A study and identification of COVID-19 viruses using N-grams with Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine

Mohamed El Boujnouni (med.elbouj@gmail.com)
National School of Applied Sciences Chouaib Doukkali University El Jadida - Morocco

Research Article

Keywords: Genomes, COVID19, N-grams, Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine

Posted Date: August 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-40344/v2

Abstract

Coronavirus disease 2019 or COVID-19 is a global health crisis caused by a virus officially named as severe acute respiratory syndrome coronavirus 2 and well known with the acronym (SARS-CoV-2). This very contagious illness has severely impacted people and business all over the world and scientists are trying so far to discover all useful information about it, including its potential origin(s) and inter-host(s). This study is a part of this scientific inquiry and it aims to identify precisely the origin(s) of a large set of genomes of SARS-COV-2 collected from different geographic locations in all over the world. This research is performed through the combination of five powerful techniques of machine learning (Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine) and a widely known tool of language modeling (N-grams). The experimental results have shown that the majority of techniques gave the same global results concerning the origin(s) and inter-host(s) of SARS-COV-2. These results demonstrated that this virus has one zoonotic source which is Pangolin.

Introduction

Nucleic acids are the biopolymers that carry all genetic information of living organisms. The two main classes of nucleic acids are deoxyribonucleic acid commonly named as (DNA) and ribonucleic acid known by the acronym (RNA). DNA is a double stranded that consists of four nucleotide bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine(C). While RNA is a single stranded that contains: Guanine, Uracil (instead of Thymine), Adenine, and Cytosine, denoted by the letters G, U, A, and C respectively.

In the fields of molecular biology and genetics, a genome is the complete set of genetic information of an organism. It consists of a large set of the aforementioned letters arranged in particular order (e.g Human genome is 3.2 billion of them). These letters contain instructions or genes that control all of the fundamental biological processes of life. A genome can be seen as a collection of information or a text written in a particular language of a simple alphabet, not with 26 letters but just four.

Biological sequence analysis, a subfield of bioinformatics and computational biology, aims at computationally process and decode the information stored in genomes. The analysis brings together several fields, from computer science to probability and statistics. Biological sequence analysis has many goals for example: Search of similarity between sequences of different organisms, Identification of intrinsic features of a sequence, Determination of sequence differences, Identification of molecular structure.

This paper falls within the scope of biological sequence analysis and has as a goal to search the origin(s) and the inter-host(s) of each genome of SARS-CoV-2 by comparing it with all members of Coronaviridae family. The experimental study is performed thought two steps that work successively: The first one uses N-grams; its role is to extract relevant information from a given sequence and to present it in a numeric form. The second one uses machine learning techniques to find biological homology between the genomes of different viruses.

The rest of this paper is organized as follows: Section 2 introduces briefly SARS-COV-2 and presents some pioneer previous researches about its possible origin(s) and inter-host(s). Section 3 gives a short presentation that includes: supervised machine learning, the five proposed techniques, and N-grams method. Section 4 gives the experimental protocol and the results of applying the aforementioned techniques to analyze the genomes of this virus. Finally, the conclusions are discussed in Section 5.

A Brief Overview About Coronaviruses, Its Origin(S) And Inter-host(S)

The virus that causes COVID-19 belongs to the family Coronaviridae, which includes a group of enveloped,
positive-sensed, single-stranded RNA viruses. This virus officially named as SARS-CoV-2 by International Committee on Taxonomy of Viruses [9] is rapidly spreading from its origin in Wuhan City of China to the rest of the world. This virus is mainly transmitted through droplets that are produced when an infected person coughs, sneezes, or exhales and that fell on floors or surfaces due to their heavy weight. A person can be infected by touching a contaminated surface and then his eyes, nose or mouth or inhaling directly these droplets. The symptoms of this disease are fever, dry cough, breathing difficulties, headache, nasal congestion, runny nose and pneumonia. This illness that has neither an approved drugs nor vaccines is significantly impacting people's health, businesses and the economy in all over the word. Unfortunately, the only ways to prevent it are to avoid being exposed directly to the virus and to the other people.

