bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Accurate Identification of SARS-CoV-2 from Viral Genome Sequences using Deep Learning

Alejandro Lopez-Rincon[1], Alberto Tonda[2], Lucero Mendoza-Maldonado[3], Eric Claassen[5], Johan Garssen[1,4] and Aletta D. Kraneveld[1]

[1]Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Universiteitsweg 99, 3584 CG Utrecht, the Netherlands; [2]UMR 518 MIA-Paris, c/o 113 rue Nationale, 75103, Paris, France; [3] Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara, Sierra Mojada No. 950, Col. Independencia C.P. 44348 Guadalajara, Jalisco, Mexico; [4]Department Immunology, Danone Nutricia research, Uppsalalaan 12, 3584 CT Utrecht, the Netherlands; [5]Athena Institute, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, the Netherlands.

Abstract

One of the reasons for the fast spread of SARS-CoV-2 is the lack of accuracy in detection tools in the clinical field. Molecular techniques, such as quantitative real-time RT-PCR and nucleic acid sequencing methods, are widely used to identify pathogens. For this particular virus, however, they have an overall unsatisfying detection rate, due to its relatively recent emergence and still not completely understood features. In addition, SARS-CoV-2 is remarkably similar to other Coronaviruses, and it can present with other respiratory infections, making identification even harder. To tackle this issue, we propose an assisted detection test, combining molecular testing with deep learning. The proposed approach employs a state-of-the-art deep convolutional neural network, able to automatically create features starting from the genome sequence of the virus. Experiments on data from the Novel Coronavirus Resource (2019nCoVR) show that the proposed approach is able to correctly classify SARS-CoV-2, distinguishing it from other coronavirus strains, such as MERS-CoV, HCoV-NL63, HCoV-OC43, HCoV-229E, HCoV-HKU1, and SARS-CoV regardless of missing information and errors in sequencing (noise). From a dataset of 553 complete genome non-repeated sequences that vary from 1,260 to 31,029 bps in length, the proposed approach classifies the different coronaviruses with an average ac-

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

curacy of 98.75% in a 10-fold cross-validation, identifying SARS-CoV-2 with an AUC of 98%, specificity of 0.9939 and sensitivity of 1.00 in a binary classification. Then, using the same basis, we classify SARS-CoV-2 from 384 complete viral genome sequences with human host, that contain the gene *ORF1ab* from the NCBI with a 10-fold accuracy of 98.17% , a specificity of 0.9797 and sensitivity of 1.00. These preliminary results seem encouraging enough to identify deep learning as a promising research venue to develop assisted detection tests for SARS-CoV-2. At this end the interaction between viromics and *deep learning*, will hopefully help to solve global infection problems. In addition, we offer our code and processed data to be used for diagnostic purposes by medical doctors, virologists and scientists involved in solving the SARS-CoV-2 pandemic. As more data become available we will update our system.

Keywords: convolutional neural networks, coronavirus, deep learning, SARS-CoV-2

---

1. Introduction

The Coronaviridae family presents a positive sense, single-strand RNA genome. This viruses have been identified in avian and mammal hosts, including humans. Coronaviruses have genomes from 26.4 kilo base-pairs (kbps) to 31.7 kbps, with G + C contents varying from 32% to 43%, and human-infecting coronaviruses include SARS-CoV, MERS-CoV, HCoV-OC43, HCoV-229E, HCoV-NL63 and HCoV-HKU1 [1]. In December 2019, SARS-CoV-2, a novel, human-infecting Coronavirus was identified in Wuhan, China, using Next Generation Sequencing [2].

As a typical RNA virus, new mutations appears every replication cycle of Coronavirus, and its average evolutionary rate is roughly 10-4 nucleotide substitutions per site each year [2]. In the specific case of SARS-CoV-2, RT-qPCR testing using primers in ORF1ab and N genes have been used to identified the infection in humans. However, this method presents a high false negative rate (FNR), with a detection rate of 30-50% [3, 4]. This low detection rate can be

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

explained by the variation of viral RNA sequences within virus species, and the viral load in different anatomic sites [5]. Population mutation frequency of site 8,872 located in ORF1ab gene and site 28,144 located in ORF8 gene gradually increased from 0 to 29% as the epidemic progressed [6].

As of March 6th of 2020, the new SARS-CoV-2 has 98,192 confirmed cases across 88 countries, with 17,481 cases outside of China [7]. In addition, SARS-CoV-2 has an estimated mortality rate of 3-4%, and it is spreading faster than SARS-CoV and MERS-CoV [8]. SARS-CoV-2 assays can yield false positives if they are not targeted specifically to SARS-CoV-2, as the virus is closely related to other Coronavirus organisms. In addition, SARS-CoV-2 may present with other respiratory infections, which make it even more difficult to identify [9, 10]. Thus, it is fundamental to improve existing diagnostic tools to contain the spread. For example, diagnostic tools combining computed tomography (CT) scans with deep learning have been proposed, achieving an improved detection accuracy of 82.9% [11]. Another solution for identifying SARS-CoV-2 is additional sequencing of the viral complementary DNA (cDNA). We can use sequencing data with cDNA, resulting from the PCR of the original viral RNA; e,g, Real-Time PCR amplicons (Fig. 1) to identify the SARS-CoV-2 [12].

