time of viral strains is available. For each type of coronavirus, the samples generated from strains isolated before the year N are used to build the model, while those generated after the year N are for testing. The year N is determined based on the condition that divides training and testing set in a rough 0.8:0.2 ratio. The 5-fold cross-validation is performed in the training process and the independent testing set is used for validation of our models. This test can truly reflect the ability of the models in applications to predict viral lethality for future strains. The parameters are set by default with traditional machine learning models (Supplementary Materials **S1**). For all deep learning-based models, we apply stochastic gradient descent with a minimum batch size of 64 for optimization. The drop-out (rate = 0.5) strategy is carried out with a 0.001 learning rate and all the models are fit for 50 training epochs. The predictive performance is evaluated by accuracy, precision, sensitivity, and F1 score of all models in the prediction tasks of coronavirus lethality.

3. RESULTS

3.1. Genome Composition of SARS-CoV, MERS-CoV and SARS-CoV-2

We first analyzed the composition of the RNA genome of the three human-adapted coronaviruses. Fig. (**1**) portrays the average distribution and variance of the nucleotides. We can observe that the proportion of A and T (in replacement of U) is high, while C and G are relatively low for all human coronaviruses. Interestingly, it is suggested that the high T and low C proportions of human coronaviruses are quite variable and act like communicating vessels. T goes up when C decreases and *vice versa*. The composition of T ranges from 0.139 to 0.552 while C makes the opposite movement from 0.374 to 0.107, respectively, among all human coronavirus. If we look into individual types, the SARS-CoV-2 as a novel human pathogen follows some typical composition of nucleotides but it is also characterized by some differences. We found that SARS-CoV-2 presents a higher variance compared with MERS-CoV and SARS-CoV. This is probably the rapid and widespread transmission of SARS-CoV-2 accelerates its evolution when infected with humans. More strains are generated differently from their ancestor clade. However, it is more pronounced of the nucleotide bias in the unpaired regions of the structured RNA genome, which may indicate a certain biological function of these special sequence signatures. Some studies have revealed that a clear difference in the magnitude of the nucleotide bias of the coronavirus genomes is likely to relate to the mechanism of subgenomic mRNA synthesis and the exposure of single-stranded RNA domains [57, 58]. The evidence shows that cytosine discrimination and deamination against CpG dinucleotides are the driving force that outlines the coronaviruses over evolutionary times [59]. It is indicated that the atypical nucleotide bias could reflect distinct biological functions that are the direct cause of the characteristic codon usage in these viruses [60]. Therefore, the analysis of the nucleotide and codon usage in coronaviruses can not only exhibit the clues on potential viral evolution but also improves the

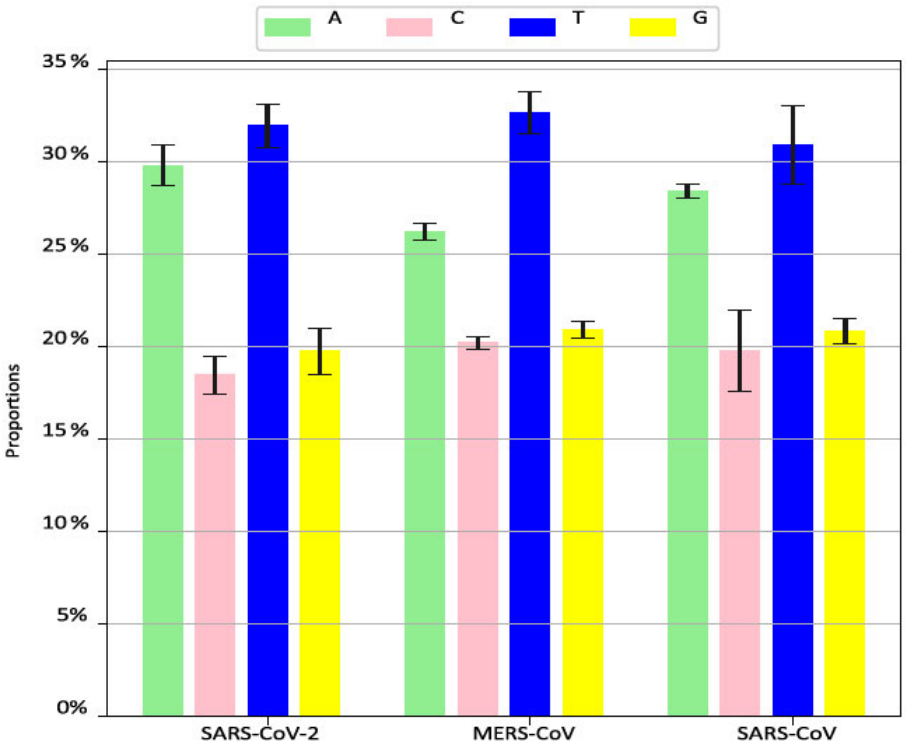

Fig. (1). Composition of nucleotide sequence for SARS-CoV, MERS-CoV and SARS-CoV-2. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

understanding of the viral regulation and promotes vaccine design.

3.2. K-mer-based Classifier Study

Studies have investigated the role of *k*-mer frequency for the fast and accurate classification of viral genomes [61]. We experiment with values for *k*-mer of length $k = 1, 2, ..., 7$ on different classifiers to measure the prediction accuracy. To explore the *k*-mer frequency patterns in distinct coronaviruses, we curate independent testing set to assess the performance. Fig. (**2**) shows the predictive accuracy across seven machine learning algorithms, at different values of *k*. Fig. (**2**) portrays the performance curves by deep learning models (in the left), and the results *via* traditional machine learning models (in the right). Overall, our proposed methods obtain an average accuracy of 0.956. However, we can observe the traditional machine learning methods exceed 0.98 in accuracy at all *k* values, whereas there is a different story for deep learning models. It is shown that VGG-19 achieves the best results, while the accuracy could be as low as 0.8 using ResNet-34 when $k = 2$. We can conclude that for these data the traditional machine learning methods outperform deep learning models almost at all levels of *k* with less fluctuation. As a result, the *k*-mer value 6 is used for the results of experiments with CGR representation.

