Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 24 (2021) 100577

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Identification of SARS-CoV-2 origin: Using Ngrams, principal component analysis and Random Forest algorithm

Hamoucha El Boujnouni [a,*], Mohamed Rahouti [a], Mohamed El Boujnouni [b]

[a] Research Center of Plant and Microbial Biotechnologies, Biodiversity, and Environment, Faculty of Sciences, Mohammed V University in Rabat, PO Box 1014, Morocco
[b] Laboratory of Information Technologies, National School of Applied Sciences, Chouaib Doukkali University in El Jadida, PO Box 1166, Morocco

ARTICLE INFO

Keywords:
Bioinformatics
Genomes
SARS-CoV-2
COVID-19
Ngrams
Principal component analysis
Random forest algorithm

ABSTRACT

COVID-19 is an infectious disease caused by the newly discovered SARS-CoV-2 virus. This virus causes a respiratory tract infection, symptoms include dry cough, fever, tiredness and in more severe cases, breathing difficulty. SARS-CoV-2 is an extremely contagious virus that is spreading rapidly all over the world and the scientific community is working tirelessly to find an effective treatment. This paper aims to determine the origin of this virus by comparing its nucleic acid sequence with all members of the coronaviridae family. This study uses a new approach based on the combination of three powerful techniques which are: Ngrams (For text categorization), Principal Component Analysis (For dimensionality reduction) and Random Forest algorithm (For supervised classification). The experimental results have shown that a large set of SARS-CoV-2 genomes, collected from different locations around the world, present significant similarities to those found in pangolins. This finding confirms some previous results obtained by other methods, which also suggest that pangolins should be considered as possible hosts in the emergence of the new coronavirus.

1. Introduction

A novel member of human coronavirus, newly identified in Wuhan, China, in late December 2019, officially named as SARS-CoV-2 (Severe Acute Respiratory Syndrome COronaVirus 2) by the International Committee on Taxonomy of Viruses [1]. It is a new strain of RNA viruses that has not been previously identified in humans. The disease caused by this virus was named COVID-19 (COronaVIrus Disease 2019) by the World Health Organization. Its most common symptoms are fever [2], tiredness [3], and dry cough [4]. Some patients may have aches and pains [4], nasal congestion, runny nose [5], sore throat, diarrhea [4], or loss of taste or smell [5]. The disease can spread through the respiratory droplets produced when an infected person coughs or sneezes [6]. These droplets land on objects and surfaces around the person. Other people may acquire SARS-CoV-2 by touching these contaminated objects or surfaces, then touching their eyes, nose, or mouth [6].

Several scientists around the world have carried out researches to fight against this virus, in particular by determining its origin(s), symptoms, causes, diagnosis, treatment, etc. Understanding the origin of this virus is a very important issue not only to identify the causes of this pandemic and to avoid future ones, but it has serious consequences on our interactions with ecosystems, on breeding of wild and domestic animals, on some laboratory practices, etc. Concerning the origin of this virus, many research papers have been published since the appearance of this pandemic. These studies, which were based on theoretical and experimental approaches, have shown various origins of SARS-COV-2. For example, Paraskevis et al. [7] found that the new corona virus is most closely related with the BatCoV RaTG13 detected in bats. In their study they used various methods and software: RDP4, Simplot v3.5.1 and phylogenetic analysis with maximum likelihood and Bayesian methods. The same result was found by Zhou et al. [8], who showed that SARS-COV-2 is 96% identical at the whole-genome level to a bat coronavirus. They used an experimental approach based on virus isolation, cell infection, electron microscopy and neutralization assay, followed by RNA extraction and PCR (Polymerase Chain Reaction), serological test, examination of ACE2 receptor for 2019-nCoV infection. Also, High-throughput sequencing, pathogen screening and genome assembly and phylogenetic analysis were used. Another study was carried out by Luan et al. [9] and suggested that Bovidae and Cricetidae should be included in the screening of intermediate hosts for SARS-CoV-2. Their working method started by sequence analysis of ACE2 followed by structure simulation of ACE2-RBD complex using SWISS-MODEL online

* Corresponding author.
E-mail address: helboujnouni@gmail.com (H. El Boujnouni).

https://doi.org/10.1016/j.imu.2021.100577
Received 24 January 2021; Received in revised form 12 April 2021; Accepted 13 April 2021
Available online 20 April 2021

Table 1
Example of extracting Ngrams from a random genome "TGATTTGATGCAA".

	$N = 3$		$N = 6$		$N = 9$	
	Ngrams	occurrences	Ngrams	occurrences	Ngrams	occurrences
1	TGA	2	ATGCAA	1	ATTTGATGC	1
2	GAT	2	TTGATG	1	TTTGATGCA	1
3	ATT	1	TGATGC	1	TGATTTGAT	1
4	TTT	1	ATTTGA	1	TTGATGCAA	1
5	GCA	1	GATTTG	1	GATTTGATG	1
6	ATG	1	TTTGAT	1		
7	TGC	1	TGATTT	1		
8	TTG	1	GATGCA	1		
9	CAA	1				

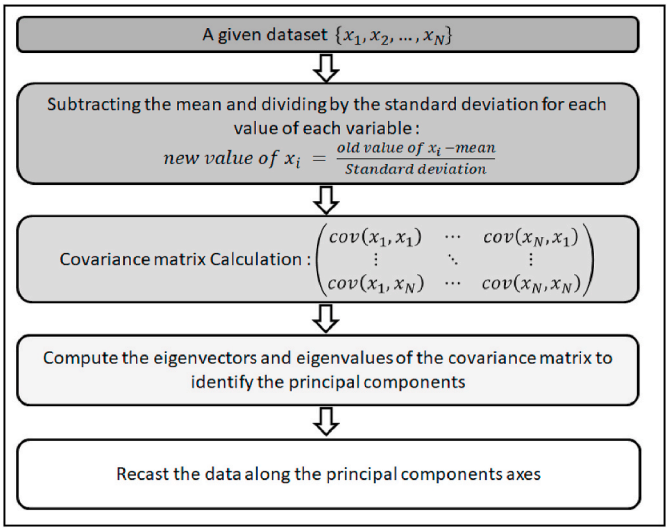

Fig. 1. Dimensionality reduction with Principal Component Analysis.

