COVID-Predictor: RNA Sequence based Prediction of Coronavirus

Jnanendra Prasad Sarkar
Larsen & Toubro Infotech Ltd., and Jadavpur University
Indrajit Saha (indrajit@nitttrkol.ac.in)
National Institute of Technical Teachers' Training & Research
Arijit Seal
Cognizant Technology Solutions
Debasree Maity
MCKV Institute of Engineering

Research Article

Keywords: SARS-CoV-2, machine learning based coronavirus prediction technique, web based application

Posted Date: April 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-23913/v1

Abstract

The problem of virus classification is always a subject of concern for virology or epidemiology over the decades. Moreover, the detection of highly divergent or yet unknown viruses is a major challenge despite of its clinical importance. In this situati on, the outbreak of novel coronavirus (SARS-CoV-2) and its susceptibility in different epidemic condition around the world clearly suggest that the virus is mutating to create divergent variants and making the task of virus prediction more challenging. On the other hand, despite of novel coronavirus, two more coronaviruses such as MERS and SARS-CoV-1 are already present. Therefore, the use of machine learning technique is highly required at this moment to predict the coronaviruses by considering their divergent genetic functional characteristics. Thus, we are proposing machine learning based coronavirus prediction technique, called COVID- Predictor, where 1000 of RNA sequences of SARS-CoV-1, MERS, SARS-CoV-2 and other virus are used to train a Na¨ıve Bayes classifier so that it can predict any unknown sequence of these viruses. In order to develop the COVID-Predictor, the feature vector is constructed by the motifs of the sequence generated by k-mer and n-gram techniques. The model has been validated using 10 fold cross validation in comparison with other classification techniques. The results show the superiority of our predictor by achieving average 97% accuracy on unseen validation set. The same pre-trained model has been used to design a web based application where RNA sequences of unknown viruses can be uploaded to predict class of coronavirus. The predictor, code and datasets are available here: http://www.nitttrkol.ac.in/indrajit/projects/COVID-Predictor/

Introduction

The China Country Office of World Health Organization (WHO)1, 2 on 31st December 2019, informed that few pneumonia cases have been detected in Wuhan City, Hubei Province of China with unknown etiology. Subsequently, on 7th January 2020, Chinese authority identified a novel virus as a cause of this disease, which WHO and International Committee on Taxonomy of Viruses declared as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or novel coronavirus (COVID-2019)2, 3 on 11th February 2020. After subsequent research, it is found that the novel coronavirus belongs in the family of coronavirus. In this family, Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) are also present. The medical research community suspects that COVID-19 is more transmissible but comparatively less fatal than SARS. According to various evidences4, 5, the transmission rate of COVID-2019 from a human to another human seems higher than SARS and the virus might be bat or pangolin origin3, 6, 7. It is also suggested that mostly the transmission of this virus is via droplets. It means once an infected person coughs or sneezes, emitted droplets come into contact with another individual over a short distance, the second individual might get infected.

As on 15th April 2020, 2077453 positive cases have been registered across the world while 134355 patients died and 509741 recovered for this virus according to “worldometers.info” 1. This virus is spreading rapidly and a threat to the human population. Generally, coronaviruse can infect multiple organs in different hosts such as animal and human. It mainly attacks respiratory system in human like other two viruses, SARS and MERS, in the same family. The genetic features like potential etiological agents of the SARS-CoV-2 have recently been identified after metagenomic analysis using next-generation sequencing (NGS)2. Moreover, another study8 shows that spike protein receptor-binding domain (RBD) of SARS-CoV-2 binds with host receptor angiotensin-converting enzyme 2 (ACE2). It generally helps to regulate the transmission of COVID-2019 in cross-species and human. At present, it is observed that the virus is susceptible with the environment by mutating and creating divergent variants.

Thus the early prediction of pneumonia caused by COVID-2019 is challenging task9.

To address the above urgent requirement, here we have developed a machine learning based technique, called, COVID- Predictor, where RNA sequences of three different coronaviruses and other virus, such as Ebola and Debgue, are used from The National Center for Biotechnology Information (NCBI) 2 and Global Initiative on Sharing All Influenza Data (GISAID) 3 to train a Multinomial Naive Bayes (MNB)10 classifier so that it can predict any unknown sequence of these viruses. For this purpose, k-mer algorithm11, 12 is used to create motifs from the RNA sequences. Thereafter n-gram concept is used to create a Bag-of-Words (BoWs) in order to create a count vector. Such count vectors of sequences are used to train the MNB. Subsequently, testing is done in the same fashion with 10 fold cross validation and unseen sequences of coronaviruses from databases. The model has also been compared with other classification techniques such as kernel based Gaussian Support Vector Machine (GSVM)13 and Random Forest (RF)14. The MNB based model shows the superior performance in comparison with other classifier based model by achieving average 97% accuracy on validation data. The same pre-trained model is used to develop a web application so that scientific and diagnostic communities related to coronavirus prediction can get the benefit out of this.

Results And Discussion

In this section, we have discussed about the data preparation, the parameters and metrics which are used for COVID-Predictor and the outcome of the predictor.

