INTERNATIONAL JOURNAL OF BIOLOGY AND BIOMEDICAL ENGINEERING
DOI: 10.46300/91011.2021.15.31 Volume 15, 2021

Geographic classification and identification of SARS-CoV2 from related viral sequences

Fayroz. F. Sherif [1] *, Khaled. S. Ahmed [2],
[1] Computers and Systems Department, Electronics Research Institute (ERI), Cairo, Egypt
[2] Bio-medical Department, Benha University, Benha, Egypt

Received: January 24, 2021. Revised: June 24, 2021. Accepted: July 12, 2021. Published: July 19, 2021.

Abstract— The COVID-19 pandemic has introduced to mild the risks of deadly epidemic-prone illnesses sweeping our globalized planet. The pandemic is still going strong, with additional viral variations popping up all the time. For the close to future, the international response will have to continue. The molecular tests for SARS-CoV-2 detection may lead to False-negative results due to their genetic similarity with other coronaviruses, as well as their ability to mutate and evolve. Furthermore, the clinical features caused by SARS-CoV-2 seem to be like the symptoms of other viral infections, making identification even harder. We constructed seven hidden Markov models for each coronavirus family (SARS-CoV2, HCoV-OC43, HCoV-229E, HCoV-NL63, HCoV-HKU1, MERS-CoV, and SARS-CoV), using their complete genome to accurate diagnose human infections. Besides, this study characterized and classified the SARS-CoV2 strains according to their different geographical regions. We built six SARS-CoV2 classifiers for each world's continent (Africa, Asia, Europe, North America, South America, and Australia). The dataset used was retrieved from the NCBI virus database. The classification accuracy of these models achieves 100% in differentiating any virus model among others in the Coronavirus family. However, the accuracy of the continent models showed a variable range of accuracies, sensitivity, and specificity due to heterogeneous evolutional paths among strains from 27 countries. South America model was the highest accurate model compared to the other geographical models. This finding has vital implications for the management of COVID-19 and the improvement of vaccines.

Keywords—COVID-19, Geographic classification, Profile hidden Markov model, SARS-CoV2 identification.

I. Introduction

Coronaviruses are single-stranded positive-sense RNA viruses with genomes up to around 32-kilo base-pairs (kbps) in length. Coronaviruses have four main sub-groups: alpha, beta, gamma, and delta coronavirus [1]. Human coronaviruses were first identified in the mid-1960s. Coronaviruses mainly affect the respiratory tract of animals and humans that causes mild to severe respiratory tract infection [2]. The common types that can infect humans are 229E and NL63 (alpha coronavirus), OC43, and HKU1 (beta coronavirus). Also, the two highly pathogenic beta coronaviruses, MERS-CoV and SARS-CoV, that appeared in the last twenty years caused East Respiratory Syndrome (MERS) and severe acute respiratory syndrome (SARS), respectively [3]. Including the most recent SARS-CoV2 that causes COVID-19 pandemic disease. Extensive research efforts were made to understand the SARS-CoV2, considering its genome, origin, and evolution to stop or cure the covid-19 disease. Many studies have reported high similarities in the genomic features between SARS-CoV-2 and other coronaviruses [2]. Also, the similarity in the disease's symptoms that caused by these coronaviruses and COVID-19 infection.

Molecular techniques, like quantitative RT-PCR and DNA sequencing techniques, are typically used to detect pathogens. One main issue with the RT-PCR test is the risk of obtaining false positive and false negative outcomes [4]. It is said that many 'suspected' instances with common clinical characteristics of COVID-19 and equal computed tomography (CT) scan images have been no longer identified [4]. Therefore, a bad result does no longer excludes the possibility of COVID19 disease and should no longer be used as the only criterion for treatment or patient control decisions. It seems that using both real-time RT-PCR and clinical characteristics allows control of SARS-CoV2 outbreak. Several challenges regarding the detection of SARS-CoV2 using RT-PCR like time-consuming and calls for optimizing additional parameters [5]. Ultimately, rules need to be carefully decided to ensure the assay's reliability and detect experimental errors[4].

Due to the frequently growing rate of novel viral sequences, many studies attempt to improve diagnostic procedures and classify these viral genome sequences using computational methods that can provide rapid and reproducible outcomes.