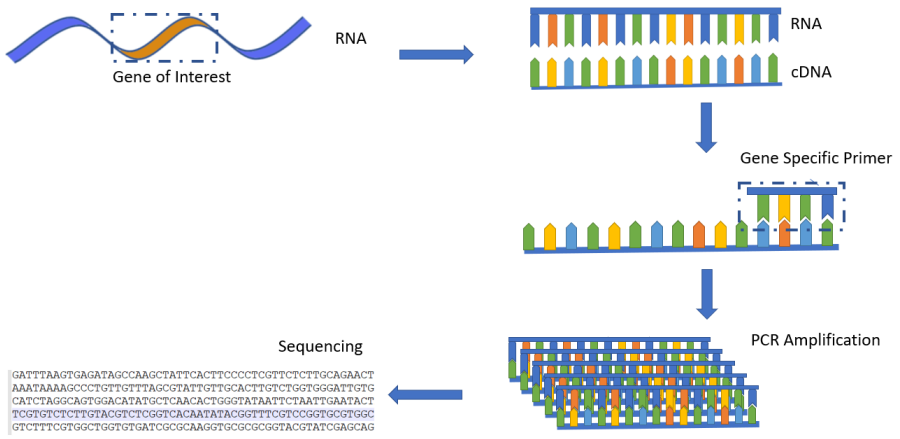

Figure 1: PCR Amplicons sequencing procedure.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Classification using viral sequencing techniques is mainly based on alignment methods such as FASTA [13] and BLAST [14]. These methods rely on the assumption that DNA sequences share common features, and their order prevails among different sequences [15, 16]. However, these methods suffer from the necessity of needing base sequences for the detection [17]. Nevertheless, it is necessary to develop innovative improved diagnostic tools that target the genome to improve the identification of pathogenic variants, as sometimes several tests, are needed to have an accurate diagnosis. As an alternative deep learning methods have been suggested for classification of DNA sequences, as these methods do not need pre-selected features to identify or classify DNA sequences. Deep Learning has been efficiently used for classification of DNA sequences, using one-hot label encoding and Convolution Neural Networks (CNN) [18, 19], albeit the examples in literature are featuring DNA sequences of length up to 500 bps, only.

In particular, for the case of viruses, Next Generation Sequencing (NGS) genomic samples might not be identified by BLAST, as there are no reference sequences valid for all genomes, as viruses have high mutation frequency [20]. Alternative solutions based on deep learning have been proposed to classify viruses, by dividing sequences into pieces of fixed lengths, from 300 bps [20] to 3,000 bps [21]. However, this approach has the negative effect of potentially ignoring part of the information contained in the input sequence, that is disregarded if it cannot completely fill a piece of fixed size.

Given the impact of the world-wide outbreak, international efforts have been made to simplify the access to viral genomic data and metadata through international repositories, such as; the 2019 Novel Coronavirus Resource (2019nCoVR) repository [6] and the National Center for Biotechnology Information (NCBI) [22], expecting that the easiness to acquire information would make it possible to develop medical countermeasures to control the disease worldwide, as it happened in similar cases earlier [23, 24, 25]. Thus, taking advantage of the available information of international resources without any political and/or economic borders, we propose an innovative system based on viral gene sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Differently from previous works in literature, that use of deep learning with fixed length features and one-hot label encoding, in this work we propose the use of a different encoding to input the full sequence as a whole. In addition, we use as base input 31,029 as an input vector, which is the maximum length of available DNA sequences for Coronavirus. Finally, we propose a novel architecture for the deep network, inspired by successful applications in cancer detection starting from miRNA [26].

2. Methods

2.1. Data

2.1.1. Classification of Coronaviruses

SARS-CoV-2 identification can give wrong results, as the virus is difficult to distinguish from other Coronaviruses, due to their genetic similarity. In addition, people with SARS-CoV-2 may present other infections besides the virus [9, 10]. Therefore, it is important to be able to properly classify SARS-CoV-2 from other Coronaviruses.