Table 1. Statistics of the refined datasets of corona and other viruses

Virus Name	Source of Sequence	After filtering out above 20K bp Sequences			
		No. of Sequence	Max Length of Sequence	Min Length of Sequence	Avg Length of Sequence
SARS-CoV-1	NCBI	515	32759	21221	29608
MERS	NCBI	291	30150	27121	29983
SARS-CoV-2	GISAID	2369	29986	20008	29520
		After filtering out above 10K bp Sequences			
Other virus	NCBI	600	19897	10735	15316

Data Preparation

The dataset of SARS-CoV-1, MERS, other kind of viruses like Ebola and Dengue were downloaded from NCBI while SARS-CoV-2 was downloaded from GISAID in fasta format on 28th March 2020. Although proposed predictor does not require sophisticated data prepossessing, only it requires complete genome sequence of viruses. As a result 515, 291, 2369 sequences of SARS-CoV-1, MERS, SARS-CoV-2 respectively of length more than 20K bp while 600 other virus such as Ebola and Dengue of length more than 10K bp are considered in our experiment. The statistics of the refined consolidated datasets are shown in Table 1, while the country wise statistics of SARS-CoV-2 is reported in Table 2. In order to visualise the virus sequences, t-distributed Stochastic Neighbor Embedding (tSNE)15 is used on count vector as generated by k-mer and n-gram techniques. k-mer is now an essential part of many methods in bioinformatics such as genome and transcriptome assembly, metagenomic sequencing, error correction of sequence reads etc.16. Solis-Reyes et.al in11 has explained that k-mer works better than other popular methods like REGA17, SCUEAL18, COMET19 etc. The embedded representation of all four virus classes and top 21 country specific sequences of SARS-CoV-2 are shown in Figure 1 and 2.

Table 2. Statistics of country wise refined sequences of SARS-CoV-2

Country	No. of Sequences	Country	No. of Sequences	Country	No. of Sequences	Country	No. of Sequences
USA	590	Spain	27	Chile	7	Mexico	1
Iceland	343	Congo	19	Ireland	6	Nepal	1
China	275	Scotland	18	Vietnam	6	Nigeria	1
Netherlands	190	Canada	17	Kuwait	4	Northern Ireland	1
England	160	Italy	17	Slovakia	4	Pakistan	1
Wales	107	Taiwan	17	Czech Republic	3	Panama	1
Japan	83	Singapore	14	Saudi Arabia	3	Peru	1
France	75	Finland	13	Fujian	2	Poland	1
Australia	64	South Korea	13	Hungary	2	Russia	1
Belgium	45	Georgia	10	India	2	South Africa	1
Portugal	44	Luxembourg	10	Thailand	2	Sweden	1
Brazil	34	Denmark	9	Cambodia	1	Turkey	1
Switzerland	31	Malaysia	8	Colombia	1		
Hong Kong	30	New Zealand	8	Ecuador	1		
Germany	27	Norway	8	Lithuania	1		

Parameters setting and Metrics

The experiments have been performed using python 3.6 and executed on an Intel Core i5-2410M CPU at 2.30 GHz Machine with 8GB RAM and Windows 7 operating system. The required input parameters are experimentally set and those are number of trees for RF = 100, decision for RF is “gini”, alpha value as smoothing factor of MNB is 0.1 and kernel used in GSVM is “rbf”. To evaluate results of COVID-Predictor, the popular performance metrics such as *Accuracy*, *Precision*, *Recall* and *F*1 −*Score* are used.

Table 3. Classification performance of different machine learning techniques after performing 10-fold cross validation with different values of k-mer and n-gram on 1000 genome sequences of SARS-CoV-1, MERS, SARS-CoV-2 and Other virus samples