Potential origins and inter-hosts of this virus were discussed in many research papers. For example, researchers [12, 26] found that the SARS-CoV-2 showed a higher sequence homology to Bat-CoV-RaTG13 and stated that the origin of this virus is bats .Another paper [6] suggested that bovidae and cricetidae should be involved in the screening of intermediate hosts for SARS-CoV-2. In the same context [30] predicted that SARS-CoV-2 utilize ACE2s of various mammals, excluding murines, and some birds, such as pigeon as intermediate hosts. Several recent studies [11, 27, 28, 29] proposed that pangolins might be the intermediate hosts between bats and humans because of the similarity of the pangolin coronavirus to SARS-CoV-2.

A Brief Overview Of The Five Supervised Learning Algorithms And N-grams Method

Machine Learning is a subfield of Artificial Intelligence and is concerned with the development of techniques and methods which enable a computer to learn. Within the field of machine learning, there are two main types of tasks: supervised, and unsupervised. In the first, we train the machine using a dataset which is labeled (i.e the classes of the objects are known). For example, a dataset of genome sequences where each one is tagged with his specie by an expert in the domain. Typical fields of supervised learning are classification, regression, and time series analysis. In the second we train the machine using unlabeled dataset. For example, a dataset of genome sequences where each one is unknown. Typical fields of unsupervised learning are projection, clustering, density estimation or generative models.

This research paper focuses on supervised learning algorithms and uses five among them to search similarities between the genomes of different viruses. The first one is: Support Vector Machine (SVM), it can be used for classification, regression and prediction challenges. SVM was first introduced by Boser, Guyon, and

Vapnik in COLT-92 in 1992 [1]. The basic idea of SVM is simple: The algorithm creates a line which separates two different classes of objects represented in two dimensions. The equation of the separator is calculated mathematically with the objective of maximizing the margin between both of them. The decision boundary is used to classify new unknown objects basing on their positions (above or below the line). In medical domain SVM was used intensively in many application fields such as: Disease diagnosis [5], detection of medical disorders in MRI images [7], medical data classification [8], etc. The second is called Artificial Neural Networks (ANN), they are inspired by the way in which the human brain learns and processes information. They consist of a collection of connected units or nodes called artificial neurons, which approximately model the neurons in a biological brain. In bioinformatics ANN was used for many tasks like classification of biological data [13], identification of functional genetic variants and the prediction of traits [14], Breast cancer image classification [15], etc. The third is called Naïve Bayes (NB), it's a probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Also this technique was applied in medical domain for example: Diagnosis of Alzheimer's disease [16], classifier for DNA barcodes [17], gene expression data [18].The fourth is K-Nearest Neighbors (k-NN), it's the simplest and the fast classification method. It searched in a given dataset the k closest points to an unknown sample, with similarity defined by a distance function. Then it uses the classes of the points to identify the one of the unknown sample by searching the most frequent class among them. K-NN has found several applications in biology. For example, predicting the protein subchloroplast locations [19], gene expression [20] and gene expression in cancer diagnosis [21]. The fifth classifier is named Decision tree (DT), it uses a tree-like model of decisions in which each internal node represents a test on an attribute. This classifier was used in many biological applications such as missing value imputation in DNA microarray gene [23], analyzing gene expression data [24], and classification of pathogenic gene sequences [25].

In language modeling or Text categorization, N-grams [2] are a sequence of consecutive items in a text; it can be classified into two categories: Character N-grams and Word N-grams. The first category is a set of *N* consecutive characters extracted from a word; it's used for tasks like language identification and data compression. The second category is a sequence of consecutive words extracted from a text; it's used for a wide range of tasks like modeling language statistically as well as for information retrieval. As said before a nucleotide sequence can be seen as a text (or more precisely a word because of the absence of space) and then N-grams can be applied on it. Table 1 shows the result of applying this process on a random sequence "TGATGACTGATACA". N-gram was also applied in numerous medical and biological fields like: Analysis of RNA [3], clustering DNA sequences [4], genome data classification [10], etc.