3.3. Comparative Performance

We analyzed the effect on viral lethality prediction *via* different numerical representations for RNA sequences using machine learning approaches. The dataset used is the same as those in Fig. (**2**). The results along with the average scores for all numerical representations and classifiers are summarized in Table **1**. As can be observed from Table **1**, for all numerical representations, the average scores are high over all measures. The best performance is achieved when using CGR representation, which yields an average accuracy of 0.985 in the testing set. Surprisingly, we can obtain an average accuracy of 0.967 even with a single nucleotide numerical representation "Just-A". At the individual classifier level, traditional machine learning methods display an apparent advantage over deep learning models. Logistic regression and neural network can achieve 100% accuracy for all numerical representations, whereas the prediction accuracy ranges from 0.679 to 0.993 implemented by Resnet34, VGG19 and AlexNet. At this point, this is probably because deep learning algorithms need a large amount of data to understand the pattern. In addition to performing higher accuracy, machine learning models are computationally cheaper in this task, *e.g.* in CGR representation, it takes much longer time for deep learning models than classical

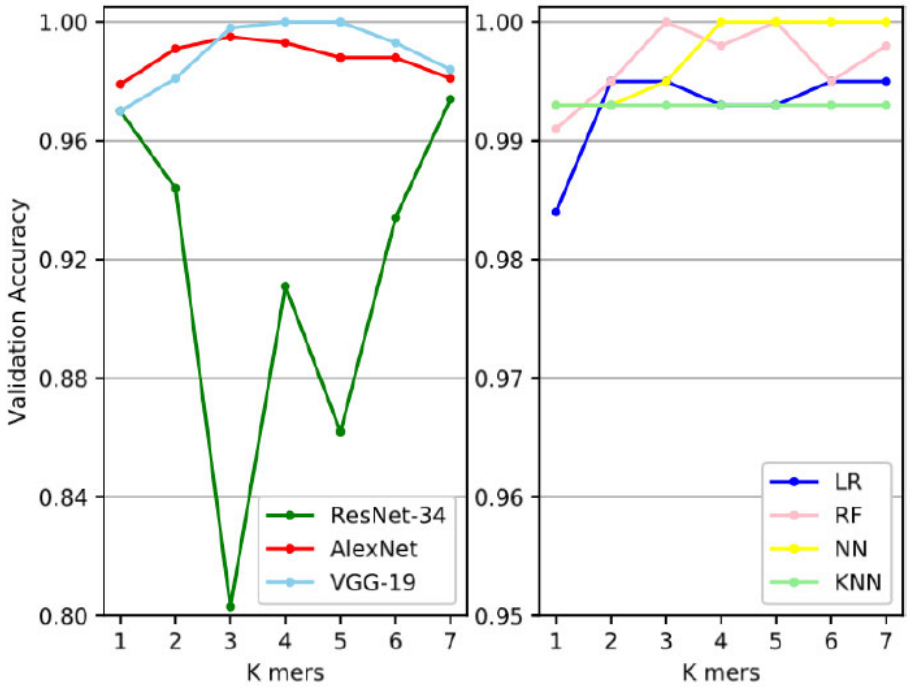

Fig. (2). The prediction accuracy across seven machine learning classifiers at different values of *k* (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

machine learning methods on average (Supplementary Materials **S2**). Overall, our results suggest that all these numerical representations are effective for modeling to differentiate the degree of the lethality of human coronavirus.

3.4. Validation on other Human Coronaviruses

We test the ability of our models to identify the lethality of other different human coronaviruses, *i.e.,* HCoV-229E, HCoV-NL63, HCoV-OC43, and HCoV-HKU1. The training process is implemented on the former three types of coronavirus data. For every test dataset, we use CGR as the numerical representation with all classifiers to predict the lethality. Interestingly, the results show that, on average, 28 out of 32 HCoV-229E, 59 out of 64 HCoV-NL63, 134 out of 142 HCoV-OC43 and 25 out of 27 HCoV-HKU1 strains identified have closer lethality with SARS-CoV-2, while the rest strains are labeled middle or high. This suggests that, overall, other test human coronaviruses have lower severity than MERS-CoV and SARS-CoV. Evidence has revealed that HCoV-OC43 and HCoV-HKU1 are associated with mild to moderate upper respiratory tract illness with about 0.1% fatality [62]. These infections may be asymptomatic and are considered the second common cause of cold [63]. Similarly, it has been well documented that the majority of HCoV--NL63 infections are mild in humans, though occasionally, this coronavirus causes pneumonia or central nervous system diseases in susceptible individuals [64]. During 2009 and 2016, it accounted for about 0.5% of all acute respiratory tract infections in hospitalized patients from Guangzhou, China, but few death cases are reported [65]. HCoV-229E is a close relative of HCoV-NL63 and it will lead to similar symptoms [66].

Fig. (**3**) displays the CGR plots of different sequences of human coronavirus at the value of 6 for k-mer frequency. The CGR plots visually indicate that the genomic signature of the SARS-CoV-2 isolate Wuhan-Hu-1 (Fig. **3c** is closer to the genomic signature of the SARS-CoV coronavirus isolate Canada (Fig. **3a**, followed by the strain of MERS-CoV Betacoronavirus England 1 isolate (Fig. **3b**. Moreover, the other four human coronaviruses from (Fig. **3d, e, f** and **g**) presents similar visual patterns, which are different from the former three types. Given the CGR plots of human coronaviruses, we further explore the trace of their origin and relation through phylogenetic analysis. We randomly select five complete genomes from each type containing the reference strain. The phylogenetic tree is constructed based on all pairwise distance with maximum likelihood techniques for the dataset. The results in Fig. (**4**) present a clear separation of seven clusters and relationships within the clusters. The average inter-cluster distances confirm that SARS-CoV-2 sequences are closest to the species of SARS-CoV (average distance 0.486), followed by MERS-CoV (4.782), which are far away from other four human coronaviruses. We also find that HCoV-OC43 and HCoV-HKU1, HCoV-229E and HCoV-NL63 may originate from the same ancestor with the genetic distance 1.842 and 2.779, respectively. But there is no evidence indicating the situation that the two different species of human coronavirus will present similar lethality if they are genetically close.

Table 1. The performance for the lethality prediction of human-adapted coronaviruses *via* seven different classifiers. Average results for each numerical representation are in bold.