Method	k-mer	n-gram = 2				n-gram = 3				n-gram = 4				n-gram = 5				Aggregated Score
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	
MNB	2	0.99810	0.99817	0.99810	0.99810	0.99810	0.99817	0.99810	0.99810	0.99810	0.99817	0.99810	0.99810	0.99905	0.99910	0.99905	0.99905	0.99835
GSVM		0.94857	0.95725	0.94857	0.94952	0.96762	0.97151	0.96762	0.96795	0.98190	0.98324	0.98190	0.98191	0.99238	0.99276	0.99238	0.99237	0.97359
RF		0.99429	0.99458	0.99429	0.99428	0.99429	0.99458	0.99429	0.99428	0.99429	0.99457	0.99429	0.99428	0.99619	0.99632	0.99619	0.99618	0.99482
MNB	3	0.99810	0.99817	0.99810	0.99810	0.99810	0.99817	0.99810	0.99810	0.99905	0.99910	0.99905	0.99905	0.99810	0.99817	0.99810	0.99810	0.99835
GSVM		0.96762	0.97151	0.96762	0.96795	0.98190	0.98324	0.98190	0.98191	0.99238	0.99276	0.99238	0.99237	0.99905	0.99909	0.99905	0.99905	0.98561
RF		0.99429	0.99458	0.99429	0.99428	0.99524	0.99548	0.99524	0.99523	0.99810	0.99816	0.99810	0.99809	0.99714	0.99725	0.99714	0.99714	0.99623
MNB	4	0.99810	0.99817	0.99810	0.99810	0.99905	0.99910	0.99905	0.99905	0.99810	0.99817	0.99810	0.99810	1.00000	1.00000	1.00000	1.00000	0.99882
GSVM		0.98190	0.98324	0.98190	0.98191	0.99238	0.99276	0.99238	0.99237	0.99905	0.99909	0.99905	0.99905	1.00000	1.00000	1.00000	1.00000	0.99344
RF		0.99429	0.99456	0.99429	0.99428	0.99714	0.99725	0.99714	0.99714	0.99714	0.99725	0.99714	0.99714	0.99810	0.99817	0.99810	0.99810	0.99670
MNB	5	0.99905	0.99910	0.99905	0.99905	0.99810	0.99817	0.99810	0.99810	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99929
GSVM		0.99238	0.99276	0.99238	0.99237	0.99905	0.99909	0.99905	0.99905	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99765
RF		0.99619	0.99633	0.99619	0.99618	0.99714	0.99725	0.99714	0.99714	0.99810	0.99816	0.99810	0.99809	0.99810	0.99816	0.99810	0.99809	0.99740
MNB	6	0.99810	0.99817	0.99810	0.99810	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99929
GSVM		0.99905	0.99909	0.99905	0.99905	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99714	0.99724	0.99714	0.99714	0.99882
RF		0.99714	0.99725	0.99714	0.99714	0.99810	0.99816	0.99810	0.99809	0.99810	0.99816	0.99810	0.99809	0.99714	0.99724	0.99714	0.99714	0.99764
MNB	7	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99905	0.99908	0.99905	0.99905	0.99953
GSVM		1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99714	0.99724	0.99714	0.99714	0.99619	0.99632	0.99619	0.99619	0.99811
RF		0.99810	0.99816	0.99810	0.99809	0.99810	0.99816	0.99810	0.99809	0.99714	0.99724	0.99714	0.99714	0.99810	0.99816	0.99810	0.99809	0.99787

Outcome of the Predictor

The dataset consisting all four types of virus sequences such as SARS-CoV-1, MERS, SARS-CoV-2 and Other viruses has been divided into two sets - one for training set and other for validation purpose. Stratified sampling method has been applied to prepare training dataset to ensure that representative from all four types of virus classes are present. As a result 1000 of virus sequences are used in training. Moreover, data samples are carefully selected from each category to avoid imbalance class problem. The validation dataset contains those sequences which are not present in training dataset. The training dataset is used in three independent machine learning techniques viz. MNB, GSVM and RF. For each machine learning technique, the motifs of virus sequences are created using k-mer method. Thereafter, such motifs are combined using n-gram technique to create count vector which is used to train the classifiers. In our experiments the value k of k-mer varies between 2 to 7, while the value of n-gram varies between 2 to 5. Each classifier has been evaluated with 10-fold cross validation followed by further validation on unseen dataset taken from NCBI and GISAID on 8th April 2020. The performance metrics of each machine learning technique

Table 4. Classification performance of COVID-Predictor on validation data

Source	Data Samples	Accuracy	Precision	Recall	F1-Score
NCBI + GISAID	2043 Sequences (262 SARS-CoV-1, 41 MERS, 1440 SARS-CoV-2, 300 Other virus)	0.92217	0.92991	0.92217	0.90726
NCBI	493 Sequences (Only SARS-CoV-2)	1.00000	1.00000	1.00000	1.00000
GISAID	4747 Sequences (Only SARS-CoV-2)	1.00000	1.00000	1.00000	1.00000

with 10 fold cross validation for different values of k-mer and n-gram have been reported in Table 3. Four quantitative metrics are further consolidated as single aggregated score for ease of comparison. The aggregated score has been computed simply by taking average of all the scores following the similar approach of what is used in20. The boundary of aggregated score is [0,1] where higher value signifies better result. It is evident from the Table 3 that MNB based COVID-Predictor produces higher aggregated score, i.e. 0.99953 for value of k-mer as 7. Similar results are also observed for MNB based COVID-Predictor for other values of k-mer. Thus, according to the results, we have prepared the pre-trained model of COVID-Predictor with 1000 genomic sequences of four virus classes for values of k-mer and n-gram as 7 and 3 respectively. To gain further confidence, we have used additional validation set of sequences as reported in Table 4. While validated with 2043 samples, it is observed that 159 cases are false positive considering prediction of SARS-CoV-2 is positive. After further investigation, it has been found that these 159 sequences are SARS-CoV-1 and misclassified by COVID-Predictor as SARS-CoV-2. As our primary objective is to predict SARS-CoV-2, we further wanted to examine the rate of false negative. For this purpose, additional two sets of SARS-CoV-2 sequences are used separately, one with 493 samples from NCBI and another with 4747 samples from GISAID. Both the cases, the COVID-Predictor predicted SARS-CoV-2 with 100% accuracy. This experiment establishes that COVID-Predictor with the proposed feature building approach has potential to predict SARS-CoV-2 with higher accuracy. The same pre-trained model is used to build the web based application where the unknown sequences can be uploaded to predict the class of coronavirus. The screen shot of the web based predictor is shown in Figure 3 and 4.