Table 1: Example of extracting N-grams from a nucleic acid sequence "TGATGACTGATACA"

	N-Grams	occurrences	N-Grams	occurrences	N-Grams	occurrences
1	TGA	3	GATGAC	1	GATGACTGA	1
2	GAT	2	TGACTG	1	GACTGATAC	1
3	ACT	1	ATGACT	1	ACTGATACA	1
4	ATG	1	CTGATA	1	TGACTGATA	1
5	CTG	1	GATACA	1	ATGACTGAT	1
6	GAC	1	TGATGA	1	TGATGACTG	1
7	TAC	1	GACTGA	1		
8	ACA	1	ACTGAT	1		
9	ATA	1	TGATAC	1		

Experiments And Results

Datasets and experimental protocol

The experimental protocol is performed through three steps that are illustrated in figure 1:

- **Collecting the dataset**: A dataset containing 12962 genomes belonging to 96 families of coronavirus was collected from [22]. The dataset includes 10313 genomes of SARS-COV-2.
- **Preprocessing of the dataset**: After fixing the value of we used N-grams technique to extract the number of occurrences of each subsequence of nucleotides of size in all of a given dataset. Then, we form a common base by selecting the first subsequences that are the most shared between all of the viruses. Next, we project each virus onto this base of size and we normalize the results. An example of this preprocessing with two genomes “ATGATGGATTG” and “ATGGATGTGGG” is given in table 2.

Table 2: Example of applying N-grams on two nucleic acid sequences, the values between brackets represents the normalization

N-grams of genome N°1 with ATGATGGATTG		N-grams of genome N°2 with ATGGATGTGGG		Common base of size	projection of genome N°1 onto the common base	projection of genome N°2 onto the common base
A T G	2	A T G	2	ATG	2 (0.8660254)	2 (0.8660254)
G A T	2	T G G	2	GAT	2 (0.8660254)	1 (-0.8660254)
A T T	1	G G A	1	GGA	1 (-0.8660254)	1 (-0.8660254)
G G A	1	G G G	1	TGG	1 (-0.8660254)	2 (0.8660254)
T G A	1	G A T	1			
T T G	1	T G T	1			
T G G	1	G T G	1			

- **Identifying SARS-CoV-2**: We considered that all of the 10313 genomes of SARS-CoV-2 that are collected from different geographic locations are unknown. Then we used the aforementioned five supervised machine learning algorithms to identify the specie (class) of each one.

The step 3 is performed in two phases (Figure 1). The first is called the training in which each machine learning algorithm will learn the genomes of the whole dataset (except SARS-COV-2). In this phase we selected the parameters of each algorithm using a grid search: SVM has two parameters (The regularization parameter) and (the width of the Gaussian kernel); ANN has two parameters the number of hidden layers and the number of neurons by each layer; KNN has two parameters k (The number of neighbors) and the distance measure that was chosen as Euclidian; DT has one parameter which is the criteria to split a given node (two possible impurity functions can be used Gini index and Information gain); and BN hasn’t practically any parameter to estimate. The second is named testing in which the algorithms will be used to predict the species of the genomes of SARS-COV-2 that are considered previously as unknown.

Experimental results

Before starting the research of the origins and inter-hosts of COVID-19 we propose to visualize the set of 10313 genomes of this virus. Since each genome, after applying N-grams, will be represented in higher dimensions (i.g dimensions if) it is very difficult if not impossible to visualize them directly. So reducing our space to 2D or 3D (without loss of information) may allow us to plot and observe patterns more clearly. Figure 2 shows the results of applying a dimensionality reduction technique, from 64D to 2D, known as principal components analysis on the aforementioned set. It can be seen clearly that the majority of genomes of SARS-COV-2 belong to the same cluster or family (except some outliers).

The Figures 3, 4, 5, 6 and 7 show that the origins of 10313 genomes of SARS-COV-2 vary with respect to the machine learning technique used in the experiment. The majority of classifiers (KNN, SVM, DT and ANN) gave practically the same result that shows that Pangolin is the most probable inter-host of SARS-COV-2. Concerning the result illustrated in Figure 7, which corresponds to NB, it can be seen that all of the genomes of SARS-COV-2 are categorized by this classifier as Alphacoronavirus 1. This unacceptable result, which is completely different from the others, can be explained by the fact that NB is based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. The features here are the set of 3Grams that are evidently dependent because their succession determinates a nucleic acid sequence.