There was an attempt to assemble viral sequences from various viral taxonomic groups into coordinated databases in recent years. The taxonomic classification primarily based on alignment approaches is considerably successful in locating sequence similarities and grouping the relatively correlated viruses as approved by ICTV (International Committee on Taxonomy of Viruses) [5].

.

BLAST algorithm is a common way to discover sequence similarity. However, BLAST search algorithm has some issues associated with remote homology detection described in [6]. Profile hidden Markov model (Profile HMM) is one of Sequence similarity searches' most important methods, especially among viruses. HMMs are statistical models that convert Multiple sequence alignment (MSA) data into a set of probability values that reflect the position-specific scores about how conserved each residue of the Alignment is and which residues are likely. These models display higher sensitivity than BLAST for the detection of remote homologs [7]. One of the interesting uses of viral profile HMMs is the sequence recognition and reconstruction of precise viral genomes from metagenomic information. Prior studies used HMM models to build highly accurate models for classifying Influenza (A) pandemics according to their outer surface proteins [8, 9].

Authors had proposed other solutions primarily based on deep learning to classify viruses by splitting the sequences into portions of fixed lengths, from 300 bps to 3000 bps [10]. However, this method hardly achieves 0.923 AUC as they ignored a part of the data that cannot fill the fixed-length number. Another study used a deep learning method to separate 50 sequences of SARS-CoV2 from 352 sequences of related viruses, considering all DNA genome sequences. This study achieved 0.97 AUC, 0.9939 specificity and 0.9 sensitivity [11]. The latest study [12] used a convolutional neural network (CNN) and a bidirectional long short-term memory (Bi-LSTM) neural network CNN-Bi-LSTM to identify the SARS CoV-2 virus from coronaviruses and predict the short regulatory motifs bound to the proteins. They used an unbalanced dataset with 10.3% (SARS CoV-2) and 89.7% negative samples belonged to other viruses among the Coronavirus family, and their binary classification achieved an accuracy of 99% as SARS CoV-2 or non-SARS CoV-2. One More binary classification study [13] had been done by extracting the genome characteristics of SARS CoV-2 vs. other forms of coronaviruses. They used SVM, KNN, Naïve Bayes, and Random Forest for classifying the samples however, they hardily achieved accuracy of 87%, 92%, 88%, and 93% respectively.

The main goal of our research is twofold: first, to study the diversity of Coronavirus family and build a classifier model for each one of the seven known coronaviruses (SARS-CoV2, HCoV-OC43, HCoV-229E, HCoV-NL63, HCoV-HKU1, MERS-CoV, and SARS-CoV), to facilitate patient diagnosis and manage the pandemic spread. Second, to investigate and characterize the SARS-CoV2 strains from different geographical regions. We build six classifiers for each world's continent (Africa, Asia, Europe, North America, South America, and Australia) to compare and classify the virus strains from different areas. This finding has vital implications for the management of COVID-19 and the improvement of vaccines.

Table I. The number of genome sequences for each virus as NCBI naming convention.

	Organism	Number of sequences
1	SARS-CoV2	860
2	MERS-CoV	466
3	HCoV-OC43	401
4	HCoV-NL,.63	170
5	HCoV-229E	122
6	HCoV-HKU1	117
7	SARS-CoV	20

Table II. The number of SARS-CoV2 genome sequences from six as retrieved from the NCBI virus database.

	Location	Number of sequences	Countries
1	North America	815	USA, Canada
2	Asia	569	China, Taiwan, India, Vietnam, South Korea, Pakistan, Kazakhstan, Sri-Lanka
3	Europe	620	Italy, France, Germany, Sweden, Spain, DEU, Greece, Czechia, Poland
4	Australis	430	Australia
5	Africa	262	Egypt, Tunis, Morocco, Nigeria
6	South America	45	Brazil, Colombia, Argentina

II. Materials and Methods

A. Dataset

This study downloaded all the available isolates of the common Coronavirus types from the NCBI virus database [9] till 8 Oct 2020. A total of 2136 human coronaviruses sequences for seven types were filtered and downloaded in Fasta file format. The downloaded sequences were forced to be unique, in complete isolation, and from different geographical origins. The samples were organized and labeled for each virus, as

summarized in Table I. The average sequence length is 30 Kbps for each type of Coronaviruses. Finally, we divided the available sequences into two parts, 90% of each virus's data is used for training the HMM models, and the remaining part is used for testing.