From the repository 2019 Novel Coronavirus Resource (2019nCoVR) [6], we downloaded all the available sequences with the query *Nucleotide Completeness="complete" AND host="homo sapiens"*, for a total of 588 samples. Next, we removed all repeated sequences, resulting in 553 unique sequences of variable length (1,260-31,029 bps). The data was organized and labeled as summarized by Table 1. We grouped HCoV-229E and HCoV-OC43 in the same class, as they are mostly known as Coronaviruses responsible for the common cold [27]; the two available samples of HCoV-4408 were also added to the same class, as it is a Betacoronavirus 1, as HCoV-OC43. In a similar fashion, we grouped HCoV-NL63 and HCoV-HKU1, as they are both associated with acute respiratory infections (ARI) [28]. Finally, we grouped SARS-CoV/SARS-CoV-P2/SARS-CoV HKU-39849 [29]/SARS-CoV GDH-BJH01 organisms together, as they are all strains of SARS.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 1: Organism, assigned label, and number of samples in the unique sequences obtained from the repository [6]. We use the NCBI organism naming convention [30].

Organism	Label	Number of Samples
SARS-CoV-2	0	66
MERS-CoV	1	240
HCoV-OC43	2	140
HCoV-229E	2	22
HCoV-4408	2	2
HCoV-NL63	3	58
HCoV-HKU1	3	17
SARS-CoV	4	7
SARS-CoV P2	4	1
SARS-CoV HKU-39849	4	1
SARS-CoV GDH-BJH01	4	1

To encode the cDNA data into an input tensor for the CNN, we assigned numeric values to the different bases; C=0.25, T=0.50, G=0.75, A=1.0 (see Fig. 2). All missing entries were assigned the value 0.0. This procedure is different from previous methods, that relied upon one-hot encoding [21, 20], and has the advantages of making the input more human-readable and do not multiply the amount of memory required to store the information. We divide the available samples in two parts, 90% for training and validation (80% training, 10% validation), and 10% for testing, in a 10-fold cross-validation scheme. k-fold cross-validation is a procedure by which available data is divided into k parts, called *folds*. At each iteration i, the i-th fold is used as a test set, while all the other folds are used as training. At the end of the k-th iteration, the average performance of the model in test over all folds provides a good estimate of the generality of the results. In this particular case, we use stratified folds, that preserve the same proportion of classes in every fold. The procedure is summarized by Fig. 3.

2.1.2. Separating SARS-CoV-2 from other viruses containing gene ORF1ab

Two thirds of the Coronaviruses' genome contain the ORF1ab gene [1]. Therefore, it is important that we are able to differentiate SARS-CoV-2 from similar viruses, like Astroviruses. From the NCBI repository [30], we down-

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 2: Coding for the input sequences.

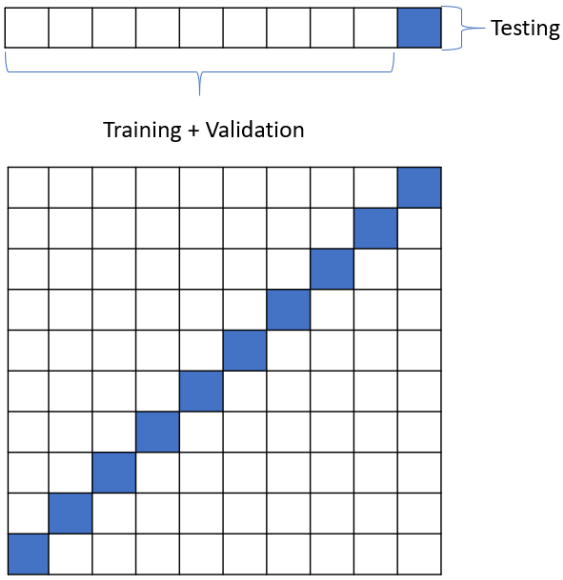

Figure 3: Scheme of a k-fold cross-validation. Available data is divided into k parts. At each iteration i, the i-th fold is used for testing, while all the others are used as a training set.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

loaded the genome sequences corresponding to the following search: *gene="ORF1ab" AND host="homo sapiens" AND "complete genome"*. This resulted in 402 sequences, distributed as described in Table 2. For this data, we assigned SARS-CoV-2 label 0, and grouped the rest of the organisms together in label 1. Next, we removed all the repeated sequences, obtaining a total of 384 unique sequences, with 45 samples belonging to SARS-CoV-2. The genomic data was translated to digits using the encoding previously described in Subsection 2.1.1.

Table 2: Organism, assigned label, and number of samples in the unique sequences obtained from the repository NCBI [30].