server and Chimera software version 1.14. In another research, Qiu et al. [10] predicted that SARS-CoV-2 tends to utilize ACE2s of various mammals (considered as intermediate hosts), except murines, and some birds, such as pigeon. To achieve these results they combined phylogenetic analysis and critical site marking. Another relevant research about the intermediate animal sources of this virus was conducted by several scientists who agreed that pangolin should be considered as an interhost of the novel coronavirus. They employed various strategies to investigate the origin of SARS-CoV-2.For example; Wong et al. [11] used genomic analysis with bbduk.sh v38.71, bbmap.sh v38.71 and Vir MAP. Lam et al. [12] applied the sequencing of pangolins genomes followed by phylogenetic and recombination analyses. Zhang et al. [13] used genome assembly and gene prediction followed by phylogeny. and Han [14] applied phylogenitic analysis of RBD. In another research paper written by Shi et al. [15] who performed an experimental study composed of four successive steps (i) isolation of SARS-CoV-2 (ii) inoculation of SARS-CoV-2 to (ferrets, cats, dogs, pigs, chickens and ducks) (iii) quantification by PCR in different organ and tissues and finally detection of antibody against SARS-CoV-2 by ELISA and neutralization assay. The authors showed that SARS-CoV-2 replicates poorly in dogs, pigs, chickens, and ducks, but ferrets and cats are permissive to infection which can act as intermediate hosts of this virus.

This paper presents a new method to ascertain the origin of SARS-CoV-2 by comparing its genomes with other viruses from Coronaviridae family [16]. The analysis is performed through three powerful techniques from machine learning and data mining that work successively: The first one is Ngrams [17]; its role is to extract relevant information from a given genome sequence and to present it in an exploitable numerical form. The second one is principal component analysis [18], it reduces the dimensionality of the extracted information by projecting them onto a lower-dimensional space, and the last one is Random Forest (RF) algorithm [19] that will be used to classify the reduced information and to find biological homology between different sequences.

2. Materials and methods

2.1. Ngrams analysis of genomes

In language modeling, Ngrams [17] are sequences of characters or words extracted from a text. It can be divided into two categories: Character based and Word based. The former is a set of *N* consecutive characters extracted from a word. The main motivation behind this approach is that similar words will have a high proportion of Ngrams in common. The second is a set of *N*consecutive words extracted from text. Word level Ngrams models are quite robust for modeling language statistically as well as for information retrieval. Ngrams has been applied in numerous medical and biological fields such as: Protein classification [20], Prediction of human immunodeficiency [21], interpreting the Hidden Information of DNA Sequences [22], etc. Table 1 shows the result of applying Ngrams on a random sequence "TGATTTGATGCAA" with different values of $N = \{3, 6, 9\}$.

2.2. Dimensionality reduction with principal component analysis

As is known, the possible number of nitrogenous bases in a given DNA or RNA is four (A, G, C, T or A, G, C, U).Statistically, Ngrams with $N = 3$ will give $4^3 = 64$ possible 3-g and when *N* equals to 6 the total possible 6-g grows exponentially and becomes $4^6 = 4096$, and if we further increase the value of*N* to 9 then the number of 9-g reaches $4^9 = 262144$.This rapid increase will slow down the performance of RF (next process) because this classifier must separate objects with very large numbers of attributes. The solution is to transform this large set of attributes into a smaller one that still contains most of the information. This task can be performed by a well-known technique called principal component analysis abbreviated with the acronym (PCA). It's a commonly used method for dimensionality reduction invented by Karl Pearson [18]. It projects each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much

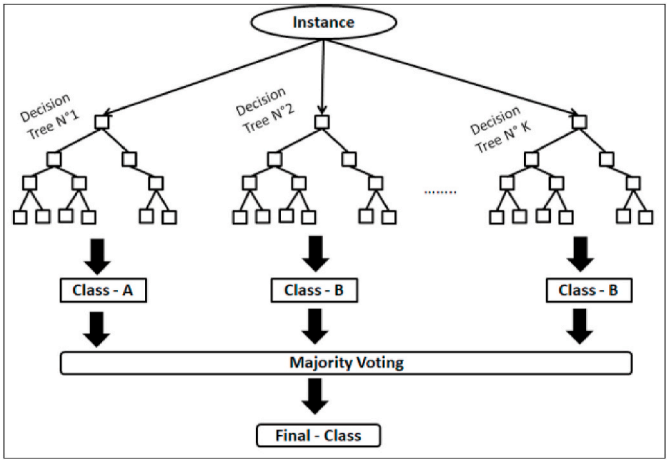

Fig. 2. Binary classification (Class-A or Class-B) with a RF. consisting of *K* decision trees.

of the data's variation as possible.Fig. 1 describes the basic steps needed to apply PCA to a given dataset that contains *N*samples $\{x_1, x_2, \ldots, x_N\}$.

2.3. Comparison of genomes using RF algorithm

Among the panoply of algorithms used in machine learning we chose RF classifier to identify the origin of COVID19. This machine learning technique was first introduced by Leo Breiman [19] who was inspired by the work of Amit and Geman [23]. It is a robust algorithm that consists of a large number of individual decision trees that operate as an ensemble. A decision tree is a non-parametric supervised learning method used for classification and regression. It is a flowchart similar to a tree structure, where each internal node represents a test on an attribute, each branch represents a result of the test, and each terminal node holds a class label. The advantage of RF algorithm is a prediction by committee which is more accurate and stable than that of any individual tree. Akin to a decision tree, RF can be used for classification (the output is the mode of the predictions), regression (the output is the mean of the predictions). In bioinformatics, RF is used intensively in many fields such as: Chronological age prediction based on DNA [24], Identification of species based on DNA barcode [25], outcome prediction in oesophageal cancer [26]. Fig. 2 shows an example of a RF that consists of *K* decision trees used for a classification task. It can be seen from this figure that each tree produces an individual decision (Class A or B), then RF merges them together to get a more accurate and stable prediction.