Since, SVM is well known by its high discrimination capacity we will use its detailed results to further investigate this virus (Figure 8).

Figure 8 shows the detailed results of the belonging degree of a sample of SARS-COV-2 to each member of Coronaviridae family (A total of 96). This belonging degree is a natural number that expresses the number of times where a member is voted as a potential origin of this virus. It can be seen that the most voted origin of COVID-19 is pangolin followed directly by Alphacoronavirus-1. This latter includes: Canine coronavirus, Feline coronavirus and Transmissible gastroenteritis virus that are linked to dogs, cats and porcine respectively [22]. The combination of these results with those obtained previously (Figure 3,4,5,6) show that even if alphacoronavirus-1 is classified in the second position and is linked to the most domestic animals (direct contact with humans) it still a very weak competitor for pangolin. This conclusion is due to the fact that four classifiers among five used in the previous experiment didn't give practically alphacoronarirus-1 as origin despite the variety of genomes that are collected from different geographical locations. In the third position we find bat alphacoronavirus which is not given by any classifier as a potential origin of SARS-COV-2 (except ANN with a very negligible number of genomes). In conclusion, pangolin has the greatest possibility to be the actual origin of SARS-COV-2.

Conclusion

In biology and in particular in genetics there are many problems of complex nature and often computationally difficult. Analyzing biological sequence to produce meaningful information is one of these problems that require efficient tools to be approached. In this paper two tools were proposed to analyze and compare the genomes of SARS-COV-2 with a large set of coronaviruses family. The first one is N-gram widely used in text categorization; its role is to create feature vectors from different genomic sequences. The second tool is a set of five supervised machine learning classifiers (Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree, and Support Vector Machine) that was proposed to classify these vectors and to detect similarities between the genomes SARS-COV-2 and the other viruses from coronaviridae family. The experimental results have shown that the virus causing coronavirus disease 2019 (COVID-19) has one potential origin which is Pangolin.

Conflict Of Interest Statement

The author states that there is no conflict of interest.

References

1. Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory 1992, Pittsburgh Pennsylvania USA, 144-152.
2. Cavnar W.B. and Trenkle J. M. N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval 1994, 161-175, LasVegas, NV.
3. Chen C-C, Qian X, and Yoon B-J. Effective computational detection of piRNAs using n-gram models and support vector machine. BMC Bioinformatics 2017; 18(suppl 14):103-109.
4. Huang H-H and Yu C. Clustering DNA sequences using the out-of-place measure with reduced n-grams. Journal of Theoretical Biology 2016; 406:61-72.
5. Mathew J, Vijayakumar R, and John J .Principal Component Analysis with SVM for Disease Diagnosis. International Journal of Innovative Technology and Exploring Engineering 2019; 8(8):615-620.
6. Luan J, Jin X, Lu Y, et al. SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. Journal of Medical Virology.2020 doi: 10.1002/jmv.25817
7. Reddy U.J, Dhanalakshmi P, and Reddy P.D.K. Image Segmentation Technique Using SVM Classifier for Detection of Medical Disorders. Ingénierie des Systèmes d'Information 2019; 24(2):173-176.
8. Shen L, Chen H, Yu Z et al. Evolving support vector machines using fruit fly optimization for medical data classification, Knowledge-Based Systems 2016; 96:61-75
9. International Committee on Taxonomy of Viruses: https://talk.ictvonline.org/
10. Tayde S.S., Nagamma Patil (2016). A Novel Approach for Genome Data Classification Using Hadoop and Spark Framework. 333-343, In: Shetty N., Prasad N., Nalini N. (eds), Emerging Research in Computing, Information Communication and Applications. Springer, Singapore.
11. Wong M-C, Cregeen S-J-J, Ajami N-J, et al. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019, bioRxiv. 2020.
12. Paraskevis D, Kostaki EG, Magiorkinis G, et al. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infection, Genetics and Evolution2020; 79:104212. Epub
13. Senthilselvan N, Rajarajan S, Subramaniyaswamy V. Optimized artificial neural network for classification of biological data, International Journal of Engineering & Technology 2018, 7(2):817-822
14. Telenti A, Lippert C, Chang P-C et al. Deep learning of genomic variation and regulatory network data, Human Molecular Genetics 2018; 27(R1):63-71
15. Kaymak S, Helwan A, Uzun D. Breast cancer image classification using artificial neural networks, Procedia Computer Science 2017, 120:126-131