Virus	Label	Number of Samples
SARS-CoV-2	0	50
MERS-CoV	1	191
HCoV-OC43	1	105
HCoV-NL63	1	29
HCoV-HKU1	1	14
HCoV-4408	1	3
HCoV-229E	1	3
HAstV-VA1	1	2
HAstV-BF34	1	2
HAstV-SG	1	1
MAstV 8	1	1
HMO-A	1	1

2.2. Convolutional Neural Network

The deep learning model used for the experiments is a CNN with 3 convolutional layers and one fully connected layer, as described in Fig. 4. The input is a vector of 31,029 elements, which is the maximum size of the genome sequences in the dataset. Each convolutional layer is characterized by 3 hyperparameters, as shown in Fig. 5. The architecture is summarized by hyper-parameters $w_0 = 130, w_1 = 204, w_2 = 150, w_3 = 196, h_0 = 148, h_2 = 236, h_2 = 81, wd_0 = 9, wd_1 = 106, wd_3 = 121$ where w_3 is the number of units in the fully connected layer. To improve generality, the fully connected layer is set with a dropout

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with probability $p_d = 0.5$ during training; moreover, a $l2$ regularization is applied to the cross-categorical entropy loss function, considering all weights in the convolutional layers, with $\beta = 10^{-3}$. The optimizer used for the weights is Adaptive Moment Estimation (Adam) [31], with learning rate $lr = 10^{-5}$, run for 500 epochs. The hyper-parameters used in the experiments were selected after a set of preliminary trials. All the necessary code was developed in Python 3, using the `keras` library for deep learning [32], and has been made available on an open GitHub repository[1].

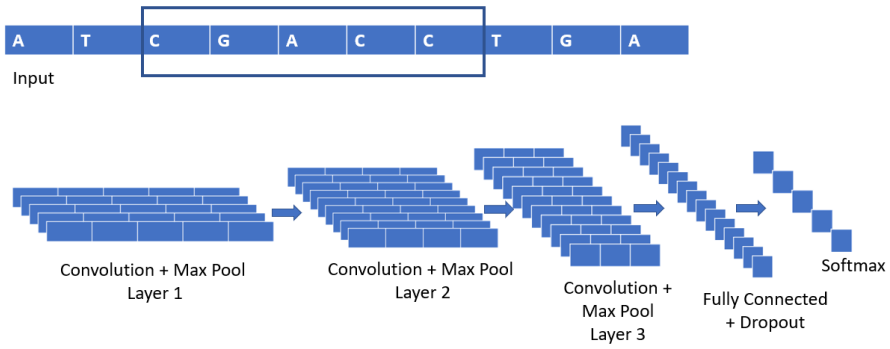

Figure 4: CNN Architecture.

3. Results

3.1. Classification of SARS-CoV-2 among Coronaviruses

In the first test, we separated the SARS-CoV-2 from other sequences available at the repository 2019 Novel Coronavirus Resource (2019nCoVR) [6]. We obtained a 10-fold average test accuracy of $\mu = 0.9875$ with $\sigma = 0.0160$. The resulting confusion matrix (Fig. 6) shows that only 3 out of the 66 SARS-CoV-2 sequences were mistakenly assigned to another class. The binarized curve of the test (Fig. 7) has an area under the curve (AUC) of 0.98, with a specificity of

[1] `http://github.org/albertotonda/deep-learning-coronavirus-genome`

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

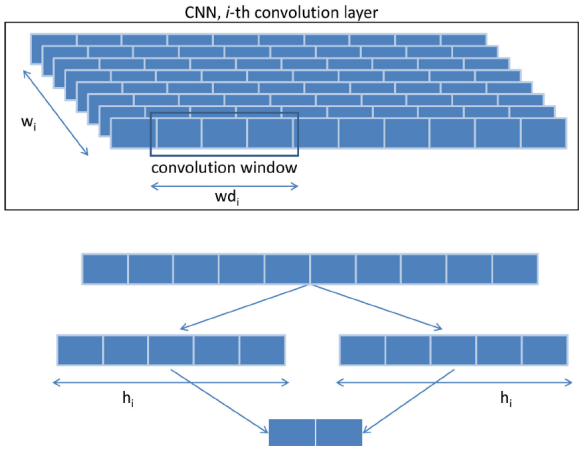

Figure 5: CNN layer description.

0.9939 and sensitivity of 1.00. This is considered an outstanding performance, according to the guidelines provided by [33, 34].

As viruses are characterized by high mutation frequencies, to assess the robustness of our approach, we performed further experiments where we added noise to the dataset, simulating possible future mutations. 5% noise was added by randomly selecting 1,551 positions from each sequence, from the 31,029 available, and modifying each selected base to another, or to a missing value, randomly. A new 10-fold cross-validation classification run on the noisy dataset yields an average accuracy $\mu = 0.9674$ with a $\sigma = 0.0158$. Figs. 8 and 9 show the resulting confusion matrix and ROC curve, respectively. This gives a AUC of 0.97, with a specificity of 0.9939 and sensitivity of 0.90.