2.4. Datasets

This study was conducted on a large dataset that consists of two parts: The former contains the complete coronaviridae nucleic acid sequences represented with2649 genomes belonging to 96 species. The second includes10313SARS-CoV-2 genomes collected from homo-sapiens from different locations around the world. The genomes used in this study correspond to the firsts RNA sequences of SARS-COV-2 collected before any mutations were found. Both parts have been downloaded from the National Center for Biotechnology Information [16]. A detailed description of the dataset used to identify SARS-CoV-2 is given in Table 2.

2.5. Experimental protocol

The experimental protocol was performed through five steps as shown in Fig. 3. The first step is gathering which aims to collect the genomes of all Coronaviridae viruses. The second step is a preprocessing of the set of genomes previously collected. This step starts by extracting

Table 2
The detailed content of the experimental dataset of coronaviridae family.

Alphacoronavirus 1	Coronavirus BtRt-BetaCoVGX2018	Night heron coronavirus HKU19
Alphacoronavirus Bat-CoVP. kuhliiItaly206645-412011	Coronavirus BtSk-AlphaCoVGX2018A	Nyctalus velutinus alphacoronavirus SC-2013
Alphacoronavirus Bat-CoVP. kuhliiItaly206679-32010	Coronavirus BtSk-AlphaCoVGX2018B	Pangolin coronavirus
Alphacoronavirus Bat-CoVP.kuhliiItaly3398-192015	Coronavirus BtSk-AlphaCoVGX2018C	Pipistrellus abramus bat coronavirus HKU5-related
Alphacoronavirus BtMs-AlphaCoVGS2013	Coronavirus BtSk-AlphaCoVGX2018D	Pipistrellus bat coronavirus HKU5
Alphacoronavirus MinkChina12016	Coronavirus HKU15	Porcine epidemic diarrhea virus
Alphacoronavirus sp.	Coronavirus cya-BetaCoV2019	Quail deltacoronavirus
Avian coronavirus	Coronavirus cyb-BetaCoV2019	Rabbit coronavirus HKU14
Bat Hp-betacoronavirus Zhejiang2013	Coronavirus cyc-BetaCoV2019	Rhinolophus affinis bat coronavirus HKU2-related
Bat alphacoronavirus	Erinaceus hedgehog coronavirus HKU31	Rhinolophus bat coronavirus HKU2
Bat coronavirus	Feline alphacoronavirus 1	Rhinolophus bat coronavirus HKU32
Bat coronavirus 1A	Ferret coronavirus	*Rhinolophus ferrumequinum* alphacoronavirus HuB-2013
Bat coronavirus BM48-31BGR2008	Hedgehog coronavirus 1	Rodent coronavirus
Bat coronavirus CDPHE15	Hipposideros pomona bat coronavirus CHB25	Rousettus aegyptiacus bat coronavirus 229E-related
Bat coronavirus HKU10	Hipposideros pomona bat coronavirus HKU10-related	Rousettus bat coronavirus GCCDC1
Beluga whale coronavirus SW1	Human coronavirus 229^E	Rousettus bat coronavirus HKU9
Betacoronavirus 1	Human coronavirus HKU1	SARS-like coronavirus WIV16
Betacoronavirus sp.	Human coronavirus NL63	Scotophilus bat coronavirus 512
Bottlenose dolphin coronavirus	Hypsugo bat coronavirus HKU25	Scotophilus kuhlii bat coronavirus 512-related
BtRf-AlphaCoVYN2012	Lucheng Rn rat coronavirus	Shrew coronavirus
Bulbul coronavirus HKU11	Magpie-robin coronavirus HKU18	Sparrow coronavirus HKU17
Camel coronavirus HKU23	Middle East respiratory syndrome-related coronavirus	Sparrow deltacoronavirus
Canada goose coronavirus	Miniopterus bat coronavirus 1	Swine acute diarrhea syndrome coronavirus
China Rattus coronavirus HKU24	Miniopterus bat coronavirus HKU8	Swine acute diarrhea syndrome related coronavirus
Common moorhen coronavirus HKU21	Miniopterus pusillus bat coronavirus HKU8-related	Swine enteric alphacoronavirus
Coronavirus AcCoV-JC34	Miniopterus schreibersii bat coronavirus 1-related	Thrush coronavirus HKU12-600
Coronavirus BtRl-BetaCoVSC2018	Mink coronavirus 1	Tylonycteris bat coronavirus HKU33
Coronavirus BtRs-AlphaCoVYN2018	Mink coronavirus strain WD1133	Tylonycteris bat coronavirus HKU4
Coronavirus BtRs-BetaCoVYN2018A	Munia coronavirus HKU13	Tylonycteris pachypus bat coronavirus HKU4-related
Coronavirus BtRs-BetaCoVYN2018B	Murine coronavirus	Wencheng Sm shrew coronavirus
Coronavirus BtRs-BetaCoVYN2018C		White-eye coronavirus HKU16

(*continued on next page*)

Table 2 (*continued*)

Alphacoronavirus 1	Coronavirus BtRt-BetaCoVGX2018	Night heron coronavirus HKU19
	Myotis ricketti alphacoronavirus Sax-2011	
Coronavirus BtRs-BetaCoVYN2018D	NL63-related bat coronavirus strain BtKYNL63-9b	Wigeon coronavirus HKU20

Ngrams and their frequencies from each genome using different values of *N*.The values are chosen in the interval $\{2,3,4\}$, an example of the extraction was described in Table 1. The Ngrams that are shared between different genomes will form a common basis. Next, the Ngrams representation of each genome will be expressed on this common basis. The third step is a dimensionality reduction; it begins by normalizing the dataset, then it uses principal component analysis to reduce its dimension. In this experiment, the cumulative explained variance was chosen to be less than 95%. The fourth step is an automatic learning process in which Coronaviridae family (except SARS-COV-2) will be used to train RF algorithm and to tune its hyper-parameters [27]. The last step uses the best model of RF to identify the origin of 10313 of SARS-CoV-2 samples. Fig. 3 describes in detail the procedure followed to identify SARS-CoV-2.