Method

The primary objective of the proposed COVID-Predictor is to correctly classify the RNA sequences of coronaviruses. In this regard, the complete RNA sequences are split into motifs using popular k-mer technique. Such motifs for four class of viruses are shown in Figure 6(a)-

(d) as word cloud. Thereafter, the n-gram technique is used to create a feature by considering *n* number of motifs. Top 10 n-grams for different viruses is shown in Figure 7(a)-(d). These n-grams/features are used call as Bag-of-Words(BoW). Such BoWs are further used to create count vector for a virus sequence. The count vectorization computes the frequencies of n-grams in a particular sequence and creates a numeric feature vector which is used in subsequent machine learning techniques.

This is reported and visualize from Figure 1 that the RNA sequences of SARS-CoV-1 and SARS-CoV-2 are similar and challenging to distinguish. Therefore, the machine learning technique can play an important role to predict such sequences. As we have broken the RNA sequence into motifs, grouped into n-grams, the features behave like sequence of texts. Therefore, it becomes a text classification problem. In this regard, we have considered probabilistic based Multinomial Naive Bayes (MNB), kernel based support vector machine (SVM) and tree based technique like random forest (RF) to evaluate. Independently, all three machine learning techniques are evaluated with features generated by count vectorization after considering different values of k-mer and n-gram. Based on the performance of three machine learning techniques over 10-fold cross validation on training data, we have finalised MNB as underlying technique for building COVID-Predictor as used in our web application. The pipeline of the proposed COVID-Predictor is described in Figure 5.

Conclusion

In current world wide context, it has become very much essential for mankind to predict coronavirus as early as possible, because SARS-CoV-2 infection has become pandemic and the both infection & death rate is getting increased world wide almost exponentially at every day while we are writing this article. As a contribution to mankind, in this study, we have proposed COVID-Predictor for predicting the coronaviruses viz. SARS-CoV-1, MERS and SARS-CoV-2 based on their RNA sequences. The same is also provided as web application so that scientific and diagnostic communities related to coronavirus prediction can get the benefit out of this. In order to achieve better performance, we have carefully performed data preprocessing, paid careful attention to build efficient features vectors, leveraging appropriate machine learning technique by performing 10-fold cross validation. Experimentally, the Multinomial Naive Bayes technique has been finalized for building web based predictor as MNB performed better on different real life datasets of RNA sequences which are collected from NCBI and GISAID as on 8th April, 2020. Summarizing: The proposed COVID-Predictor has been shown to have capability to predict the coronaviruses viz. SARS-CoV-1, MERS and SARS-CoV-2 based on their RNA sequences. A web based application of this COVID-Predictor has also been built, so that SARS-CoV-2 can be predicted as early as possible to save mankind.

Declarations

Acknowledgements (not compulsory)

We thank all those who have contributed sequences to NCBI and GISAID database.

Author contributions statement

J.P.S, I.S conceived the study design and wrote the manuscript. J.P.S, I.S, A.S and D.M conducted the experiments and wrote code. J.P.S, I.S. and A.S analysed the results. All authors reviewed the manuscript.

Additional information

The authors declared no conflicts of interest. All the processed datasets, supplementary and the software are available at http://www.nitttrkol.ac.in/indrajit/projects/COVID-Predictor/

Footnotes

1https://www.worldometers.info/coronavirus/

2https://www.ncbi.nlm.nih.gov/

3https://www.gisaid.org/

References

1. WHO. Coronavirus disease (covid-19) pandemic. *World Heal. Organ. West. Pac. China* (2020). URL https://www. who.int/china.
2. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in china, 2019. *The New Engl. J. Medicine* **382**, 727–733 (2020).
3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nat.* **579**, 270–273 (2020).

4. Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* **395**, 514–523 (2020).
5. Li, Q. *et al.* Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *The New Engl. Medicine* **382**, 1199–1207 (2020).
6. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of sars-cov-2. *Nat. Medicine* (2020).
7. Zhang, Y.-Z. & Holmes, E. C. A genomic perspective on the origin and emergence of sars-cov-2. *Cell* (2020).
8. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of sars coronavirus. *J. Virol.* **94** (2020).
9. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet* **395**, 497–506 (2020).
10. George, H. & Langley, J. P. Estimating Continuous Distributions in Bayesian Classifiers. *Proc. Eleventh Conf. on Uncertain. Artif. Intell.* **69**, 338–345 (1995).
11. Solis-Reyes, S., Avino, M., Poon, A. & Kari, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes. *Plos One* **13**, e0206409 (2018).
12. Manekar, S. C. & Sathe, S. R. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience* **7**, 1–13 (2018).
13. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
14. Breiman, L. Random Forests. *Learn.* **45**, 5–32 (2005).
15. Hinton, G. & Roweis, S. T. Stochastic neighbor embedding. *In Proc. 15th Int. Conf. on Neural Inf. Process. Syst.* 857–864 (2002).
16. Melsted, P. & Pritchard, J. K. Efficient counting of k-mers in dna sequences using a bloom filter. *BMC Bioinforma.* **12** (2011).
17. Pineda-Pena, A. C. *et al.* Automated subtyping of hiv-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new rega version 3 and seven other *Infect. Genet. Evol.* **19**, 337–348 (2013).
18. Pond, S. L. K. *et al.* An evolutionary modelbased algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in hiv-1. *PLoS Comput. Biol.* **5**, e1000581 (2009).
19. Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C. & Bercoff, D. P. Comet: adaptive context-based modeling for ultrafast hiv-1 subtype identification. *Nucleic Acids Res.* **42**, e144–e144 (2014).
20. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *The New Engl. J. Medicine* **9**, 471–472 (2012).