16. Bhagya Shree, S.R., Sheshadri, H.S. Diagnosis of Alzheimer's disease using Naive Bayesian Classifier. Neural Computing and Applications 2018, 29(1):123-132.
17. Anderson, M. P., & Dubnicka, S. R. A sequential naïve Bayes classifier for DNA barcodes. Statistical Applications in Genetics and Molecular Biology 2014; 13(4): 423-434
18. Chandra B , Gupta M. Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data, Expert Systems with Applications 2011, 38(3):1293-1298
19. Du P, Cao S, Li Y, SubChlo: Predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm, Journal of Theoretical Biology 2009, 261(2) : 330-335

20 Li L, Darden T. A, Weingberg C. R. et al . Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Combinatorial chemistry & high throughput screening 2001, 4(8): 727-739.

21. Lee C-P, Lin W-S, Chen Y-M, et al. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method, Expert Systems with Applications 2011, 38(5):4661-4667
22. National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/
23. Saha S, Ghosh A, Bandopadhyay S. et al. Missing value imputation in DNA microarray gene expression data: a comparative study of an improved collaborative filtering method with decision tree based approach, International Journal of Computational Science and Engineering 2019,18(2)130-139
24. Ludwig S-A, Picek S, Jakobovic D. Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm. In: Kahraman C., Topcu Y. (eds) Operations Research Applications in Health Care Management. International Series in Operations Research & Management Science, Vol 262. Springer, Cham, 2018
25. Sudha V.P., Vijaya M.S. (2018) Decision Tree Based Model for the Classification of Pathogenic Gene Sequences Causing ASD. In: Deshpande A. et al. (eds) Smart Trends in Information Technology and Computer Communications. SmartCom 2017. Communications in Computer and Information Science, Vol 876. Springer, Singapore
26. Zhou P, Yang X-L, Wang X-G et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020. Vol 579. https://doi.org/10.1038/s41586-020-2012-7
27. Lam T, Shum M, Zhu H et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins, nature 26 March 2020. https://doi.org/10.1038/s41586-020-2169-0

28 Zhang T, Wu Q and Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.03.022

29. Han G-Z, Pangolins Harbor SARS-CoV-2-Related Coronaviruses, Trends in Microbiology 2020. https://doi.org/10.1016/j.tim.2020.04.001

30. Qiu Y, Zhao Y-B, Wang Q, et al. Predicting the angiotensin converting enzyme 2 (ACE2) utilizing capability as the receptor of SARS-CoV-2. Microbes and infection 2020. https://doi: 10.1016/j.micinf.2020.03.003