Numerical Representation	Model	Training Data			-	Testing Data			
		Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
Real	LR	0.999	0.999	1.000	0.999	1.000	1.000	1.000	1.000
	KNN	0.999	1.000	0.999	0.999	0.984	0.994	0.983	0.988
	NN	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	RF	0.998	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	ResNet34	0.964	0.990	0.990	0.990	0.979	0.992	0.986	0.989
	VGG19	0.961	0.989	0.989	0.989	0.981	0.988	0.988	0.988
	AlexNet	0.671	0.841	0.841	0.841	0.679	0.893	0.670	0.765
	Average	**0.941**	**0.973**	**0.974**	**0.973**	**0.946**	**0.981**	**0.946**	**0.961**
Nearest neighbor based doublet	LR	0.999	0.999	1.000	0.999	1.000	1.000	1.000	1.000
	KNN	0.998	0.999	0.996	0.998	0.981	0.993	0.981	0.987
	NN	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	RF	0.998	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	ResNet34	0.966	0.991	0.991	0.991	0.977	0.991	0.988	0.989
	VGG19	0.946	0.981	0.981	0.981	0.967	0.987	0.984	0.986
	AlexNet	0.857	0.936	0.936	0.936	0.714	0.902	0.712	0.796
	Average	**0.966**	**0.986**	**0.986**	**0.986**	**0.948**	**0.981**	**0.952**	**0.964**
EIIP	LR	0.999	0.999	1.000	0.999	1.000	1.000	1.000	1.000
	KNN	0.998	0.999	0.995	0.997	0.981	0.993	0.981	0.987
	NN	0.998	0.997	0.999	0.998	1.000	1.000	1.000	1.000
	RF	0.999	0.997	0.998	0.997	1.000	1.000	1.000	1.000
	ResNet34	0.962	0.989	0.989	0.989	0.972	0.989	0.980	0.984
	VGG19	0.940	0.978	0.978	0.978	0.979	0.992	0.989	0.990
	AlexNet	0.839	0.927	0.927	0.927	0.848	0.949	0.936	0.942
	Average	**0.962**	**0.983**	**0.983**	**0.983**	**0.968**	**0.989**	**0.983**	**0.986**
PP	LR	0.999	0.999	1.000	0.999	0.995	0.998	0.995	0.997
	KNN	0.999	1.000	0.998	0.999	0.981	0.993	0.981	0.987
	NN	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	RF	0.999	0.999	0.998	0.999	0.998	0.999	0.998	0.998
	ResNet34	0.963	0.989	0.989	0.989	0.977	0.991	0.985	0.988
	VGG19	0.943	0.980	0.980	0.980	0.993	0.997	0.994	0.996
	AlexNet	0.662	0.837	0.837	0.837	0.681	0.894	0.669	0.765
	Average	**0.937**	**0.971**	**0.971**	**0.971**	**0.946**	**0.981**	**0.946**	**0.961**
Just-A	LR	0.999	0.999	1.000	0.999	1.000	1.000	1.000	1.000
	KNN	0.998	0.999	0.996	0.998	0.986	0.994	0.985	0.990
	NN	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	RF	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	ResNet34	0.969	0.992	0.992	0.992	0.960	0.984	0.969	0.977
	VGG19	0.969	0.992	0.992	0.992	0.984	0.989	0.992	0.991
	AlexNet	0.842	0.928	0.928	0.928	0.841	0.942	0.933	0.938
	Average	**0.967**	**0.986**	**0.986**	**0.986**	**0.967**	**0.987**	**0.982**	**0.985**
CGR	LR	1.000	1.000	1.000	1.000	0.995	0.998	0.995	0.997
	KNN	0.999	1.000	0.999	0.999	0.993	0.997	0.993	0.995
	NN	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000
	RF	0.999	0.997	0.998	0.999	0.995	0.998	0.995	0.997
	ResNet34	0.975	0.996	0.996	0.996	0.934	0.975	0.933	0.954
	VGG19	0.948	0.982	0.982	0.982	0.993	0.997	0.994	0.996
	AlexNet	0.955	0.986	0.986	0.986	0.988	0.995	0.992	0.994
	Average	**0.982**	**0.994**	**0.994**	**0.994**	**0.985**	**0.994**	**0.986**	**0.990**