Figures

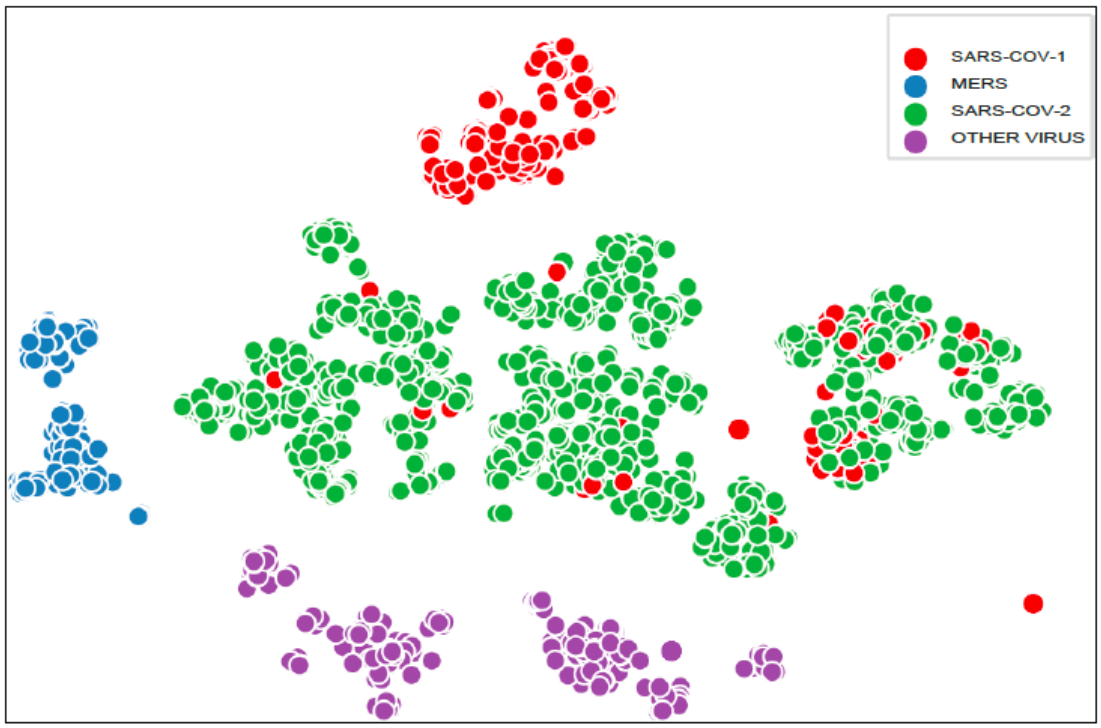


Figure 1

Embedded representation of SARS-CoV-1, MERS, SARS-CoV-2 and Other virus classes.