Figures

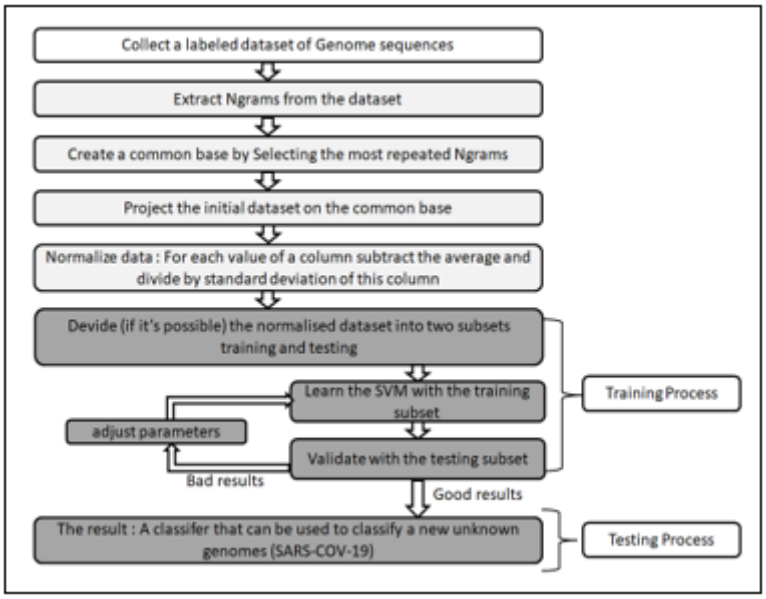

Figure 1

Description of the experimental protocol

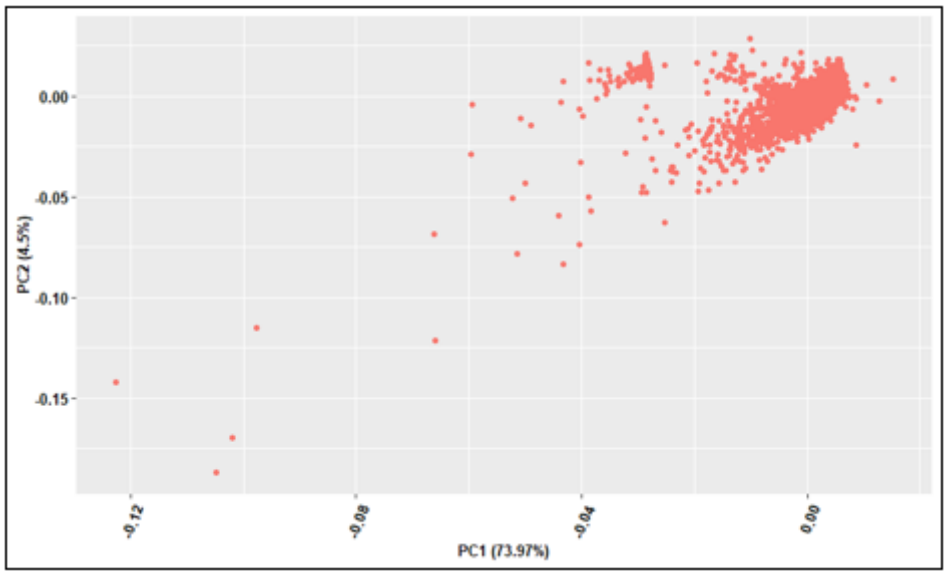

Figure 2

Visualization in 2D of the set of 10313 genomes of SARS-COV-2

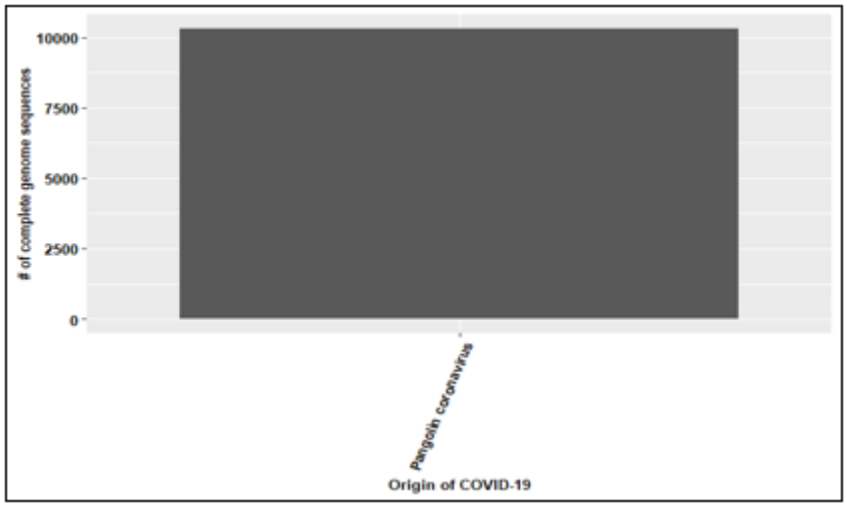

Figure 3

Origins of COVID-19 using KNN with k=1 and N-gram with N=3 and M=64

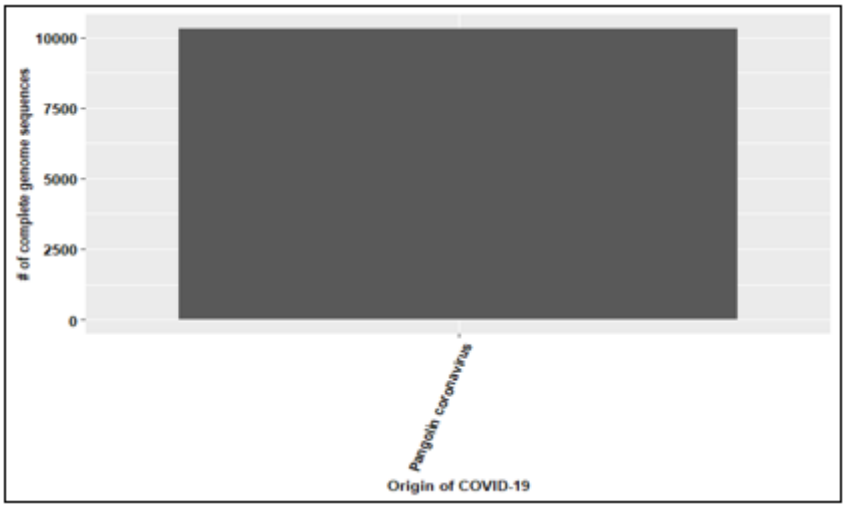

Figure 4