To classify SARS-CoV2 geographical origins, we downloaded all the newly available isolates of the SARS-CoV2 genome from the NCBI virus database [14] till 15 Dec 2020. All the previous considerations, such as uniqueness, completeness, and differential geographical origins, have been considered. The downloaded sequences were from 27 countries organized and grouped according to the world's continents, as summarized in Table II. The study downloaded a total of 2741 human SARS-CoV2 sequences in Fasta file format for geographical analysis.

B. Multiple Sequence Alignment (MSA)

Multiple sequence alignment MSA is the Alignment or comparison of three or more sequences (DNA, RNA, or protein) to investigate their diversity. MSA can compare homologous sequences and place them above each other in a matrix with a minimum number of spaces (gaps), where each column in the matrix represents a set of characters that are homologous. So that the characters may coincide when the sequences are closely related but may conflict as the sequences diverge [15]. Such MSA representation can produce a consensus sequence showing the preserved regions from the aligned sequences. These alignments mainly support exploring evolutionary relationships, phylogenetic tree reconstruction, and Profile hidden Markov model.

Algorithms like ClustalW[13] and MUSCLE[16] are well-known and widely used in MSA. In our study, MSA is done for each Coronaviruses strain separately using the CLC genomics workbench [17].

C. Modeling using Profile HMM

Profile HMMs are powerful statistical models that transform MSA data into a set of probability values that reflect the position-specific scores for each residue in the group of sequences. Profile analysis has long been a helpful tool in finding and aligning distantly associated sequences and identifying known sequence domains in new sequences[18]. A profile is a description of the consensus of a multiple sequence alignment. It uses a position-specific scoring system to capture information about conservation at various positions in multiple Alignment. This makes it a much more sensitive and specific method for searching the database than pairwise methods such as regular BLAST to recognize remote homologs.

Within the following, we have two essential steps: building profile HMM model and database searching. Building a model means changing a multiple alignment of every group of sequences right into a probabilistic model, while searching the database means scoring a sequence to the Profile HMM. One of the most broadly used profile HMM applications is the HMMER package [19].

a. Model Building

Profile HMM is represented as a series of states (begin, match, insert, delete, and end states) and arrows indicating state order as shown in figure1. The "Match" states (M1, M2, ...M5) each represents one column in the alignment that emits residue with probabilities learned through model estimation. Simultaneously, the "insert" state (I0, I1, ...I5) means placing new residues preceding each "match" state. While "delete" states (D1, D2, ...D5) are silent that can be used to pass or skip the "match" state. The transition arrows indicating that we can go through the insert states, placing new residues or go through the silent states, skipping one or extra of the match states. This architecture guarantees that we can examine each new sequence and, at the same time, reduce the range of parameters in the constructed model.

In this study, we construct a unique HMM profile for each type of coronaviruses using the 'Hmmbuild' program in HMMER package v3.3.1[19]. The 'Hmmbuild' program's input was the pre-aligned sequences of the training dataset summarized in table I. To improve the database search sensitivity, we used the 'hmmcalibrate' program in HMMER to calculate the expectation values E-value. E-value is the statistical significance of the match to this sequence. The more significant sequence, the lower the E-value, and the larger the database you seek, the more the range of expected false positives. Finally, HMM database has been constructed using concatenating HMM files that are already built and calibrated [22].

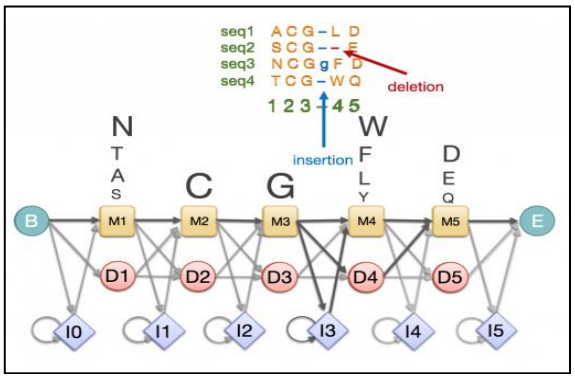


Figure 1. A profile HMM modelling a multiple sequence alignment. https://www.ebi.ac.uk/training/online/sites

b. Database Searching

Searching database asks if the complete target sequence is homologous (or now not) to a query profile. Every sequence in the test dataset can be matched to a profile model by calculating its probability generated by that model.