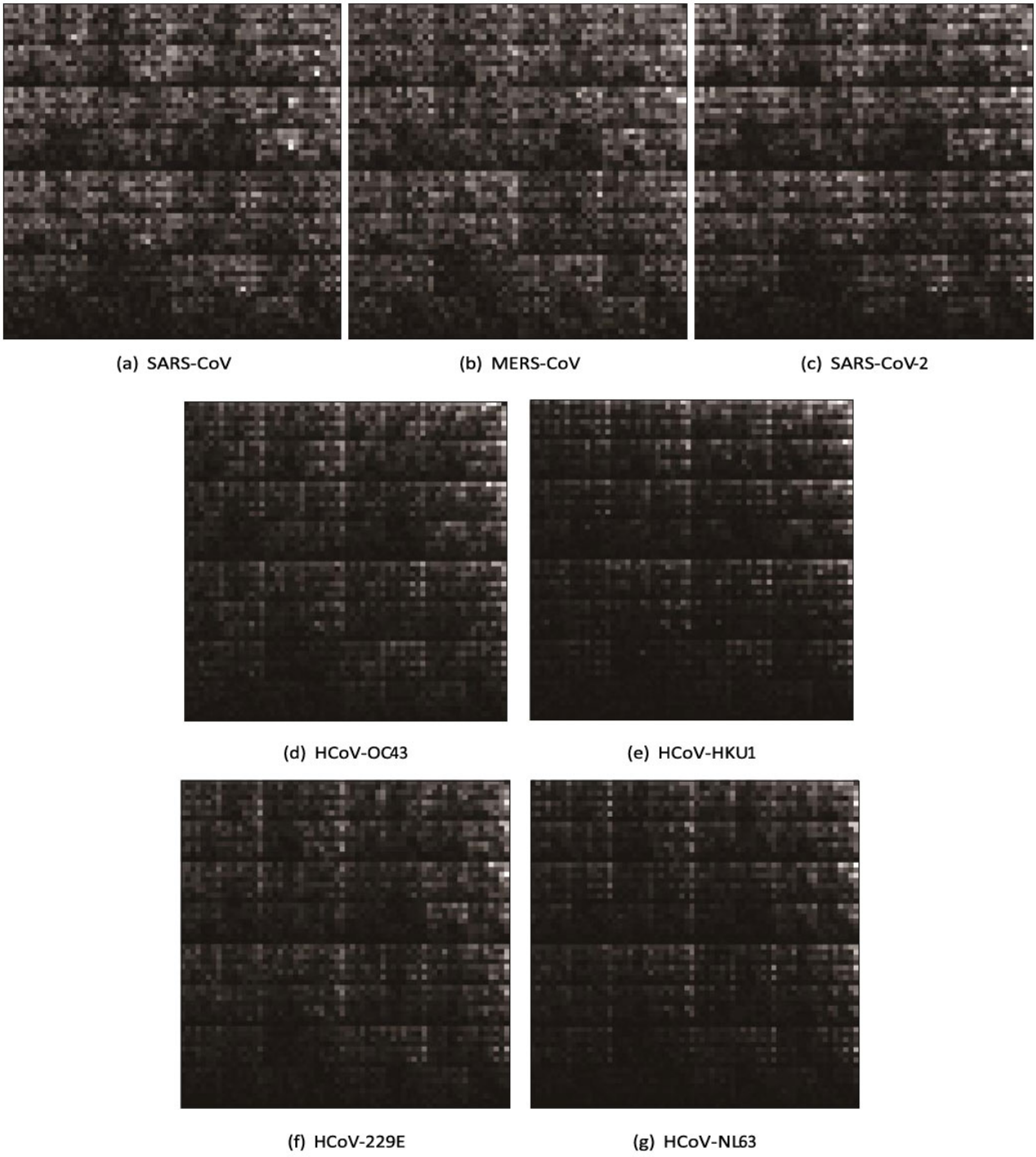

Fig. (3). The visualization of Chaos Game Representation plots of references of seven human coronavirus at k = 6 of (**a**) SARS-CoV-/NC_004718.3/Severe acute respiratory syndrome-related coronavirus/Canada, (**b**) MERS-CoV/NC_038294.1/Betacoronavirus England 1, Middle East respiratory syndrome-related coronavirus/United Kingdom, (**c**) SARS-CoV-2/NC_045512.2/Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1/China, (**d**) HCoV-OC43/JN129835.1/Human coronavirus OC43 strain HK04-02, China/Betacoronavirus 1, (**e**) HCoV-HKU1/NC_006577.2/Human coronavirus HKU1, (**f**) HCoV-229E/JX503060.1/Human coronavirus 229E isolate 0349, Netherlands, (**g**) HCoV-NL63/JQ765575.1/Human coronavirus NL63 strain NL63/DEN/2005/1876, USA. The vertices of the plot are assigned A (top left), T (top right), C (bottom left), G (bottom right). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

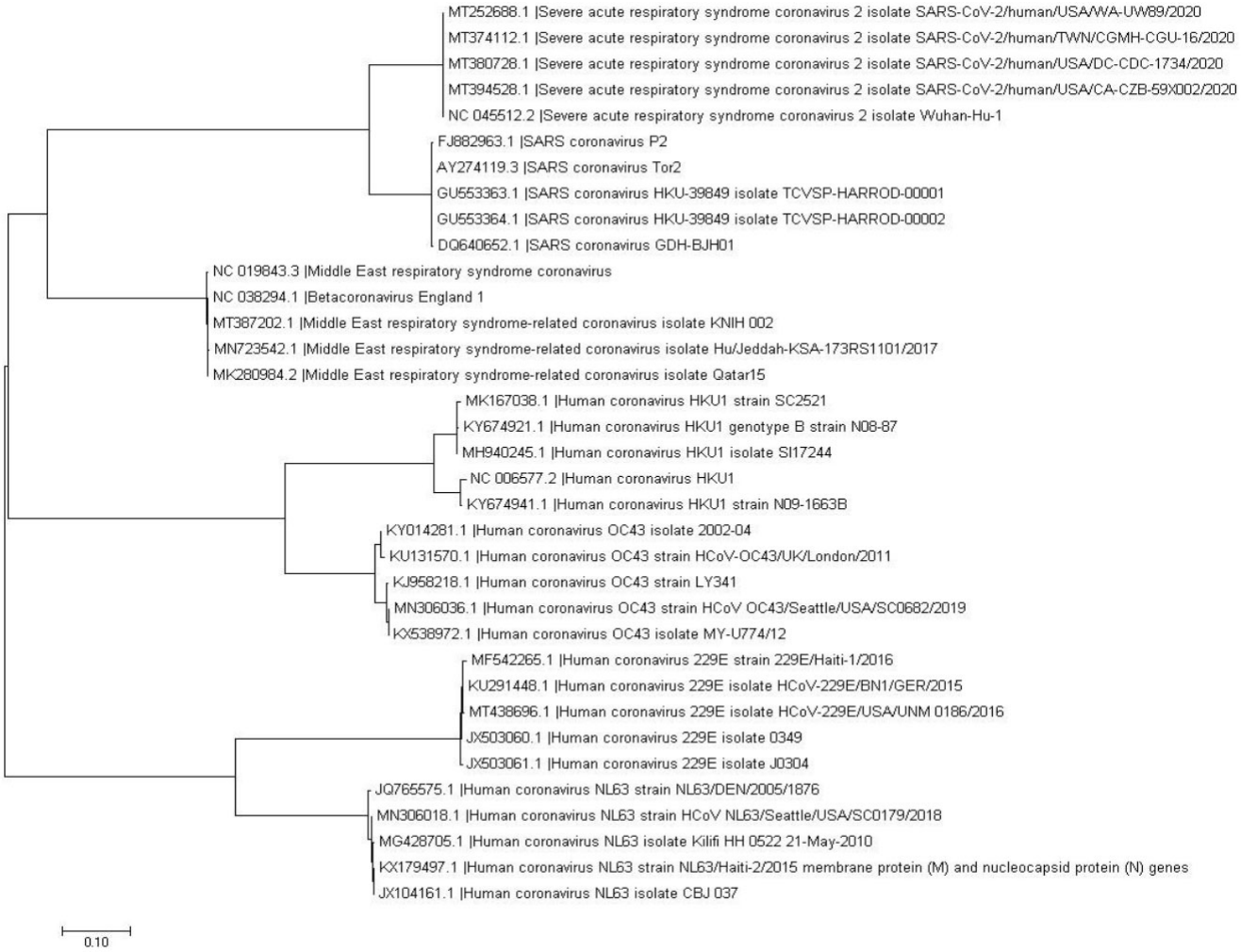

Fig. (4). The phylogenetic tree using maximum likelihood generated pairwise distance matrix shows the hierarchical classification between different human coronaviruses.

4. DISCUSSION

We provided a comprehensive quantitative analysis to predict the lethality of human-adapted coronavirus with six different numerical representations of RNA sequences applied in machine learning models. The models are computational efficiency because they are alignment-free. Compared with alignment-based methods, multiple sequence alignment is not needed with the leverage of DFT techniques. The experiment results show that most of the models have achieved rapid and accurate predictions for the lethality of new human-adapted coronavirus. We validated our results by a quantitative analysis based on the construction of the phylogenetic tree, which reveals the evolutionary relationships among all human coronaviruses based upon genetic information. Coronaviruses are usually thought to cause mild and non-lethal symptoms in humans before the outbreak of SARS-CoV in 2003. The high pathogenicity of SARS-CoV, MERS-CoV and newly SARS-CoV-2 captures surgent interests and concerns of the family of coronavirus. Timely analysis of genomic sequences of novel strains requires quick sequence similarity comparison with thousands of known species, which are generally performed by alignment-based methods. However, these methods are time-consuming and sometimes challenging in cases where homologous sequence continuity cannot be ensured. The application of alignment-free approaches has addressed this issue that can handle a large number of sequences effectively.