The experiment was conducted using R as a programming language and five packages namely: "seqinr" to read the nucleic acid sequences of different viruses stored in fasta format, "ngram" to construct the vectors of Ngrams from different genomes, "stats" to reduce the dimension of the vectors of Ngrams using PCA, "randomForest" to learn the dataset of coronaviridae with RF and to find the possible origin(s) of SARS-CoV-2 and "graphics" to represent graphically the statistical results. The experiments are conducted on a computer with an i5-7200U CPU @ 2.50 GHz (4CPUs) having 12 GB of RAM.

3. Result and discussion

3.1. Ngrams extraction

Table 3 shows an example of the application of Ngrams with N = $2,3,4$ on SARS-COV-2. It can be seen that, when$N = 2$, the sequence of this virus is represented with $4^2 = 16$ attributes that are the occurrences of all possible 2-g that can be found in SARS-COV-2 genome (e.g "AA","TG", etc.). Also, it can be noted that, the occurrence of a given 2-g is very high. This can be explained by the fact that it is highly probable to find a sequence consists of only two specific nucleotides in a given genome. When N increases $N = \{3,4\}$ the size of the vectors that represent the genomes grows $(4^3, 4^4)$and the occurrences become lower. However, the description of the nucleic acid sequence becomes more accurate.

3.2. Dimensionality reduction using PCA

Figs. 4–6 show the cumulative proportion of variance explained (CPVE) by each principal component with respect to the value of N which varies from 2 to 4. The horizontal lines with red color represent thresholds that correspond to a CPVE equals to 95%. They will be used to decide on the number of principal components to keep to adequately reducing the dimension of our dataset. Take, for example, Fig. 4 where Ngram with $N = 2$ was used to extract features from genomes. As said before, the number of possible Ngrams is $4^2 = 16$by consequence each genome will be represented by a vector of 16 values. With PCA, we can reduce this number by computing the principal components and using them to perform a change of basis on the dataset. It can be seen through this figure that the first two principal components explain about 85.5% of the total variation in the dataset. Thus, the new representation of each genome becomes a vector with only two components. The same approach was applied when $N = 3$ (Fig. 5) and $N = 4$ (Fig. 6) which allow as to represent each genome with a vector of, respectively,4 values instead of 64 and 7 values instead of 256.Table 4 shows the result of the application of Ngrams followed by PCA on the same SARS-COV-2 genome represented with Ngrams in Table 3. It can be seen that PCA has successfully reduces the number of attributes obtained by Ngrams which will ease the task of RF.

3.3. Training RF and hyperparameters selection using a grid search

The identification of SARS-CoV-2 origin with our method requires a preliminary step in which the optimal values of the hyperparameters of RF algorithm must be determined [27]. These values correspond to the maximum of classification accuracy of coronaviridae members (2649 samples). In this training step, a grid search space for tuning those hyperparameters that takes into account the influence of the number of N-grams and the effect of PCA was performed. The hyperparameters concerned are:

- ntree: The number of trees in RF algorithm, it is obvious that larger number of trees produce more stable models, but require more memory and a longer run time. This hyperparameter will be varied in the interval$\{3^1, 3^2, 3^3\}$.

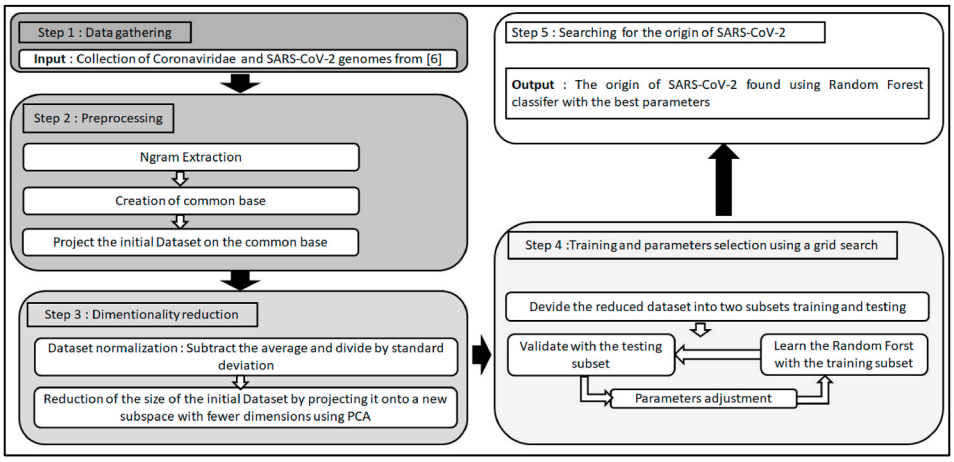

Fig. 3. Flow diagram for SARS-CoV-2 identification. Caption details the different steps from the inputs (SARS-CoV-2 and Coronaviridae complete genomes) to the output (The possible origin of SARS-CoV-2).

Table 3
The application of N-grams with $N = 2, 3, 4$ on a sample of SARS-COV-2.