The 'nhmmer' program in HMMER3 can scan each comparison between a test sample and target searching for high-score un-gapped alignment segments. A window around each such segment is extracted, merging overlapping windows. HMMER3 implements two search algorithms to

score a sequence fit to HMM: Viterbi algorithm that provides the score of the most likely alignment with the sequence or forward algorithm that calculates the score as the sum over all possible alignments to the HMM of the Profile the total probability of a sequence aligning to the HMM [20]. The classification is done by scoring the whole test-sets for each group (Table I), with each coronavirus HMM model, using the 'nhmmer' program in HMMER3. For classifying SARS-CoV-2 from other related coronaviruses, we labeled SARS-CoV2 by "1" and grouped the other viruses in label "0". Matches to the right coronavirus type are classified as true hits. The results of this program are the sequence top hits list, ranked with the E-value. The rankings and E-values here reflect the certainty that this target sequence consists of one or more domain names belonging to the hmm family [19].

III. Results and Discussion

Multiple sequence alignments (MSA) were done for SARS-CoV2 sequences and each type of coronaviruses (table I) separately using the CLC genomics workbench [17]. Then we built HMM specific model for each aligned group of coronaviruses, followed by HMM calibration and database searching using the HMMER.3 package. Regarding SARS-CoV2 geographic classification, MSA was done for each organized continent group as summarized in Table II. As the previous structure, each aligned group of SARS-CoV2 continent was used to build a specific HMM model for that continent, followed by HMM calibration and database searching in the same package. In the following subsections, we would present and discuss the classification results of Coronavirus family and geographic classification results of SARS-CoV2, respectively.

A. Classification results of Coronavirus family

Large The proposed HMM models improved the accuracy of Covid-19 diagnosis and accomplished classification accuracy of 100%. The classification of each proposed virus model vs. others in the Coronavirus family like (SARS-CoV2 vs. non-SARS-CoV2), (MERS-CoV vs. non- MERS-CoV), (HCoV-OC43 vs. non- HCoV-OC43) achieved 100% sensitivity and specificity. Besides, the proposed models improved the accuracy of Covid-19 diagnosis compared to literature studies. The authors in [13] performed one binary classifier (COVID-19 vs. other types of coronaviruses) and hardly achieved an accuracy of 87%, 92%, 88%, and 93% using SVM, KNN, Naïve Bayes, and Random Forest, respectively. The study in [10] used deep learning to classify SARS-CoV2. However, their results hardly achieved 0.923 AUC as they ignored a part of the data that cannot fill the fixed-length number. While the authors in [11] used a small dataset of complete sequences to separate (50 SARS-CoV2 among 352 from others), the result achieved a specificity of 99% and a sensitivity of 90% using a convolutional neural network.