Previous studies have elucidated that the origin of this SARS-CoV-2 stems from bats [10, 67]. Early sequencing of SARS-CoV-2 strains revealed over 99% similarity with some bat-like coronavirus, indicating these infections result from a recent cross-species event [68]. Bats are regarded as the natural reservoir of viruses and cross-species transmission to mammals [4, 69]. Before the emergence of SARS-

CoV-2, it was uncovered that the coronavirus SARS-Cov and MERS-CoV have also originated from bats [70, 71]. The phylogenetic analyses assist in identifying the relationships between SARS-CoV-2 and other coronaviruses through the nucleotide and amino acid sequence similarities. The continuous human-to-human transmission has been confirmed and asymptomatic cases have continued to increase [72, 73]. There is a desperate need for strict precautions to prevent the spread of the virus and protect public health. Vaccines and miracle drugs are the most efficient ways of fighting against this crisis. Currently, the development of vaccines has been into Phase 3 trials in some countries, while the human ACE2 receptor has been identified as the potential receptor for COVID-19 and serves as a potential target for treatment [74, 75]. Nevertheless, with the circulation of bat-related coronavirus and geographic coverage, it is critical to monitor the evolution of coronavirus. Currently, seven known types of coronavirus can infect humans. Novel strains of these coronaviruses can likely arise and attack human again through reassortment and mutation when two different or more strains co-infect the same host. Preparation is necessary to prevent potential epidemics and pandemics caused by a novel coronavirus. As a result, our work paves the basis for surveillance by inferring the lethality of any potential human coronaviruses that may emerge in the future.

This study is subject to a variety of limitations. The definition of classifying the degree of coronavirus lethality is mainly based on the mortality rate. We assume that the higher the mortality, the more lethal for the virus, and thus make three categories of the lethality level for all viruses with a different threshold. However, our estimation for these values lies within the range of fatality rate from the literature, which we do not have sufficient data to obtain and parameterize the case-structured model, especially for viruses with few samples. Besides, some other factors such as innate immune system and comorbidities could make a significant impact on the lethality of virus when infecting humans. Moreover, the limited data points for the human coronavirus pale the high predictive accuracy, as most of the machine learning algorithms possess a superb generation ability to discover inherent patterns from training samples, particularly in the small dataset. But like typical machine learning approaches, our models are not qualified to provide a direct and accessible explanation that explicitly interprets why a certain coronavirus strain is more lethal to humans. Some rule-based methods or clinical study might provide a better rationale for their results.

CONCLUSION

We provide a comprehensive analysis through alignment-free machine learning-based methods for the prediction of the lethality of existing human-adapted coronavirus. The results show that on average, CGR, EIIP, and Just-A representations perform better than others, with an average accuracy of 0.985, 0.968 and 0.963, respectively. Interestingly, traditional machine learning methods display obvious merit both in computational efficiency and performance than deep learning models on this task. Validation of other types of human coronavirus in combination with phylogenetic analysis further demonstrates our predictive results. We hope this work would facilitate the research of COVID-19 for biologists and clinicians that are in the frontline to detect the lethality of new emerging variants of SARS-CoV-2. Future work includes the construction of novel coronavirus surveillance and *in vitro* evaluation of the computational models.

ETHICALS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The source code, data and supplementary can be found at: https://github.com/Rayin-saber/Alignment-free-lethality-predictionof-coronavirus.

FUNDING

This project is supported by AcRF Tier 2 grant MOE2014-T2-2-023, Ministry of Education, Singapore and A*STAR-NTU-SUTD AI Partnership grant, RGANS1905.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Li, G.; Fan, Y.; Lai, Y.; Han, T.; Li, Z.; Zhou, P.; Pan, P.; Wang, W.; Hu, D.; Liu, X.; Zhang, Q.; Wu, J. Coronavirus infections and immune responses. *J. Med. Virol.,* **2020**, *92*(4), 424-432. http://dx.doi.org/10.1002/jmv.25685 PMID: 31981224

[2] Wertheim, J.O.; Chu, D.K.; Peiris, J.S.; Kosakovsky Pond, S.L.; Poon, L.L. A case for the ancient origin of coronaviruses. *J. Virol.,* **2013**, *87*(12), 7039-7045. http://dx.doi.org/10.1128/JVI.03273-12 PMID: 23596293

[3] Vijaykrishna, D.; Smith, G.J.; Zhang, J.X.; Peiris, J.S.; Chen, H.; Guan, Y. Evolutionary insights into the ecology of coronaviruses. *J. Virol.,* **2007**, *81*(8), 4012-4020. http://dx.doi.org/10.1128/JVI.02605-06 PMID: 17267506

[4] Cui, J.; Li, F.; Shi, Z-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.,* **2019**, *17*(3), 181-192. http://dx.doi.org/10.1038/s41579-018-0118-9 PMID: 30531947

[5] Woo, P.C.Y.; Lau, S.K.P; Huang, Y.; Yuen, K.-Y. Coronavirus diversity, phylogeny and interspecies jumping. *Exp. Biol. Med.,* **2009**, *234*(10), 1117-1127. http://dx.doi.org/10.3181/0903-MR-94

[6] Kahn, J.S.; McIntosh, K. History and recent advances in coronavirus discovery. *Pediatr. Infect. Dis. J.,* **2005**, *24*(11)(Suppl.), S223-S227. http://dx.doi.org/10.1097/01.inf.0000188166.17324.60 PMID: 16378050

[7] Victor, M.; Muth, D.; Niemeyer; Drosten, C. Hosts and sources of

endemic human coronaviruses. *Adv. Virus Res.,* **2018**, *100*, 163-188.