Species	Value of *N*	Ngrams
SARS-COV-2	2	3213,2587,2843,2303,2368,1989,2085,2072,2020,1739,1611,1410,1166,1089,883,437
	3	1005,819,857,888,874,769,722,700,759,630,810,620,703,716,736,579,553,674,552,536, 606,547,537,617,454,485,605,439,556,550,520,507,470,515,494,371,426,438,458,299,341,340,268,354,372,279,295,340,285,210,169, 163,187,114,468,327,223,112,96,96,133,88,76,74
	4	330,245,302,246,289,255,231,199,250,239,216,244,206,174,164,187,203,161,176,158,168,170,164,78,85,157,138,137,144,109,94,78,80, 257,234,189,273,208,207,209,204,258,217,186,241,195,190,192,166,169,156,213,196,188,127,141,151,103,195,157,98,173,142,151,161, 113,164,117,95,125,98,102,100,89,138,108,105,103,81,98,87,71,45,191,195,210,179,214,165,199,153,183,128,156,172,130,145, 163,120,127,149,130,121,157,119,114,167,107,133,115,116,91,131,89,76,77,70,52,40,285,207,168,146,179,143,138,167,135,124,122, 143,107,110,117,130,112,109,81,84,93,90,85,81,78,97,67,62,212,166,159,121,138,131,118,80,98,117,79,63,29,52,42,21,137,135,94,111, 49,80,97,56,42,34,32,61,65,56,75,293,82,47,37,129,91,101,62,48,32,30,29,29,200,20,29,28,53,38,31,19,109,26,15,111,53,17,41,30,22,77, 45,17,17,21,16,249,22,244,261,153,95,54, 49,27,33,257,115,115,15,222,37,54,94,55,22,13,48,32,61,95,17,10,58,38,13,12,29,13,14,32,14,21,37,14,10,22,20

- mtry: Number of features (predictors) randomly sampled as candidates at each split, its default value is the round of the square root of the number of features [27]. To be more careful, mtry will be varied in an interval constructed by three values that are: (square root of total number of all predictors), (half of this square root value), and (twice of the square root value).

Table 5 shows a sensitivity analysis on the controlling hyper-parameters of RF algorithm taking into account the effects of *N* in N-grams and the use or not of PCA. For each combination of parameters, the table gives the accuracy of RF to learn the coronaviridae family (except SARS-COV-2). The classification accuracy is evaluated using *k* fold cross validation with $k = 10$ [28]. It means that the dataset was divided randomly into ten subsets. Each part was used as a testing subset for the RF trained on other nine subsets. The average of the 10 error terms obtained will be taken. This process will be repeated 10 times and the final result will be re-averaged the table also shows a crucial information which is the run-time required to search for the best hyper-parameters of RF in a given intervals of ntree and mtry.

From Table 5, it can be seen that the best accuracy is always achieved when ntree equals27 which corresponds to the maximum value in its interval of variation$\{3, 9, 27\}$. It's an expected result because when the number of trees increases the classification results become more accurate (many trees participate in the collective decision). However, this promising result is not without downsides as it is clearly visible on the run-time column level. Besides, it can be observed that the best accuracy doesn't correspond constantly to the value by default of mtry. Consequently, it was justifiable to vary mtry in an interval. Moreover, it can be highlighted that the accuracy increases when the parameter of Ngrams (*N*) increases. This can be explained by the fact that Ngrams with a large *N*will extract more information from the sequence of genomes. But, this satisfactory result isn't without cost in terms of run-time which increases with *N*(ntree and mtry are constant). Additionally, the effect of PCA is very remarkable. It allowed the reduction of the number of the features extracted from each genome while achieving comparable results in terms of accuracy to those found with all features (without PCA).Also, the use of PCA allows to reduce the run-time needed to tune the hyperparameters of RF.

In the next step, the best hyper-parameters found using the grid search will be used to build a robust RF to identify the origin of SARS-CoV-2.

3.4. Searching for the origin of SARS-CoV-2

Figs. 7–9 show the origin(s) of SARS-CoV-2 without use of PCA. It can be seen that when $N = 2$, RF gives a one origin to 10313SARS-COV-2

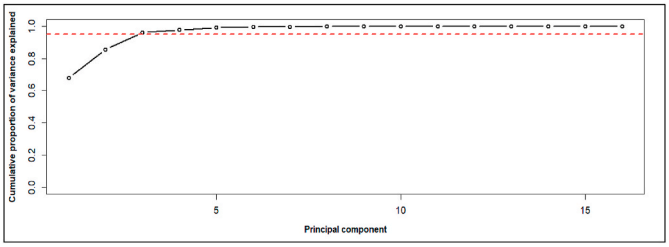

Fig. 4. Cumulative proportion of explained variance by each principal component when.$N = 2$

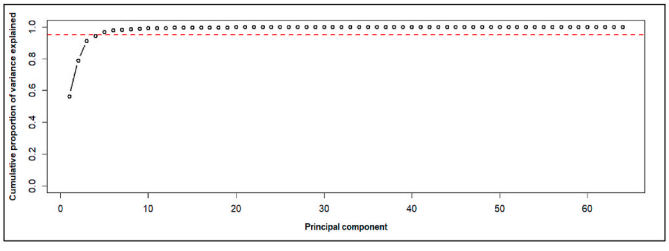

Fig. 5. Cumulative proportion of explained variance by each principal component when.$N = 3$

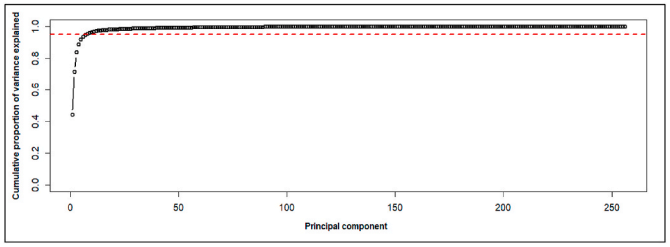

Fig. 6. Cumulative proportion of explained variance by each principal component when.$N = 4$

Table 4
Application of N-grams with $N = 2, 3, 4$followed by PCA on a sample of SARS-COV-2.