Origins of COVID-19 using SVM with σ=0.1,C=10 and N-gram with N=3 and M=64

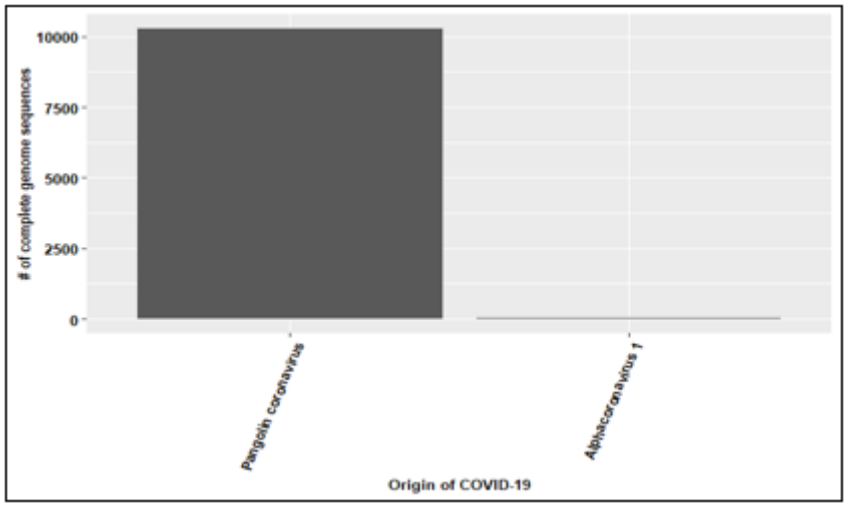

Figure 5

Origins of COVID-19 using Decision tree with Gini index as a measures of impurity and N-gram with N=3 and M=64

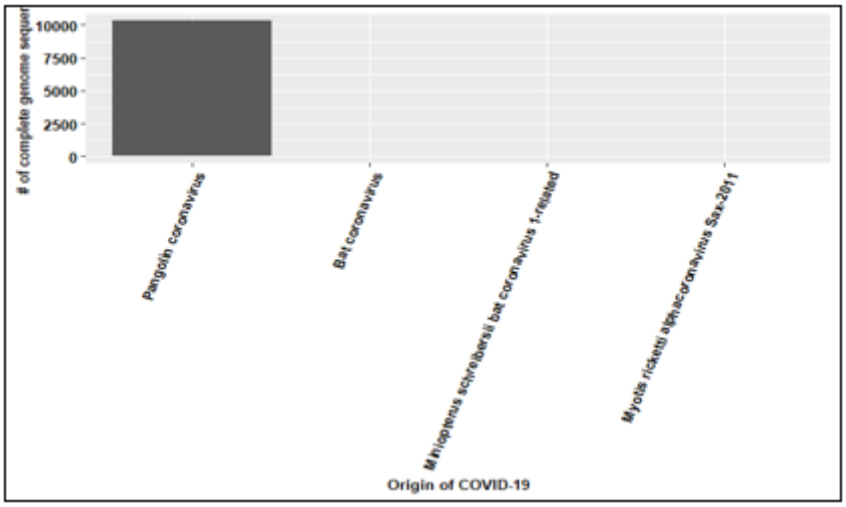

Figure 6

Origins of COVID-19 using ANN with 2 hidden layers each consists of 12 neurons and N-gram with N=3 and M=64