B. Geographic classification results of SARS-CoV2

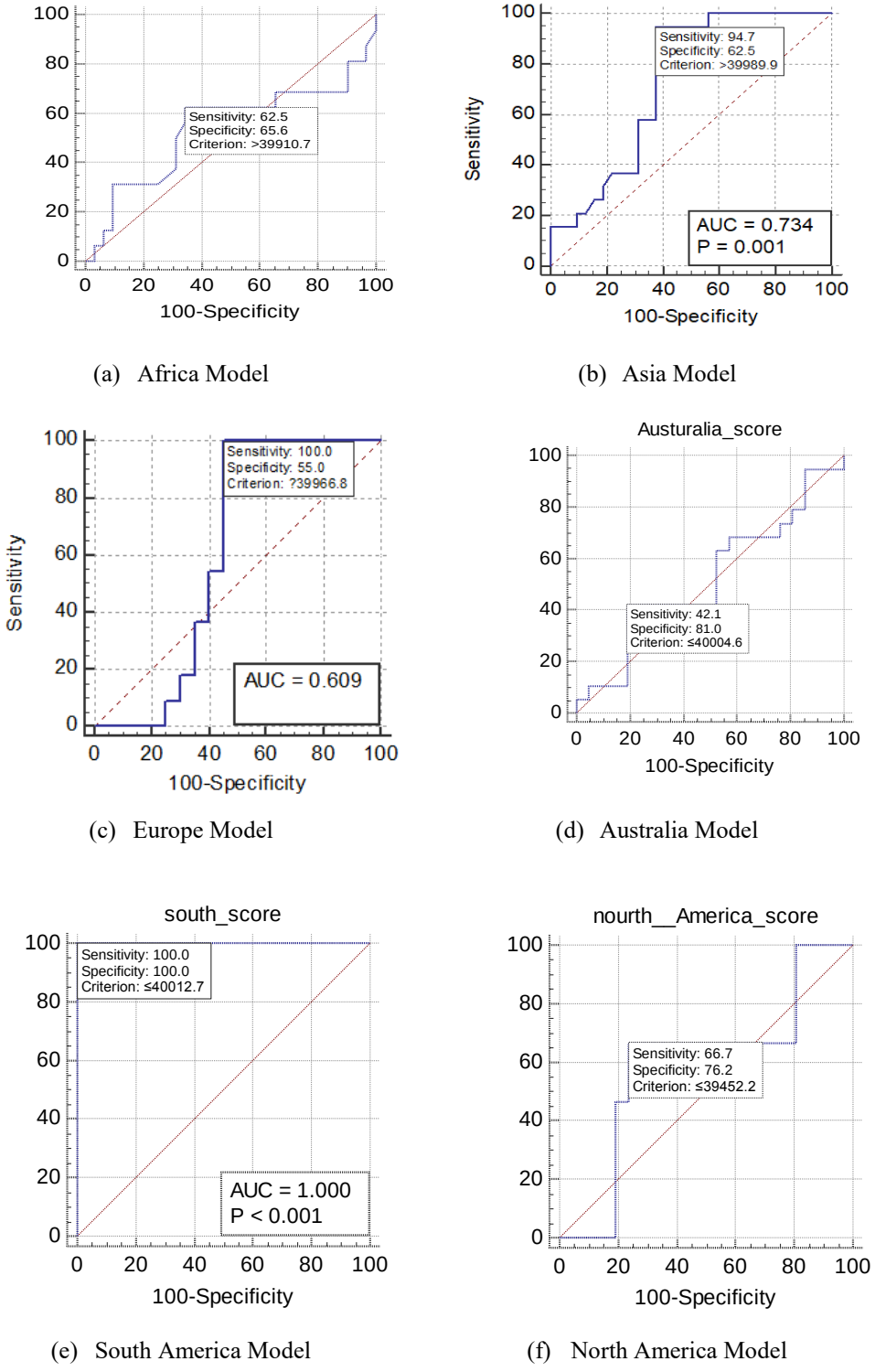
According to the geographic classification of Covid-19, identifying the geographical origin of SARS-CoV2 strains and classifying them according to the six world continents has been done by scoring each continent model with its corresponding test-set (Table I). We used ‘nhmmer' program in HMMER3 to search DNA test queries against each continent model. Receiver operating characteristic (ROC) curve plot the true positive TP rate (sensitivity) versus the false-positive FP rate (100-specificity) for different cut-off points. Each point on the ROC figure denotes a sensitivity-specificity pair consistent with a particular decision or threshold. The following ROC curves were plotted using the MedCalc program [21]. The curves show the observed threshold values that achieved maximum sensitivity and specificity. Figure 2 showed the ROC curves of each continent model according to the geographical origin of SARS-CoV2 sequences. The area under the ROC curve (AUC) varied between 0.526 and 1. Table III summarizes the geographic classification results of continent models in terms of AUC, sensitivity, and specificity.

Table III Summary of geographic classification results of SARS-CoV2 using HMM.

	Model	AUC	Sensitivity	Specificity	Significant level P
1	South America	1	100%	100%	0.001
2	Asia	0.734	94.74%	62.5%	0.001
3	North America	0.73	66.7%	76.2%	0.05
4	Europe	0.609	100%	55%	0.03
5	Africa	0.539	62.5%	65.62%	0.07
6	Australia	0.526	42.11%	80.95%	0.07

IV. Conclusions

With the high transmissibility of the SARS-CoV2, the proper diagnosis of Covid-19 is urgent to prevent the virus from spreading also. Considering the false negatives given by RT-



(a) Africa Model (b) Asia Model

(c) Europe Model (d) Australia Model

(e) South America Model (f) North America Model

Figure 2. ROC curves for geographical classification results of SARS-CoV2 using HMM

PCR, higher implementations such as the viral classification model usage are vital to come across the virus properly. Here, we improved the accuracy of the SARS-CoV2 diagnosis to reach 100% using our HMM models. In conclusion, our results indicate that SARS-CoV2 sequences are susceptible to HMM models. Moreover, our geographic classification models of SARS-CoV2 showed a virus genome diversity associated with the geographical distribution across the world. However, our preliminary results showing a variable range of accuracies, sensitivity, and specificity. South America was the highest accurate model, followed by Asia, North America, Europe, Africa, and Australia models. The geographic classification models may be further improved and validated by adding more samples for more countries that have not been available in NCBI until the date of data retrieval. We concluded that profile HMMs could effectively detect and identify the diversity of SARS-CoV2 and the other six known coronaviruses. Moreover, this geographic analysis of the human COVID-19 shows possibly heterogeneous evolutional paths among strains from 27 countries. This finding has vital implications for controlling COVID-19 and developing vaccines.

