Computational Predictions for Protein Sequences of COVID-19 Virus via Machine Learning Algorithms

Walaa Alkady
Ain Shams University
Muhammad Zanaty
Ain Shams University
Heba M. Afify (hebaaffify@pg.cu.edu.eg)
Cairo university

Research Article

Keywords: COVID-19 protein sequences, Conjoint triad (CT), Linear Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM).

Posted Date: June 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34004/v1

Version of Record: A version of this preprint was published at Medical & Biological Engineering & Computing on July 22nd, 2021. See the published version at https://doi.org/10.1007/s11517-021-02412-z.

Abstract

The coronavirus infection is increasingly evolving to be an international epidemic in 27 countries as a serious respiratory disease. Therefore, the computational biology carrying this virus that correlated with the human population is urgently needed. In this paper, the classification of the human protein sequences of COVID-19 according to the country is applied by machine learning algorithms. The proposed model is based on the distinguishing of 9238 sequences by three stages including data preprocessing, data labeling, and classification. In the first stage, the function of data preprocessing converts the amino acids of COVID-19 protein sequences to eight groups of numbers based on volume and dipole of the amino acids. In the second stage, there are two methods for data labeling of 27 countries from 0 to 26. The first method is based on the selection of one number for each country according to the code number of countries while the second method is based on binary elements only for each country. The classification algorithms are executed to discover different COVID-19 protein sequences according to their countries. The findings are concluded that the accuracy of 100% performed by country based binary labeling method with Linear Regression (LR) or K-Nearest Neighbor (KNN) or Support Vector Machine (SVM) classifiers. Further, it found that the USA with large data records in infection rate has more priority for correct classification compared to other countries with a low data rate. The unbalanced data for COVID-19 protein sequences is considered a major issue, especially the available data in USA represented 76% from a total of 9238 sequences. As a consequence, this proposed model will help as a diagnostic bioinformatics tool for the COVID-19 protein sequences among different countries.

1. Introduction

The resistance against the coronavirus is still a challenging phase due to limited information on this virus. The available cases of coronavirus protein sequences encouraged the research for the taxonomic classification of the COVID-19 virus [1]. The early differentiation of this virus created a low spread, especially in dynamic population growth. Many factors focused on the probabilities of developing the virus in patients such as age, hygienic behaviors, location, environment, and health status [2]. The World Health Organization (WHO) reported that the COVID-19 virus beginning from China has the chance to spread for increasing mortality in many countries [3]. Also, WHO confirmed that this virus is called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and emerged in bats which can transmit to human [4]. It means that exposure to the specific wild environment is providing more rapid infections. The infection risk of COVID-19 is controlled by accurate quarantine schedules because there are no specific drugs or vaccines for this virus [5]. This COVID-19 virus has an important influence on the financial and social aspects that cause the critical need for reducing the diffusion of this virus. Moreover, the classification task of the virus protein sequences by machine learning algorithms supported the treatment plans for this COVID-19 virus [6]. Additionally, the NCBI virus as public resources [7] provided a list of genomic datasets targeting COVID-19 virus to aid in coronavirus variant analysis. Zhang et.al [8] presented the protein sequence analysis for host identification of COVID-19 and the study of the similarity between COVID-19 and HIV (human immunodeficiency).

Randhawa et al. [9] presented the machine learning algorithms of DNA sequences to classify the COVID-19 virus as beta coronavirus. Qiang et al. [10] proposed the observation model of protein sequences of COVID-19 virus to study the genomic evolution by machine learning algorithms. Zhou et al. [11] suggested the protein-protein interaction (PPI) network for recognizing the candidate drugs for COVID-19 virus. The prediction of PPI for HIV, SARS, and pandemic influenza A (H1N1) virus [12] identified by amino acid composition as shown in [13]. On the other hand, discovering host-virus PPI proposed by the co-immunoprecipitation method [14].

The temperature divergence in different countries [15] created the mutation of COVID-19 protein sequences that promoted the researchers to study geographical variation from bioinformatics vision. Therefore, classification based country in the domain of COVID-19 protein sequences is a significant survey for tracking this virus.

In this study, the proposed model investigated the classification of COVID-19 protein sequences according to the country through machine learning techniques.