Species	Value of *N*	Ngrams
SARS-COV-2	2	0.4831962439710,-0.7451479383900
	3	0.6808870079172,-1.7158131143720,0.4418335738713,0.0570661968030
	4	−1.2299304981050,3.9345058563130,0.6165315842700,-0.2268199446313,-0.0771337837728,-0.1955765124330,-0.1608062654392

Table 5
Searching for the best values of N-grams, ntree, mtry (Best accuracy is marked in bold) with and without PCA.

	N-grams	Number of features	ntree	mtry	Accuracy of the model	Time in seconds to search the best ntree and mtry values	Best parameters (ntree, mtry)
with PCA	2	2	3	1	0.9512012	2.439439	(27,1)
			3	2	0.9615666		
			9	1	0.9794964	2.482191	
			9	2	0.9641108		
			27	**1**	**0.9820407**	2.911288	
			27	2	0.9730757		
	3	4	3	1	0.9692985	2.560438	(27,2)
			3	2	0.9666454		
			3	4	0.9666754		
				1	0.9833427	3.170673	
			9	2	0.9846146		
			9	4	0.9833425		
			27	1	0.9858970	3.843025	
			27	**2**	**0.9871789**		
			27	4	0.9794962		
	4	7	3	2	0.9705607	2.627833	(27,2)
			3	3	0.9678981		
			3	6	0.9692296		
			9	2	0.9884709	3.070224	
			9	3	0.9846048		
			9	6	0.9884609		
			27	**2**	**0.9923074**	4.028982	
			27	3	0.9910253		
			27	6	0.9884906		
Without PCA	2	16	3	2	0.9705015	2.632825	(27,2)
			3	4	0.9654128		
			3	8	0.9782538		
			9	2	0.9909957	2.972137	
			9	4	0.9846247		
			9	8	0.9884709		
			27	**2**	**0.9948617**	4.44932	
			27	4	0.9897532		
			27	8	0.9859265		
	3	64	3	4	0.9846244	3.49078	(27,4)
			3	8	0.9679769		
			3	16	0.9718233		
			9	4	0.9922875	3.895493	
			9	8	0.9859462		
			9	16	0.9871789		
			27	**4**	**0.9935894**	6.499027	
			27	8	0.9923074		
			27	16	0.9897529		
	4	256	3	8	0.9705806	6.512125	(27,8)
			3	16	0.9589921		
			3	32	0.9743876		
			9	8	0.9833917	8.84526	
			9	16	0.9884906		
			9	32	0.9897333		
			27	**8**	**0.9961537**	15.33804	
			27	16	0.9897629		
			27	32	0.9948617		

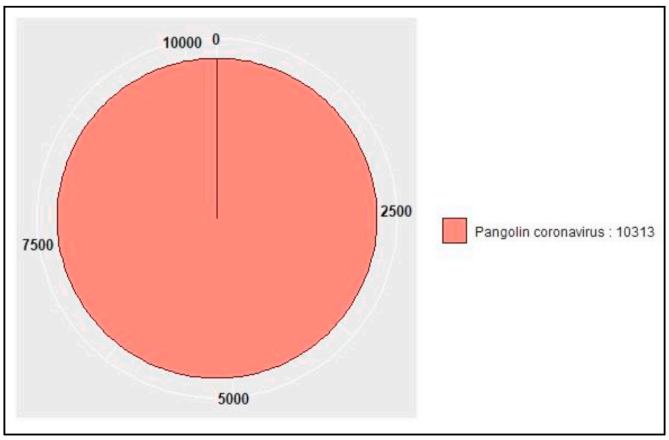

Fig. 7. Origin of 10313 samples of SARS-CoV-2 with$N = 2$, ntree = 27and mtry = 2without PCA.

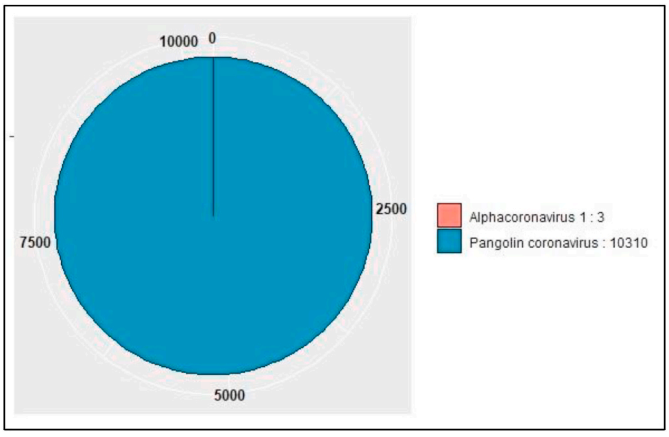

Fig. 8. Origins of 10313 samples of SARS-CoV-2 with$N = 3$, ntree = 27and mtry = 4without PCA.

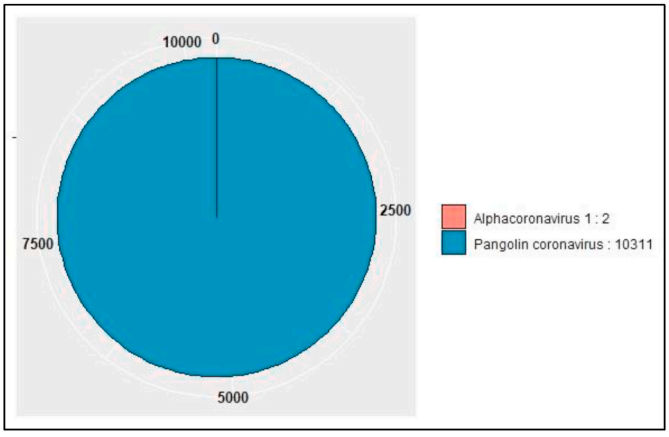

Fig. 9. Origins of 10313 samples of SARS-CoV-2 with$N = 4$, ntree = 27and mtry = 8 without PCA.