References

[1] M. Teymoori-Rad, S. Samadizadeh, A. Tabarraei, A. Moradi, M. B. Shahbaz, and A. Tahamtan, "Ten challenging questions about SARS-CoV-2 and COVID-19," *Expert Rev Respir Med,* pp. 1-8, Jun 30 2020.

[2] S. Cleemput *et al.*, "Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes," *Bioinformatics,* vol. 36, no. 11, pp. 3552-3555, 2020.

[3] A. Algaissi, A. S. Agrawal, A. M. Hashem, and C. K. Tseng, "Quantification of the Middle East Respiratory Syndrome-Coronavirus RNA in Tissues by Quantitative Real-Time RT-PCR," *Methods Mol Biol,* vol. 2099, pp. 99-106, 2020.

[4] A. Tahamtan and A. Ardebili, "Real-time RT-PCR in COVID-19 detection: issues affecting the results," *Expert Rev Mol Diagn,* vol. 20, no. 5, pp. 453-454, May 2020.

[5] A. J. Davison, "Journal of General Virology – Introduction to 'ICTV Virus Taxonomy Profiles'," vol. 98, no. 1, pp. 1-1, 2017.

[6] G. Lu *et al.*, "GenomeBlast: a web tool for small genome comparison," *BMC Bioinformatics,* vol. 7 Suppl 4, p. S18, Dec 12 2006.

[7] P. Skewes-Cox, T. J. Sharpton, K. S. Pollard, and J. L. DeRisi, "Profile hidden Markov models for the detection of viruses within metagenomic sequence data," (in eng), *PloS one,* vol. 9, no. 8, pp. e105067-e105067, 2014.

[8] M. ElHefnawi and F. F. Sherif, "Accurate classification and hemagglutinin amino acid signatures for influenza A virus host-origin association and subtyping," *Virology,* vol. 449, pp. 328-338, 2014/01/20/ 2014.

[9] F. F. SHERIF, Y. M. KADAH, and M. EL-HEFNAWI, "INFLUENZA A SUBTYPING AND HOST ORIGIN CLASSIFICATION USING PROFILE HIDDEN MARKOV MODELS," vol. 12, no. 02, p. 1240009, 2012.

[10] A. Tampuu, Z. Bzhalava, J. Dillner, and R. Vicente, "ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples," *PLoS One,* vol. 14, no. 9, p. e0222271, 2019.

[11] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado, E. Claassen, J. Garssen, and A. D. Kraneveld, "Accurate Identification of SARS-CoV-2 from Viral Genome Sequences using Deep Learning," p. 2020.03.13.990242, 2020.

[12] A. Whata and C. Chimedza, "Deep Learning for SARS COV-2 Genome Sequences," *IEEE Access,* vol. 9, pp. 2169-3536, 04/16 2021.

[13] H. Arslan, "Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data," vol. 74, no. 1, p. 20, 2021.

[14] *NCBI virus database* Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/

[15] H. Shi and X. Zhang, "Component-Based Design and Assembly of Heuristic Multiple Sequence Alignment Algorithms," *Front Genet,* vol. 11, p. 105, 2020.

[16] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics,* vol. 5, no. 1, p. 113, 2004/08/19 2004.

[17] *CLC Workbench*. Available: https://digitalinsights.qiagen.com

[18] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn, "HMMER web server: 2018 update," *Nucleic Acids Research,* vol. 46, no. W1, pp. W200-W204, 2018.

[19] *HMMER package v3.3.1*. Available: http://hmmer.org/

[20] L. Huo, H. Zhang, X. Huo, Y. Yang, X. Li, and Y. Yin, "pHMM-tree: phylogeny of profile hidden Markov models," *Bioinformatics,* vol. 33, no. 7, pp. 1093-1095, Apr 1 2017.

[21] *MedCalc program*. Available: https://www.medcalc.org/