[8] Cooke, F.J.; Shapiro, D.S. Global outbreak of severe acute respiratory syndrome (SARS). *Int. J. Infect. Dis.,* **2003**, *7*(2), 80-85. http://dx.doi.org/10.1016/S1201-9712(03)90001-4 PMID: 12839707

[9] Aleanizy, F.S.; Mohmed, N.; Alqahtani, F.Y.; El Hadi Mohamed, R.A. Outbreak of Middle East respiratory syndrome coronavirus in Saudi Arabia: A retrospective study. *BMC Infect. Dis.,* **2017**, *17*(1), 23. http://dx.doi.org/10.1186/s12879-016-2137-3 PMID: 28056850

[10] Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; Bi, Y.; Ma, X.; Zhan, F.; Wang, L.; Hu, T.; Zhou, H.; Hu, Z.; Zhou, W.; Zhao, L.; Chen, J.; Meng, Y.; Wang, J.; Lin, Y.; Yuan, J.; Xie, Z.; Ma, J.; Liu, W.J.; Wang, D.; Xu, W.; Holmes, E.C.; Gao, G.F.; Wu, G.; Chen, W.; Shi, W.; Tan, W. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet,* **2020**, *395*(10224), 565-574. http://dx.doi.org/10.1016/S0140-6736(20)30251-8 PMID: 32007145

[11] Morens, D.M.; Folkers, G.K.; Fauci, A.S. The challenge of emerging and re-emerging infectious diseases. *Nature,* **2004**, *430*(6996), 242-249. http://dx.doi.org/10.1038/nature02759 PMID: 15241422

[12] Willman, M.; Kobasa, D.; Kindrachuk, J. A comparative analysis of factors influencing two outbreaks of middle eastern respiratory syndrome (MERS) in Saudi Arabia and south korea. *Viruses,* **2019**, *11*(12), 1119. http://dx.doi.org/10.3390/v11121119 PMID: 31817037

[13] Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.,* **2020**, *20*(5), 533-534. http://dx.doi.org/10.1016/S1473-3099(20)30120-1 PMID: 32087114

[14] Heymann, D.L.; Shindo, N. WHO Scientific and Technical Advisory Group for Infectious Hazards. COVID-19: What is next for public health? *Lancet,* **2020**, *395*(10224), 542-545. http://dx.doi.org/10.1016/S0140-6736(20)30374-3 PMID: 32061313

[15] Yin, R.; Luusua, E; Dabrowski, J; Zhang, Y; Kwoh, C.K Tempel: Time-series mutation prediction of influenza a viruses *via* attention-based recurrent neural networks. *Bioinformatics,* **2020**, *36*(9), 2697-2704. http://dx.doi.org/10.1093/bioinformatics/btaa050 PMID: 31999330

[16] Morens, D.M.; Daszak, P.; Taubenberger, J.K. Escaping pandora's box-another novel coronavirus. *N. Engl. J. Med.,* **2020**, *382*(14), 1293-1295. http://dx.doi.org/10.1056/NEJMp2002106 PMID: 32101660

[17] Peeri, N.C.; Shrestha, N; Rahman, Md.S; Zaki, R; Tan, Z; Bibi, S; Baghbanzadeh, M; Aghamohammadi, N; Zhang, W; Haque, U The SARS, MERS and novel coronavirus (covid-19) epidemics, the newest and biggest global health threats: What lessons have we learned? *Int. J. Epidemiol.,* **2020**, *40*(3), 717-726.

[18] Li, Z.; Yi, Y.; Luo, X.; Xiong, N.; Liu, Y.; Li, S.; Sun, R.; Wang, Y.; Hu, B.; Chen, W.; Zhang, Y.; Wang, J.; Huang, B.; Lin, Y.; Yang, J.; Cai, W.; Wang, X.; Cheng, J.; Chen, Z.; Sun, K.; Pan, W.; Zhan, Z.; Chen, L.; Ye, F. Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. *J. Med. Virol.,* **2020**, *92*(9), 1518-1524. http://dx.doi.org/10.1002/jmv.25727 PMID: 32104917

[19] Cao, Y.; Li, L.; Feng, Z.; Wan, S.; Huang, P.; Sun, X.; Wen, F.; Huang, X.; Ning, G.; Wang, W. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.,* **2020**, *6*(1), 11. http://dx.doi.org/10.1038/s41421-020-0147-1 PMID: 32133153

[20] Yin, R.; Tran, V.H.; Zhou, X.; Zheng, J.; Kwoh, C.K. Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model. *PLoS One,* **2018**, *13*(12), e0207777. http://dx.doi.org/10.1371/journal.pone.0207777 PMID: 30576319

[21] Yin, R. **2020**, *Meta-analysis on the lethality of influenza a viruses using machine learning approaches*. Doctoral thesis, Nanyang Technological University, Singapore, **2020**.

[22] Bastolla, U. How lethal is the novel coronavirus, and how many undetected cases there are? The importance of being tested. *medRxiv,* **2020**. http://dx.doi.org/10.1101/2020.03.27.20045062

[23] Bello-Chavolla, O.Y.; Bahena-Lopez, J.P.; Antonio-Villa, N.E.; Vargas-Vázquez, A. ; GonzálezDíaz, A.; Márquez-Salinas, A.; Fermín-Martínez, C.A.; Naveja, J.J.; Aguilar-Salinas, C.A. Predicting mortality due to SARS-COV-2: A mechanistic score relating obesity and diabetes to covid-19 outcomes in Mexico. *medRxiv,* **2020**.

[24] Wang, L.; Li, J.; Guo, S.; Xie, N.; Yao, L.; Cao, Y.; Day, S.W.; Howard, S.C.; Graff, J.C.; Gu, T.; Ji, J.; Gu, W.; Sun, D. Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. *Sci. Total Environ.,* **2020**, *727*, 138394. http://dx.doi.org/10.1016/j.scitotenv.2020.138394 PMID: 32334207

[25] Aiewsakun, P.; Wongtrakoongate, P.; Thawornwattana, Y.; Hongeng, S.; Thitithanyanont, A Sarscov-2 genetic variations associated with covid-19 severity. *medRxiv,* **2020**.

[26] Jiang, X.; Coffee, M.; Bari, A.; Wang, J.; Jiang, X.; Shi, J.; Dai, J.; Cai, J.; Zhang, T.; Wu, Z. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers*. *Comput. Mater. Contin.,* **2020**, *63*(1), 537-551. http://dx.doi.org/10.32604/cmc.2020.010691

[27] Saha, I.; Ghosh, N.; Maity, D.; Seal, A.; Plewczynski, D. COVID-DeepPredictor: Recurrent neural network to predict SARS-CoV-2 and other pathogenic viruses. *Front. Genet.,* **2021**, *12*, 569120. http://dx.doi.org/10.3389/fgene.2021.569120 PMID: 33643375

[28] Lopez-Rincon, A.; Tonda, A.; Mendoza-Maldonado, L.; Mulders, D.G.J.C.; Molenkamp, R.; Perez-Romero, C.A.; Claassen, E.; Garssen, J.; Kraneveld, A.D. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Sci. Rep.,* **2021**, *11*(1), 947. http://dx.doi.org/10.1038/s41598-020-80363-5 PMID: 33441822

[29] World Health Organization et al. Middle east respiratory syndrome coronavirus (MERS-COV). **2019**. Available from: https://www.who.int/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov).