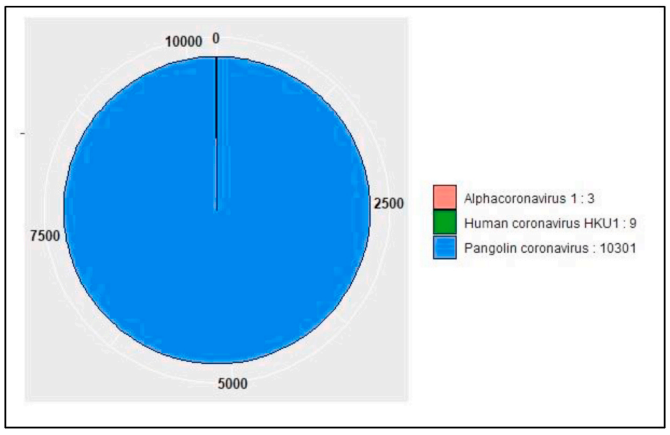

Fig. 10. Origin(s) of 10313 samples of SARS-CoV-2 using PCA. with $N = 2$, ntree = 27 and mtry = 1.

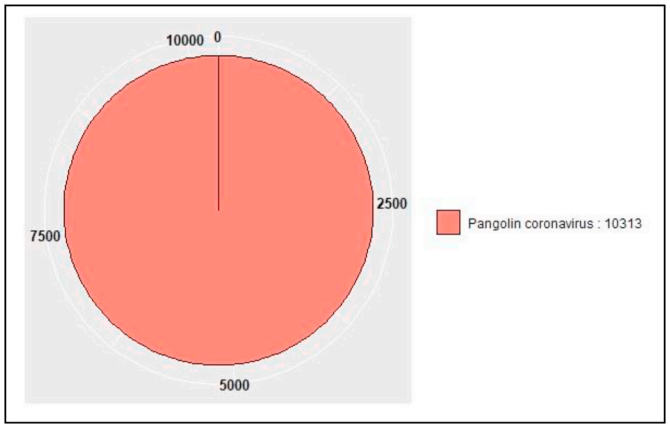

Fig. 11. Origin of 10313 samples of SARS-CoV-2 using PCA. with $N = 3$, ntree = 27 and mtry = 2.

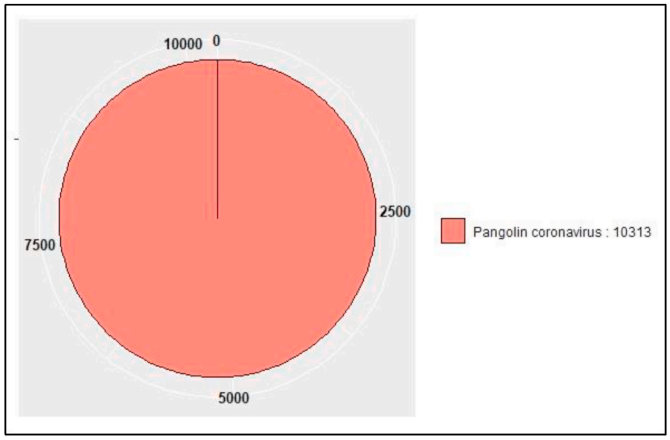

Fig. 12. Origin of 10313 samples of SARS-CoV-2 using PCA with $N = 4$, ntree = 27and mtry = 2.

genomes which is Pangolin. With the increase of$N$, a new origin appears which is Alphacoronavirus-1 with a very negligible representation. This degradation in RF's quality is due to fact that when N increases this classifier will find difficulties to select the best features to split a node (to build the trees), hence the usefulness of the preprocessing with PCA.

Figs. 10–12 show the origin(s) of SARS-CoV-2 using PCA with the best values of the hyperparameters of RF. The value of N is variable in the interval$\{2, 3, 4\}$. It can be observed that when$N = 2$, RF yields 3 origins to SARS-COV-2 namely: Pangolin (10301), Alphacoronavirus-1 (3) and Human Coronavirus HKU1 (9). When $N = 3$ or4 RF gives only one origin to 10313 samples of SARS-CoV-2 which is pangolin. This can be considered as an improvement if we assume that a one virus cannot

have more than a one origin. Also, because PCA reduces the dimensionality of each genome (e.g Table 3 Vs Table 4) it will also reduce the run-time required to predict the origin of SARS-COV-2 with RF. In conclusion, the use of Ngrams with high values of *N*followed by PCA and RF helps to better identify the origin of SARS-CoV-2 in a limited time.

4. Conclusion

In this paper, a new method based one three powerful techniques that operate successfully were proposed to identify SARS-CoV-2's origin. The former is Ngram, it is widely used in text categorization; it creates feature vectors from different nucleic acid sequences. The second is principal component analysis; it's used to cleverly reduce the large quantity of information generated by Ngrams while maintaining as much of the data's variation as possible. The last technique is a supervised machine learning algorithm called RF that is proposed to classify these reduced vectors and to detect similarities between the genome of SARS-CoV-2 and other coronaviridae viruses. This experiment conducted on a large dataset of genomes using our method has shown that SARS-CoV-2 was originated from pangolins. This study confirms some previous findings that indicate that pangolins were the culprits in COVID-19 transmission to humans and refutes others that suggest other zoonotic origins like bats, ferrets, cats, etc.

Declaration of competing interest

The authors state that there is no conflict of interest.

Acknowledgement

None. No funding to declare.

References

[1] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. On the origin and continuing evolution of SARS-CoV-2. National Science Review 2020;7(6):1012–23. https://doi.org/10.1093/nsr/nwaa036.