3.2. Separating SARS-CoV-2 from other viruses containing gene ORF1ab

In a next batch of experiments, we aim to distinguish SARS-CoV-2 from other genome sequences from NCBI [30], with the following search parameters: *gene="ORF1ab" AND host="homo sapiens" AND "complete genome"*. We get a 10-fold average accuracy of $\mu = 0.9817$ with a $\sigma = 0.0167$. The resulting confusion matrix (Fig. 6) shows that 7 out of the 45 SARS-CoV-2 sequences, were classified in another class. The ROC curve of the test (Fig. 11) has an area

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

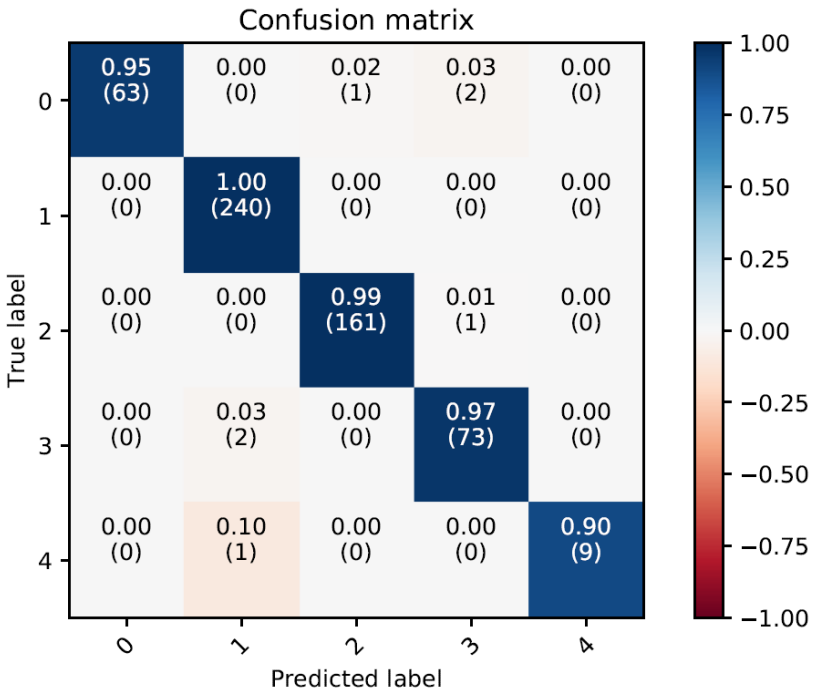

Figure 6: Confusion matrix resulting from the test of a 10-fold cross-validation, comprising 553 samples belonging to 5 different classes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

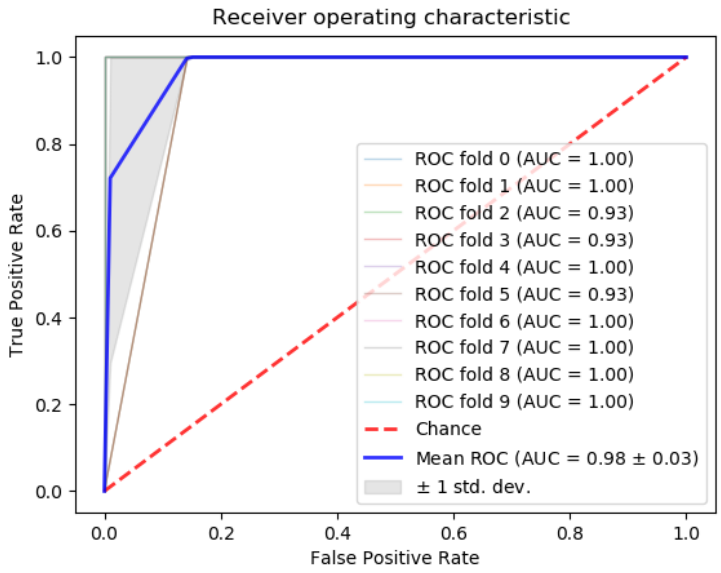

Figure 7: Binarized ROC curve of the 553 sequences, where we consider samples belonging to SARS-CoV-2 as class 0, and all the rest as class 1.

under the curve (AUC) of 0.92 , with a specificity of 0.9797 and sensitivity of 1.00.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