[30] World Health Organization et al. Summary of probable sars cases with onset of illness from 1 November 2002 to 31 july 2003. **2003**. Available from: http://www.who.int/csr/sars/country/table2004_04_21/en/index.html.

[31] Wheeler, D.L.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; Dicuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L.Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D.J.; Madden, T.L.; Maglott, D.R.; Miller, V.; Ostell, J.; Pruitt, K.D.; Schuler, G.D.; Shumway, M.; Sequeira, E.; Sherry, S.T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R.L.; Tatusova, T.A.; Wagner, L.; Yaschenko, E. Database resources of the national center for biotechnology information. *Nucleic Acids Res.,* **2008**, *36*(Suppl. 1), D13-D21. PMID: 18045790

[32] Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucleic Acids Res.,* **1985**, *13*(9), 3021-3030. http://dx.doi.org/10.1093/nar/13.9.3021 PMID: 2582368

[33] Yin, R.; Zhou, X.; Zheng, J.; Kwoh, C.K. Computational identification of physicochemical signatures for host tropism of influenza A virus. *J. Bioinform. Comput. Biol.,* **2018**, *16*(6), 1840023. http://dx.doi.org/10.1142/S0219720018400231 PMID: 30567479

[34] Zeng, M.; Zhang, F.; Wu, F-X.; Li, Y.; Wang, J.; Li, M. Protein-protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics,* **2020**, *36*(4), 1114-1120. PMID: 31593229

[35] Kwan, H.K.; Arniker, S.B. Numerical representation of DNA sequences. In: 2009 IEEE International Conference on Electro/Information Technology, 2009 Jun 7-9; Windsor, ON, Canada, pp. 307-310.

http://dx.doi.org/10.1109/EIT.2009.5189632

[36] Adetiba, E.; Olugbara, O.O.; Taiwo, T.B. Identification of pathogenic viruses using genomic cepstral coefficients with radial basis function neural network. In: *Advances in Nature and Biologically Inspired Computing*; Pillay, N.; Engelbrecht, A.P.; Abraham, A.; du Plessis, M.C.; Snášel, V.; Muda, A.K., Eds.; Springer: Cham, **2016**; pp. 281-291. http://dx.doi.org/10.1007/978-3-319-27400-3_25

[37] Borrayo, E.; Mendizabal-Ruiz, E.G.; Vélez-Pérez, H.; Romo-Vázquez, R.; Mendizabal, A.P.; Morales, J.A. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. *PLoS One,* **2014**, *9*(11), e110954. http://dx.doi.org/10.1371/journal.pone.0110954 PMID: 25393409

[38] Almeida, J.S.; Carriço, J.A.; Maretzek, A.; Noble, P.A.; Fletcher, M. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics,* **2001**, *17*(5), 429-437. http://dx.doi.org/10.1093/bioinformatics/17.5.429 PMID: 11331237

[39] Randhawa, G.S.; Hill, K.A.; Kari, L. Ml-dsp: Machine learning with digital signal processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. *BMC Genom.,* **2019**, *20*(1), 267. http://dx.doi.org/10.1186/s12864-019-5571-y PMID: 30943897

[40] Chakravarthy, N.; Spanias, A.; Iasemidis, L D. Autoregressive modeling and feature analysis of DNA sequences. *EURASIP J. Adv. Signal Process.,* **2004**, *2004*(1), 952689. http://dx.doi.org/10.1155/S111086570430925X

[41] Voss, R.F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.,* **1992**, *68*(25), 3805-3808. http://dx.doi.org/10.1103/PhysRevLett.68.3805 PMID: 10045801

[42] Berger, J.A.; Mitra, S.K; Carli, Marco; Neri, Alessandro Visualization and analysis of DNA sequences using DNA walks. *J. Franklin Inst.,* **2004**, *341*(1-2), 37-53. http://dx.doi.org/10.1016/j.jfranklin.2003.12.002

[43] Lalović, D.; Veljković, V. The global average DNA base composition of coding regions may be determined by the electron-ion interaction potential. *Biosystems,* **1990**, *23*(4), 311-316. http://dx.doi.org/10.1016/0303-2647(90)90013-Q PMID: 2322643

[44] Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.,* **1990**, *18*(8), 2163-2170. http://dx.doi.org/10.1093/nar/18.8.2163 PMID: 2336393

[45] Karamichalis, R.; Kari, L.; Konstantinidis, S.; Kopecki, S.; Solis-Reyes, S. Additive methods for genomic signatures. *BMC Bioinformatics,* **2016**, *17*(1), 313. http://dx.doi.org/10.1186/s12859-016-1157-8 PMID: 27549194

[46] Randhawa, G.S.; Soltysiak, M.P.M.; El Roz, H.; de Souza, C.P.E.; Hill, K.A.; Kari, L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS One,* **2020**, *15*(4), e0232391. http://dx.doi.org/10.1371/journal.pone.0232391 PMID: 32330208

[47] Yin, R.; Zhou, X.; Rashid, S.; Kwoh, C.K. HopPER: An adaptive model for probability estimation of influenza reassortment through host prediction. *BMC Med. Genomics,* **2020**, *13*(1), 9. http://dx.doi.org/10.1186/s12920-019-0656-7 PMID: 31973709

[48] Zhou, X.; Zheng, J.; Ivan, F.X.; Yin, R.; Ranganathan, S.; Chow, V.T.K.; Kwoh, C.K. Computational analysis of the receptor binding specificity of novel influenza a/h7n9 viruses. *BMC Genomics,* **2018**, *19*(2), 41-50.

[49] Yin, R.; Zhang, Y.; Zhou, X.; Kwoh, C.K. Time series computational prediction of vaccines for influenza A H_3N_2 with recurrent neural networks. *J. Bioinform. Comput. Biol.,* **2020**, *18*(1), 2040002. http://dx.doi.org/10.1142/S0219720020400028 PMID: 32336247

[50] Randhawa, G.S.; Hill, K.A.; Kari, L. Mldsp-gui: An alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis. *Bioinformatics,* **2020**, *36*(7), 2258-2259. http://dx.doi.org/10.1093/bioinformatics/btz918 PMID: 31834361

[51] Marsland, S. *Machine Learning: An Algorithmic Perspective*; CRC Press: Boca Raton, FL, USA, **2015**.

[52] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.,* **2012**, 1097-1105.

[53] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv,* **2014**.

[54] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016 Jun 27-3; Las Vegas, NV, USA; pp. 770-778.

[55] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.,* **2011**, *12*, 2825-2830.