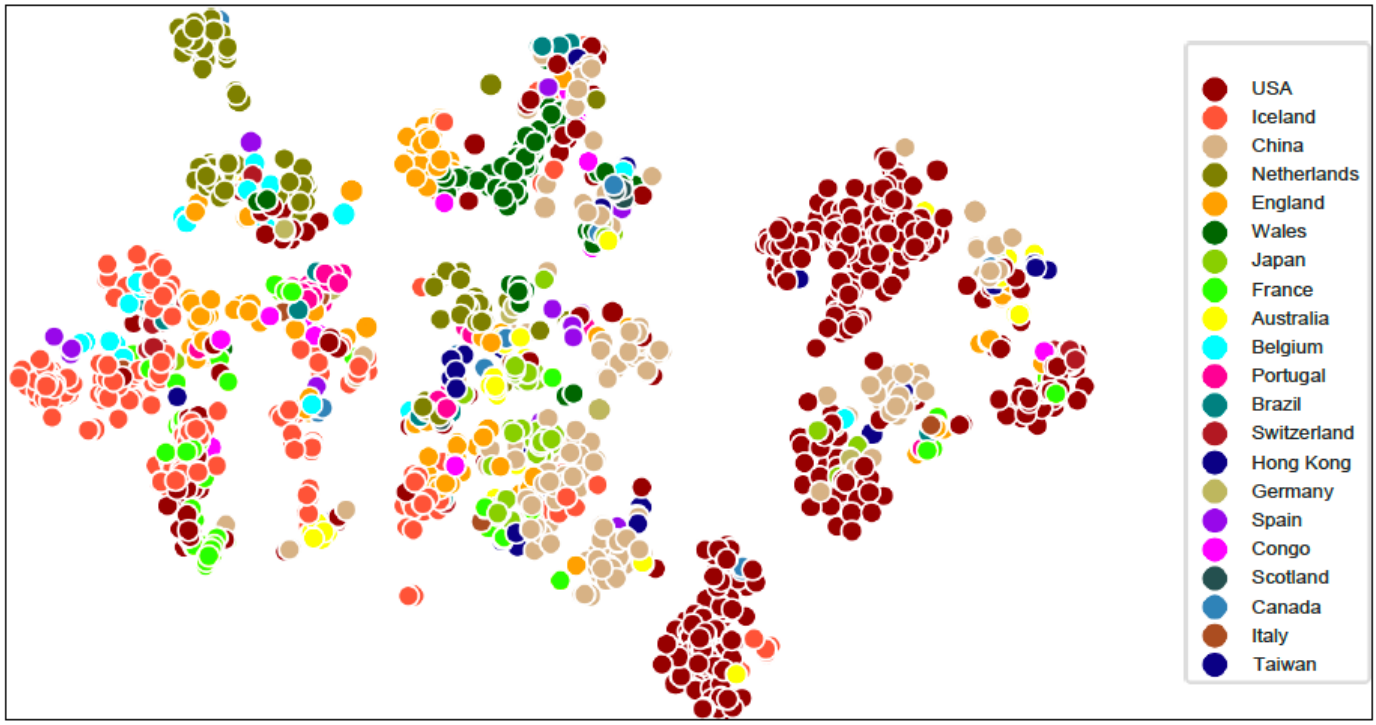


Figure 2

Embedded representation of SARS-CoV-2 sequences of top 21 countries.