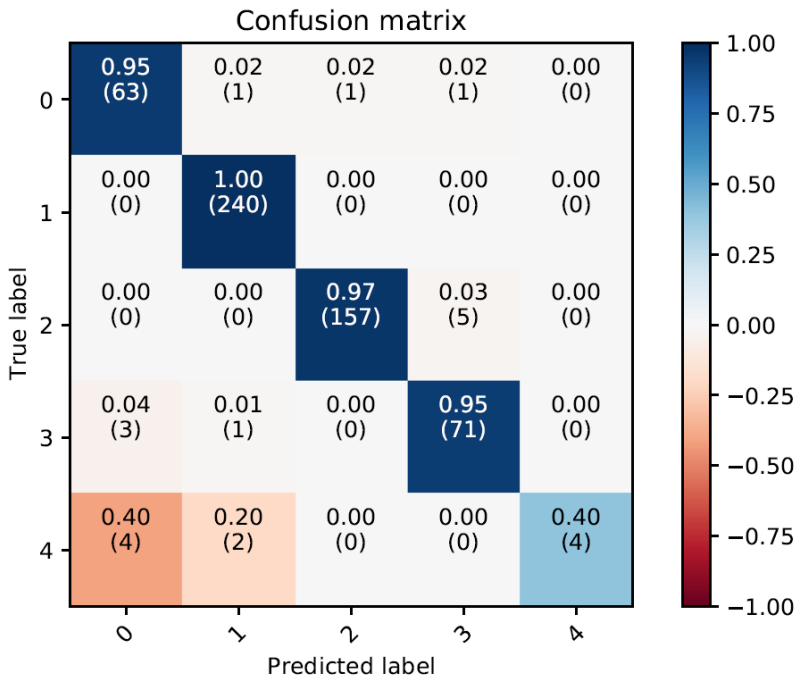

Figure 8: Confusion matrix resulting from the test of a 10-fold cross-validation, comprising 553 samples belonging to 5 different classes, with a 5% noise in the dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

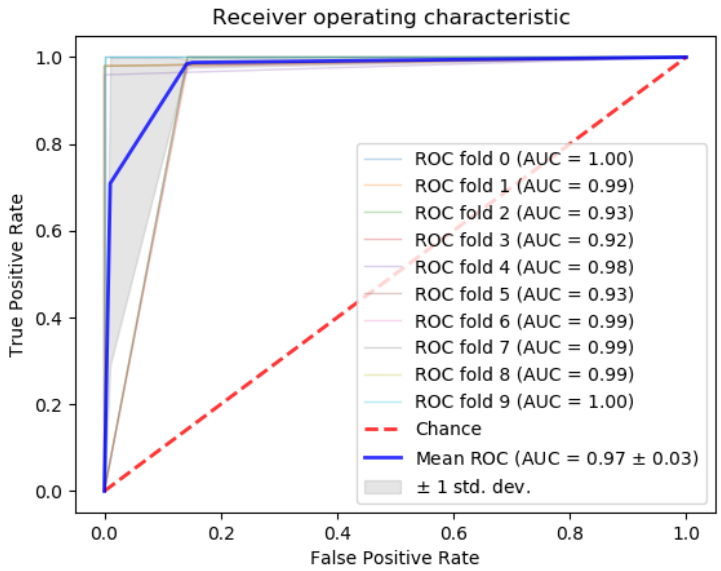

Figure 9: Binarized ROC curve of the 553 sequences, where we consider samples belonging to SARS-CoV-2 as class 0, and all the rest as class 1, with 5% added noise.

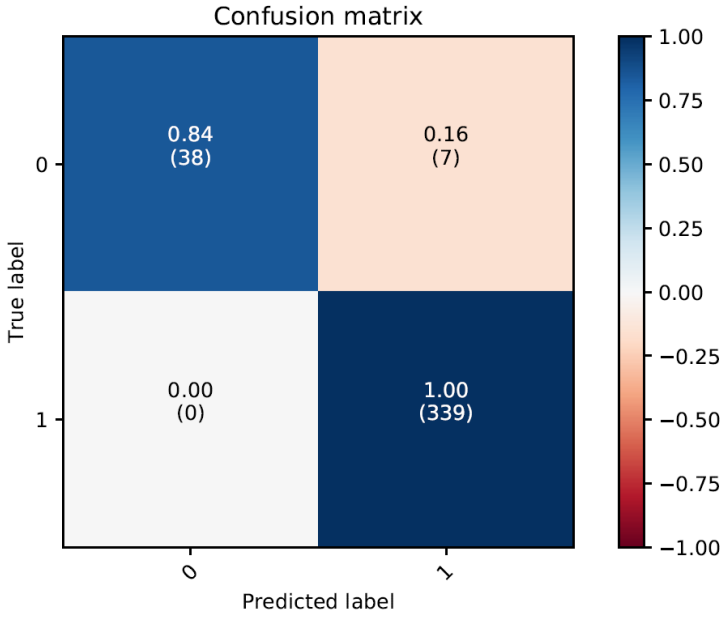

Figure 10: Confussion Matrix of the 384 NCBI sequences with 2 classes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