[56] Paszke, Adam; Gross, Sam; Chintala, Soumith; Chanan, Gregory; Yang, Edward; DeVito, Zachary; Lin, Zeming; Desmaison, Alban; Antiga, Luca; Lerer, Adam Automatic differentiation in pytorch. **2017**.

[57] Grigoriev, A. Mutational patterns correlate with genome organization in SARS and other coronaviruses. *Trends Genet.,* **2004**, *20*(3), 131-135. http://dx.doi.org/10.1016/j.tig.2004.01.009 PMID: 15049309

[58] Pyrc, K.; Jebbink, M.F.; Berkhout, B.; van der Hoek, L. Genome structure and transcriptional regulation of human coronavirus NL63. *Virol. J.,* **2004**, *1*(1), 7. http://dx.doi.org/10.1186/1743-422X-1-7 PMID: 15548333

[59] Woo, P.C.Y.; Huang, Y.; Lau, S.K.P.L.; Yuen, K.-Y. Coronavirus genomics and bioinformatics analysis. *viruses,* **2010**, *2*(8), 1804-1820.

[60] Berkhout, B.; van Hemert, F. On the biased nucleotide composition of the human coronavirus RNA genome. *Virus Res.,* **2015**, *202*, 41-47. http://dx.doi.org/10.1016/j.virusres.2014.11.031 PMID: 25656063

[61] Solis-Reyes, S.; Avino, M.; Poon, A.; Kari, L. An open-source k-MER based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS One,* **2018**, *13*(11), e0206409. http://dx.doi.org/10.1371/journal.pone.0206409 PMID: 30427878

[62] Kissler, S.M.; Tedijanto, C; Goldstein, E; Grad, Y; Lipsitch, M Projecting the transmission dynamics of SARS-COV-2 through the post pandemic period. *Science,* **2020**, *368*(6493), 860-868. http://dx.doi.org/10.1126/science.abb5793 PMID: 32291278

[63] Su, S.; Wong, G.; Shi, W.; Liu, J.; Lai, A.C.K.; Zhou, J.; Liu, W.; Bi, Y.; Gao, G.F. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.,* **2016**, *24*(6), 490-502. http://dx.doi.org/10.1016/j.tim.2016.03.003 PMID: 27012512

[64] Abdul-Rasool, S.; Fielding, B.C. Understanding human coronavirus hcov-nl63. *Open Virol. J.,* **2010**, *4*, 76-84. http://dx.doi.org/10.2174/1874357901004010076 PMID: 20700397

[65] Zeng, Z-Q.; Chen, D-H.; Tan, W-P.; Qiu, S-Y.; Xu, D.; Liang, H-X.; Chen, M-X.; Li, X.; Lin, Z-S.; Liu, W-K.; Zhou, R. Epidemiology and clinical characteristics of human coronaviruses OC43, 229E, NL63, and HKU1: A study of hospitalized children with acute respiratory tract infection in Guangzhou, China. *Eur. J. Clin. Microbiol. Infect. Dis.,* **2018**, *37*(2), 363-369. http://dx.doi.org/10.1007/s10096-017-3144-z PMID: 29214503

[66] Dijkman, R.; Jebbink, M.F.; El Idrissi, N.B.; Müller, M.A.; Kuijpers, T.W.; Zaaijer; der Hoek, L.V. Human coronavirus NL63 and 229E seroconversion in children. *J. Clin. Microbiol.,* **2008**, *46*(7), 2368-2373. http://dx.doi.org/10.1128/JCM.00533-08 PMID: 18495857

[67] Zhu, H.; Guo, Q.; Li, M.; Wang, C.; Fang, Z.; Wang, P.; Tan, J.; Wu, S.; Xiao, Y Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. *BioRxiv,* **2020**.

[68] Letko, M.; Marzi, A.; Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B beta coronaviruses. *Nat. Microbiol.,* **2020**, *5*(4), 562-569. http://dx.doi.org/10.1038/s41564-020-0688-y PMID: 32094589

[69] Li, W.; Shi, Z.; Yu, M.; Ren, W.; Smith, C.; Epstein, J.H.; Wang, H.; Crameri, G.; Hu, Z.; Zhang, H.; Zhang, J.; McEachern, J.; Field, H.; Daszak, P.; Eaton, B.T.; Zhang, S.; Wang, L.F. Bats are natural reservoirs of SARS-like coronaviruses. *Science,* **2005**, *310*(5748), 676-679. http://dx.doi.org/10.1126/science.1118391 PMID: 16195424

[70] Guan, Y.; Zheng, B.J.; He, Y.Q.; Liu, X.L.; Zhuang, Z.X.; Cheung, C.L.; Luo, S.W.; Li, P.H.; Zhang, L.J.; Guan, Y.J.; Butt,

K.M.; Wong, K.L.; Chan, K.W.; Lim, W.; Shortridge, K.F.; Yuen, K.Y.; Peiris, J.S.; Poon, L.L. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science,* **2003**, *302*(5643), 276-278. http://dx.doi.org/10.1126/science.1087139 PMID: 12958366

[71] Alagaili, A.N.; Briese, T.; Mishra, N.; Kapoor, V.; Sameroff, S.C.; Burbelo, P.D.; de Wit, E.; Munster, V.J.; Hensley, L.E.; Zalmout, I.S.; Kapoor, A.; Epstein, J.H.; Karesh, W.B.; Daszak, P.; Mohammed, O.B.; Lipkin, W.I. Middle east respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. *MBio,* **2014**, *5*(2), e00884-e14. http://dx.doi.org/10.1128/mBio.01002-14 PMID: 24570370

[72] Shao, P.; Shan, Y. Beware of asymptomatic transmission: Study on 2019-ncov prevention and control measures based on extended SEIR model. *BioRxiv,* **2020**. http://dx.doi.org/10.1101/2020.01.28.923169

[73] Zhao, S.; Lin, Q.; Ran, J.; Musa, S.S.; Yang, G.; Wang, W.; Lou, Y.; Gao, D.; Yang, L.; He, D.; Wang, M.H. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int. J. Infect. Dis.,* **2020**, *92*, 214-217. http://dx.doi.org/10.1016/j.ijid.2020.01.050 PMID: 32007643

[74] Lurie, N.; Saville, M.; Hatchett, R.; Halton, J. Developing covid-19 vaccines at pandemic speed. *N. Engl. J. Med.,* **2020**, *382*(21), 1969-1973. http://dx.doi.org/10.1056/NEJMp2005630 PMID: 32227757

[75] Zhao, Y.; Zhao, Z.; Wang, Y.; Zhou, Y.; Ma, Y.; Zuo, W Single-cell RNA expression profiling of ace2, the putative receptor of Wuhan 2019-nCov. *BioRxiv,* **2020**.