[2] Aloysius MM, Thatti A, Gupta A, Sharma N, Bansal P, Goyal H. COVID-19 presenting as acute pancreatitis. Pancreatology 2020;20(5):1026–7. https://doi.org/10.1016/j.pan.2020.05.003.

[3] Ashktorab H, Pizuorno A, Oskroch G, Alma Fierro N, Sherif ZA, Brim H. COVID-19 in Latin America: symptoms, morbidities, and gastrointestinal manifestations. Gastroenterology 2021;160(3):938–40. https://doi.org/10.1053/j.gastro.2020.10.033.

[4] Han R, Huang L, Jiang H, Dong J, Peng H, Zhang D. Early clinical and CT manifestations of coronavirus disease 2019 (COVID-19) pneumonia. Am J Roentgenol 2020;215(2).

[5] Meng X, Deng Y, Dai Z, Meng Z. COVID-19 and anosmia: a review based on up-to-date knowledge. Am J Otolaryngol 2020;41(5):102581. https://doi.org/10.1016/j.amjoto.2020.102581.

[6] Dhand R, Li J. Coughs and sneezes: their role in transmission of respiratory viral infections, including SARS-CoV-2. Am J Respir Crit Care Med 2020;202(5). https://doi.org/10.1164/rccm.202004-1263PP.

[7] Paraskevis D, Kostaki E-G, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus(2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol 2020;79:104212. https://doi.org/10.1016/j.meegid.2020.104212.

[8] Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3.

[9] Luan J, Jin X, Lu Y, Zhang L. SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. Journal ofMedical Virology 2020;92(9):1649–56. https://doi.org/10.1002/jmv.25817.

[10] Qiu Y, Zhao Y-B, Wang Q, Li J-Y, Zhou Z-J, Liao C-H, Ge X-Y. Predicting the angiotensin converting enzyme 2 (ACE2) utilizing capability as the receptor of SARS-CoV-2. Microb Infect 2020;22(4–5):221–5. https://doi.org/10.1016/j.micinf.2020.03.003.

[11] Wong M-C, Cregeen S-J-J, Ajami N-J, Petrosino J-F. Evidence of recombination in coronaviruses implicatingpangolinorigins of nCoV-2019, bioRxiv. 2020. https://doi.org/10.1101/2020.02.07.939207.

[12] Lam T-T-Y, Jia N, Zhang Y-W, Shum M-H-H, Jiang J-F, Zhu H-C, Tong Y-G, Shi Y-X, Ni X-B, Liao Y-S, Li W-J, Jiang B-G, Wei W, Yuan T-T, Zheng K, Cui X-M, Li J, Pei G-Q, Qiang X, Cheung W-Y-M, Li L-F, Sun F-F, Qin S, Huang J-C, Leung G-M, Holmes E-C, Hu Y-L, Guan Yand Cao W-C. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature 2020;583(7815):282–5. https://doi.org/10.1038/s41586-020-2169-0.

[13] Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol 2020;30(8):1578. https://doi.org/10.1016/j.cub.2020.03.022.

[14] Han G-Z. Pangolins harbor SARS-CoV-2-related coronaviruses. Trends Microbiol 2020. https://doi.org/10.1016/j.tim.2020.04.001.

[15] Shi J, Wen Z, Zhong G, Yang H, Wang C, Huang B, Liu R, He X, Shuai L, Sun Z, Zhao Y, Liu P, Liang L, Cui P, Wang J, Zhang X, Guan Y, Tan W, Wu G, Chen H, Bu Z. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS coronavirus 2. Science 2020;368(6494):1016–20. https://doi.org/10.1126/science.abb7015.

[16] National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/labs/virus.

[17] Cavnar W-B, Trenkle J-M. N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, 161-175, LasVegas, NV; 1994.

[18] Pearson K. On lines and planes of closest fit to systems of points in space. Phil Mag 1901;2(11):559–72. https://doi.org/10.1080/14786440109462720.

[19] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[20] Islam S-M-A, Heil B-J, Kearney Kearney C-M, Baker E-J. Protein classification using modified n-grams and skip-grams. Bioinformatics 2018;34(9):1481–7. https://doi.org/10.1093/bioinformatics/btx823.

[21] Masso M. Prediction of human immunodeficiency virus type 1 drug resistance: representation of targetsequence mutational patterns via an n-grams approach. In: Proceeding of the international conference on bioinformatics and biomedicine, philadelphia, PA, USA; 4-7 Oct. 2012.

[22] Le N-Q-K, Yapp E-K-Y, Nagasundaram N, Yeh H-Y. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fast text N-grams. Frontiers in Bioengineering and Biotechnology 2019;7(305):1–9. https://doi.org/10.3389/fbioe.2019.00305.

[23] Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Comput 1997;9(7):1545–88.

[24] Naue J, Hoefsloot H-C-J, Mook O-R-F, Rijlaarsdam-Hoekstra L, van der Zwalm M-C-H, Henneman P, Kloosterman A-D, Verschure P-J. Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression. Forensic Sci Int: Genetics 2017;31:19–28. https://doi.org/10.1016/j.fsigen.2017.07.015.

[25] Meher P-K, Sahu T-K, Rao A-R. Identification of species based on DNA barcode using k-mer feature vectorand Random forest classifier. Gene 2016;592(2):316–24. https://doi.org/10.1016/j.gene.2016.07.010.

[26] Paul D, Su R, Romain M, Sébastien V, Pierre Vand Isabelle G. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Comput Med Imag Graph 2017;60:42–9. https://doi.org/10.1016/j.compmedimag.2016.12.002.

[27] Probst P, Wright M-N, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowl Discov 2019. arXiv:1804.03515.

[28] Kohavi R. A study of CrossValidation and bootstrap for accuracy estimation and model selection. In: Proceeding of the 14th International joint conference on Artificial intelligence, vol. 2; August 1995. p. 1137–43.