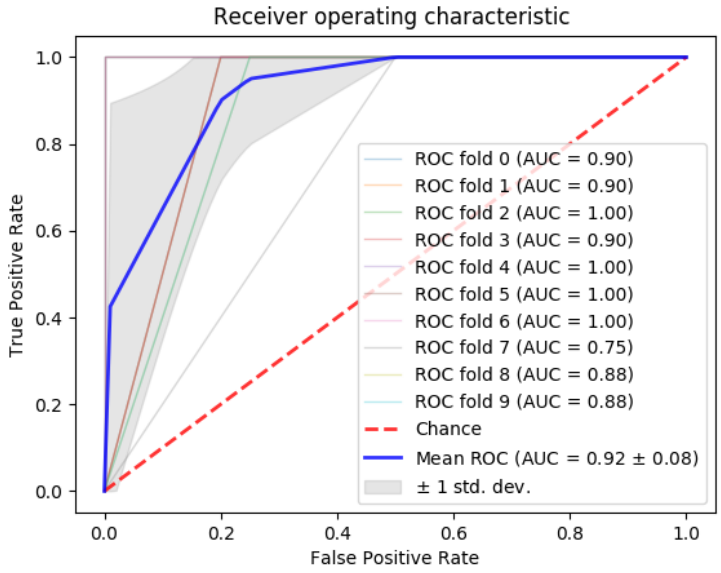

Figure 11: ROC curve of the 384 sequences, where we consider SARS-CoV-2, as class 0 and the rest as 1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4. Conclusion

Being able to reliably identify SARS-CoV-2 and distinguish it from other similar pathogens is important to contain its spread. The time of processing samples and the availability of reliable diagnostic tests is a challenge during an outbreak. Developing innovative diagnostic tools that target the genome to improve the identification of pathogens, can help reduce health costs and time to identify the infection, instead of using unsuitable treatments or testing. Moreover, it is necessary to perform an accurate classification to identify the different species of Coronavirus, the genetic variants that could appear in the future, and the coinfections with other pathogens.

Following, the high transmissibility of the SARS-CoV-2, the proper diagnosis of the disease is urgent, to stop the virus from spreading further. Considering the false negatives given by the standard nucleic acid detection, better implementations such as using deep learning are necessary in order to to properly detect the virus. While the accuracy of current nucleic acid testing is around 30-50%, and CT scans with deep learning go up at 83%, we believe that the use of a CNN-based system with sequencing has the potential to improve the accuracy of the diagnosis above 90%.

Our preliminary results using non-repeated sequences with differences in length from 1,260 to 31,029, missing information (segments with Ns) and noise (errors) do show an area under the curve of 98% in binary classification in a 10-fold cross-validation. In order to further improve the proper classification within the 7 existing coronavirus strains, more examples of full genome sequences with *host=homo sapiens* are needed, in order to make a full sub-type classification instead of grouping HCoV-229E/OC43 and HCoV-NL60/HKU1 as we were forced to do, due to the lack of samples. Thus, to further validate our results, we will increase, and accommodate the data as it becomes available in the international repositories to further improve our system.

As of March 12th 2020, China and USA have made publicly available 50 SARS-CoV-2 virus gene sequences each. In Europe, however due to the strict

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

privacy laws, only 3 sequences; Italy, Sweden and Finland, one viral genome sequence each are available and this is of great concern. We urge to consider to make more data publicly available, in order to increase the possibility to create counter-measures to the spread of the virus.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5. Bibliography

References

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

[2] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet 395 (10224) (2020) 565–574.

[3] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, et al., Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr, Eurosurveillance 25 (3) (2020).

[4] D. K. Chu, Y. Pan, S. Cheng, K. P. Hui, P. Krishnan, Y. Liu, D. Y. Ng, C. K. Wan, P. Yang, Q. Wang, et al., Molecular diagnosis of a novel coronavirus (2019-ncov) causing an outbreak of pneumonia, Clinical chemistry (2020).

[5] D. A. Marston, L. M. McElhinney, R. J. Ellis, D. L. Horton, E. L. Wise, S. L. Leech, D. David, X. de Lamballerie, A. R. Fooks, Next generation sequencing of viral rna genomes, BMC genomics 14 (1) (2013) 444.

[6] Beijing Institute of Genomics, Chinese Academy of Science, China National Center for Bioinformation & National Genomics Data Center, `https://bigd.big.ac.cn/ncov/?lang=en`, online; accessed 27 January 2020 (2013).

[7] W. H. Organization, WHO report Coronavirus disease 2019 (COVID-19), World Health Organization,, Geneva :, 2020., licence : CC BY-NC-SA 3.0 IGO.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[8] Y. Wang, H. Kang, X. Liu, Z. Tong, Combination of rt-qpcr testing and clinical features for diagnosis of covid-19 facilitates management of sars-cov-2 outbreak, Journal of Medical Virology (2020).

[9] H. C. Metsky, C. A. Freije, T.-S. F. Kosoko-Thoroddsen, P. C. Sabeti, C. Myhrvold, Crispr-based surveillance for covid-19 using genomically-comprehensive machine learning design, bioRxiv (2020).