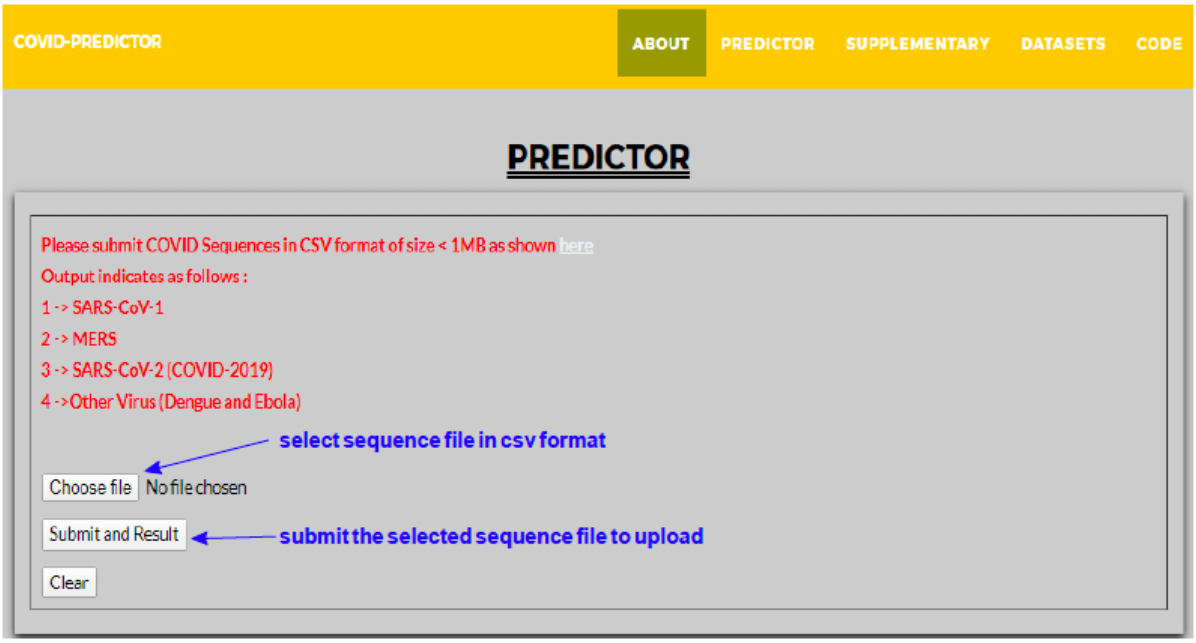


Figure 3

The screen shot of the web based COVID-Predictor to select csv file

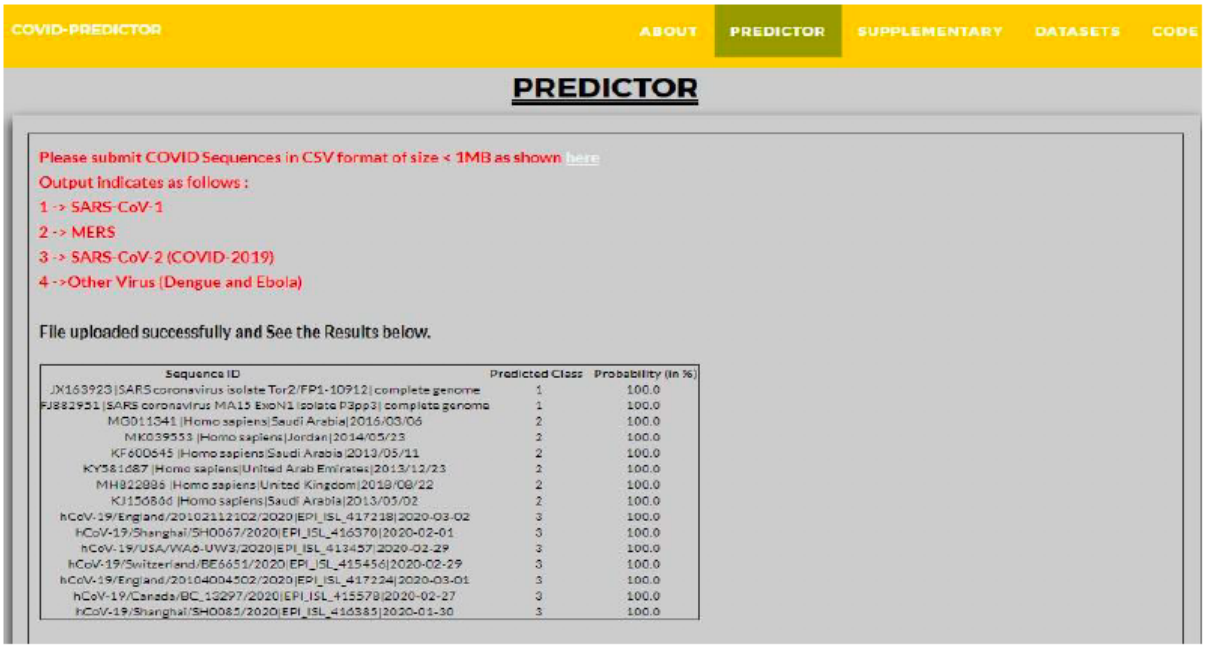


Figure 4

The screen shot of the web based COVID-Predictor after prediction

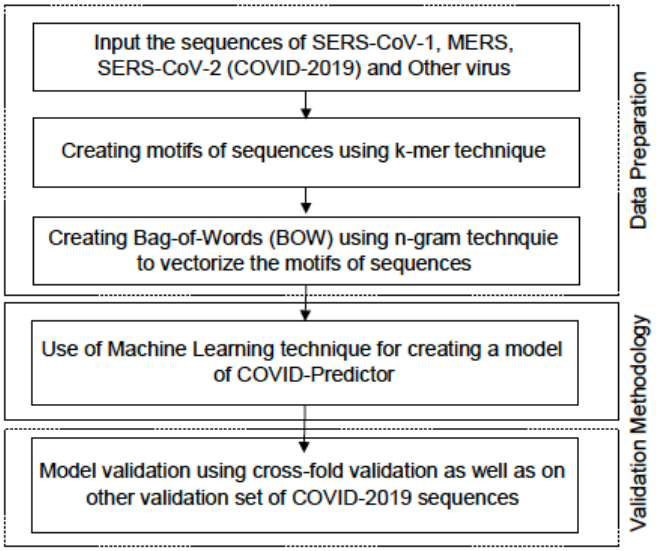
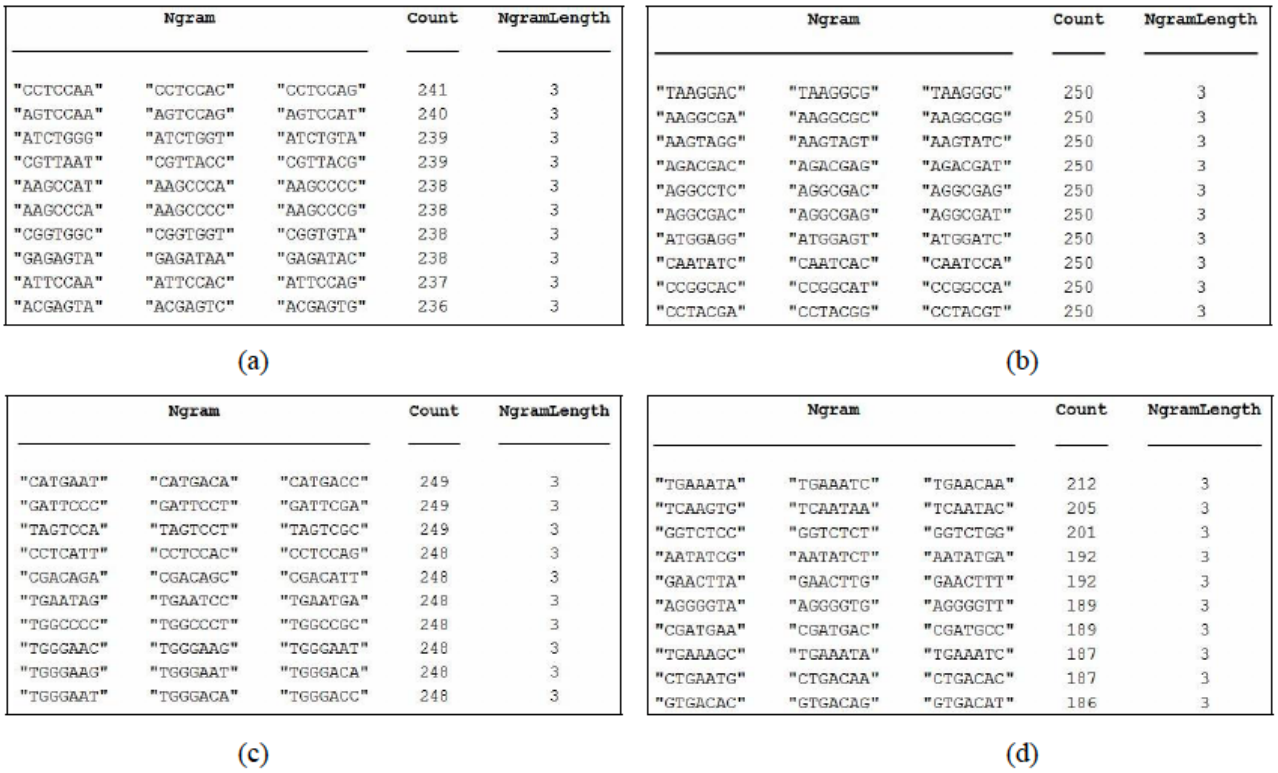