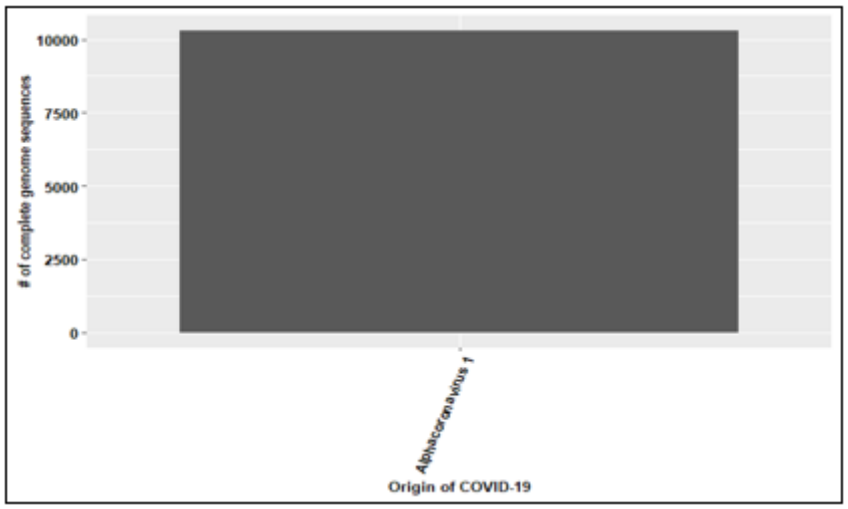

Figure 7

Origins of COVID-19 using NB and N-gram with N=3 and M=64

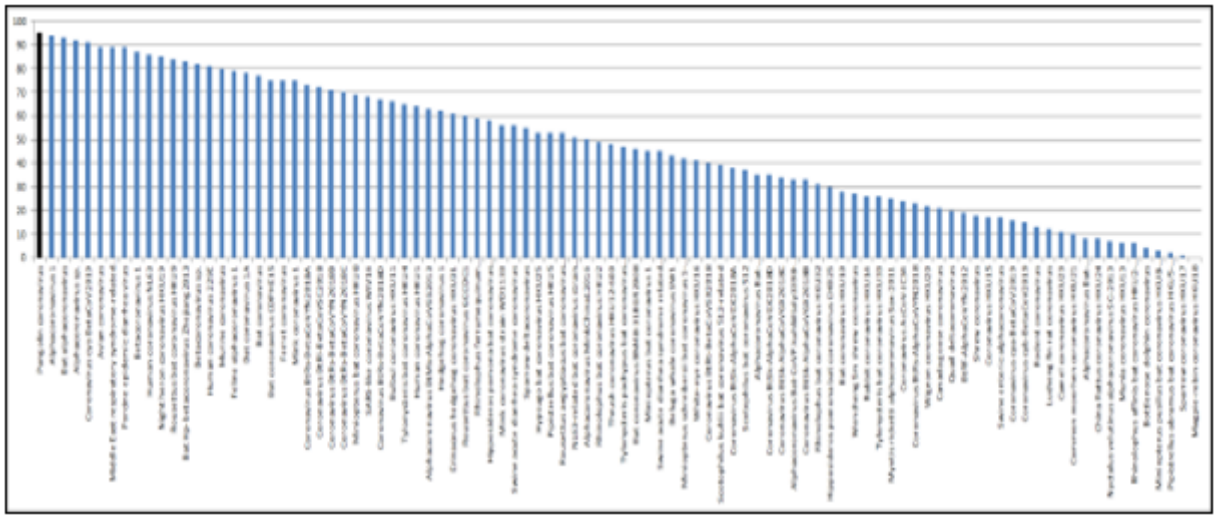

Figure 8

A detailed membership degree of a genome of SARS-COV-2