[10] M. Wang, Q. Wu, W. Xu, B. Qiao, J. Wang, H. Zheng, S. Jiang, J. Mei, Z. Wu, Y. Deng, et al., Clinical diagnosis of 8274 samples with 2019-novel coronavirus in wuhan, medRxiv (2020).

[11] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, et al., A deep learning algorithm using ct images to screen for corona virus disease (covid-19), medRxiv (2020).

[12] J. Y. Kim, P. G. Choe, Y. Oh, K. J. Oh, J. Kim, S. J. Park, J. H. Park, H. K. Na, M.-d. Oh, The first case of 2019 novel coronavirus pneumonia imported into korea from wuhan, china: implication for infection prevention and control measures, Journal of Korean Medical Science 35 (5) (2020).

[13] W. R. Pearson, [5] rapid and sensitive sequence comparison with fastp and fasta (1990).

[14] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, Journal of molecular biology 215 (3) (1990) 403–410.

[15] L. Pinello, G. Lo Bosco, G.-C. Yuan, Applications of alignment-free methods in epigenomics, Briefings in Bioinformatics 15 (3) (2014) 419–430.

[16] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (4) (2003) 513–523.

[17] D. Bzhalava, J. Ekström, F. Lysholm, E. Hultin, H. Faust, B. Persson, M. Lehtinen, E.-M. de Villiers, J. Dillner, Phylogenetically diverse tt virus viremia among pregnant women, Virology 432 (2) (2012) 427–434.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[18] N. G. Nguyen, V. A. Tran, D. L. Ngo, D. Phan, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, M. Kubo, K. Satou, et al., Dna sequence classification by convolutional neural network, Journal of Biomedical Science and Engineering 9 (05) (2016) 280.

[19] R. Rizzo, A. Fiannaca, M. La Rosa, A. Urso, A deep learning approach to dna sequence classification, in: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Springer, 2015, pp. 129–140.

[20] A. Tampuu, Z. Bzhalava, J. Dillner, R. Vicente, Viraminer: Deep learning on raw dna sequences for identifying viral genomes in human samples, PloS one 14 (9) (2019).

[21] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, F. Sun, Identifying viruses from metagenomic data by deep learning, arXiv preprint arXiv:1806.07810 (2018).

[22] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbsnp: the ncbi database of genetic variation, Nucleic acids research 29 (1) (2001) 308–311.

[23] C. d. S. Ribeiro, M. Y. van Roode, G. B. Haringhuizen, M. P. Koopmans, E. Claassen, L. H. van de Burgwal, How ownership rights over microorganisms affect infectious disease control and innovation: a root-cause analysis of barriers to data sharing as experienced by key stakeholders, PloS one 13 (5) (2018).

[24] J. H. Simon, E. Claassen, C. E. Correa, A. D. Osterhaus, Managing severe acute respiratory syndrome (sars) intellectual property rights: the possible role of patent pooling, Bulletin of the World Health Organization 83 (2005) 707–710.

[25] C. d. S. Ribeiro, M. P. Koopmans, G. B. Haringhuizen, Threats to timely sharing of pathogen sequence data, Science 362 (6413) (2018) 404–406.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.990242; this version posted March 14, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[26] A. Lopez-Rincon, A. Tonda, M. Elati, O. Schwander, B. Piwowarski, P. Gallinari, Evolutionary optimization of convolutional neural networks for cancer mirna biomarkers classification, Applied Soft Computing 65 (2018) 91–100.

[27] A. Vabret, T. Mourez, S. Gouarin, J. Petitjean, F. Freymuth, An outbreak of coronavirus oc43 respiratory infection in normandy, france, Clinical infectious diseases 36 (8) (2003) 985–989.

[28] L.-J. Cui, C. Zhang, T. Zhang, R.-J. Lu, Z.-D. Xie, L.-L. Zhang, C.-Y. Liu, W.-M. Zhou, L. Ruan, X.-J. Ma, et al., Human coronaviruses hcov-nl63 and hcov-hku1 in hospitalized children with acute respiratory infections in beijing, china, Advances in virology 2011 (2011).

[29] F. Zeng, C. Chan, M. Chan, J. Chen, K. Chow, C. Hon, K. Hui, J. Li, V. Li, C. Wang, et al., The complete genome sequence of severe acute respiratory syndrome coronavirus strain hku-39849 (hk-39), Experimental Biology and Medicine 228 (7) (2003) 866–873.

[30] I. Mizrachi, Genbank: the nucleotide sequence database, The NCBI Handbook [Internet], updated 22 (2007).

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[32] F. Chollet, et al., Keras, `https://keras.io` (2015).

[33] A.-M. Šimundić, Measures of diagnostic accuracy: basic definitions, Ejifcc 19 (4) (2009) 203.

[34] J. N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (9) (2010) 1315–1316.