Figure 5

Pipeline of the proposed COVID-Predictor

Figure 6

Word cloud of k-mer motifs of RNA sequences for (a) SARS-CoV-1 (b) MERS (c) SARS-CoV-2 and (d) Other virus as Bag-of-Words(BoW). Such BoWs are further used to create count vector for a virus sequence. The count vectorization computes the frequencies of n-grams in a particular sequence and creates a numeric feature vector which is used in subsequent machine learning techniques.



Ngram			Count	NgramLength
"CCTCCAA"	"CCTCCAC"	"CCTCCAG"	241	3
"AGTCCAA"	"AGTCCAG"	"AGTCCAT"	240	3
"ATCTGGG"	"ATCTGGT"	"ATCTGTA"	239	3
"CGTTAAT"	"CGTTACC"	"CGTTACG"	239	3
"AAGCCAT"	"AAGCCCA"	"AAGCCCC"	238	3
"AAGCCCA"	"AAGCCCC"	"AAGCCCG"	238	3
"CGGTGGC"	"CGGTGGT"	"CGGTGTA"	238	3
"GAGAGTA"	"GAGATAA"	"GAGATAC"	238	3
"ATTCCAA"	"ATTCCAC"	"ATTCCAG"	237	3
"ACGAGTA"	"ACGAGTC"	"ACGAGTG"	236	3

(a)

Ngram			Count	NgramLength
"TAAGGAC"	"TAAGGCG"	"TAAGGGC"	250	3
"AAGGCGA"	"AAGGCGC"	"AAGGCGG"	250	3
"AAGTAGG"	"AAGTAGT"	"AAGTATC"	250	3
"AGACGAC"	"AGACGAG"	"AGACGAT"	250	3
"AGGCCTC"	"AGGCGAC"	"AGGCGAG"	250	3
"AGGCGAC"	"AGGCGAG"	"AGGCGAT"	250	3
"ATGGAGG"	"ATGGAGT"	"ATGGATC"	250	3
"CAATATC"	"CAATCAC"	"CAATCCA"	250	3
"CCGGCAC"	"CCGGCAT"	"CCGGCCA"	250	3
"CCTACGA"	"CCTACGG"	"CCTACGT"	250	3

(b)

Ngram			Count	NgramLength
"CATGAAT"	"CATGACA"	"CATGACC"	249	3
"GATTCCC"	"GATTCCT"	"GATTCGA"	249	3
"TAGTCCA"	"TAGTCCT"	"TAGTCGC"	249	3
"CCTCATT"	"CCTCCAC"	"CCTCCAG"	248	3
"CGACAGA"	"CGACAGC"	"CGACATT"	248	3
"TGAATAG"	"TGAATCC"	"TGAATGA"	248	3
"TGGCCCC"	"TGGCCCT"	"TGGCCGC"	248	3
"TGGGAAC"	"TGGGAAG"	"TGGGAAT"	248	3
"TGGGAAG"	"TGGGAAT"	"TGGGACA"	248	3
"TGGGAAT"	"TGGGACA"	"TGGGACC"	248	3

(c)

Ngram			Count	NgramLength
"TGAAATA"	"TGAAATC"	"TGAACAA"	212	3
"TCAAGTG"	"TCAATAA"	"TCAATAC"	205	3
"GGTCTCC"	"GGTCTCT"	"GGTCTGG"	201	3
"AATATCG"	"AATATCT"	"AATATGA"	192	3
"GAACTTA"	"GAACTTG"	"GAACTTT"	192	3
"AGGGGTA"	"AGGGGTG"	"AGGGGTT"	189	3
"CGATGAA"	"CGATGAC"	"CGATGCC"	189	3
"TGAAAGC"	"TGAAATA"	"TGAAATC"	187	3
"CTGAATG"	"CTGACAA"	"CTGACAC"	187	3
"GTGACAC"	"GTGACAG"	"GTGACAT"	186	3

(d)

Figure 7

Top 10 n-grams of k-mer motifs generated for (a) SARS-CoV-1 (b) MERS (c) SARS-CoV-2 and (d) Other virus