2. Materials And Methods

The proposed block diagram designed procedure series including data preprocessing, data labeling, and classification algorithms to categorize the COVID-19 countries as shown in Fig.1.

2.1 Dataset Description

The used dataset for SARS-Cov-2(COVID 19) is shown in the NCBI virus [7]. This dataset contains 9238 sequences, each sequence identified using the accession number and has much information such as protein sequence as a FASTA file, geographical location, and protein sequence length. There are two forms in the dataset. In first forms, Comma Separated Values (CSV) files include the accession numbers of protein sequences and other information such as geographical location (Geo_location), host, isolation source, and gene bank title as in Fig.2. The maximum sequence length found in this dataset is QIX12193 that has 7098 amino acids. The minimum sequence length found in this dataset is YP_009725312 that has 13 amino acids. In second forms, FASTA Files contains the accessions along with the protein sequence as in Fig.3.

2.2. Data Preprocessing

Generally, the protein sequences consist of twenty amino acids. In the data preprocessing phase, classification of the amino acids' side chains to seven classes according to their volume and dipole [16] is applied to the protein sequences. Molecular modeling and density-functional theory approaches [17] used for calculating volumes and dipoles of the side chains. This classification method converts the protein sequences to numbers from zero to seven (0-7). Table 1 shows the classification of the amino acids based on their side-chains dipoles and volumes.

Table 1: Amino acids classification according to dipoles and volumes

Class Number	Dipole Scale	Volume Scale	Amino Acids
1	-	-	A, G, V
2	-	+	I, L, F, P
3	+	+	Y, M, T, S
4	++	+	H, N, Q, W
5	+++	+	R, K
6	+'+'+'	+	D, E
7	+	+	C

Some explanations of the symbols found in Table 1:

- Dipole scale: (-), Dipole<1.0 / (+), 1.0<Dipole<2.0 / (++), 2.0<Dipole<3.0 / (+++), Dipole>3.0 / (+'+'+'), Dipole>3.0 with opposite orientation

- Volume scale: (-), Volume<50; / (+), Volume> 50
- Cysteine (C) amino acid is separated from class 3 because of its ability to form disulfide bonds.

As shown in Table 2, the unknown amino acid class was added to the coding with class number zero and amino acid labeled with (X). For example, the protein sequence "ALGCERQSKXTP" would be represented as “121765435032” according to eight amino acid classes. The sample of COVID-19 protein sequences after conversion displayed in Fig.4.

Table 2: Amino acids coding according to eight Classes

Class Number	Amino Acids
0	X (Unknown Amino Acid)
1	A, G, V
2	I, L, F, P
3	Y, M, T, S
4	H, N, Q, W
5	R, K
6	D, E
7	C

After converting the protein sequences to numbers, an array of eight elements is filled with the number of the occurrences of each amino acid class for each sequence in the dataset. Finally, the frequencies of the amino acids are normalized by dividing each frequency by the length of the sequence to avoid the biased to the taller sequences.

2.3. Data Labeling

The classification based country is implemented on COVID 19 dataset. The countries' names were labeled as in Table 3. There are 27 unique countries found in this dataset after removing the cities.

Table 3: Code numbers of countries

Country Number	Country Name	Country Number	Country Name
0	Australia	14	Nepal
1	Belgium	15	Nigeria
2	Brazil	16	Pakistan
3	Spain	17	Peru
4	Colombia	18	Philippines
5	Finland	19	South Korea
6	France	20	China
7	India	21	Sweden
8	Iran	22	Taiwan
9	Iraq	23	Thailand
10	Israel	24	Tunis
11	Italy	25	USA
12	Japan	26	Viet Nam
13	Malaysia		

Table 4: One Number Format Samples

Protein Sequence ID	Country Label
1	8
2	8
3	6
4	6
5	6
6	25
7	25
8	25
9	25
10	20
11	20
12	20
13	20
14	25
15	25
16	25
17	25
18	25

Table 5: Binary Array Format Samples

ID	Country Binary Label																										
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

2.4. Classification

After converting the protein sequences of COVID 19 to numbers that belong to eight amino acid classes, different six classifiers sequentially used for predicting the protein sequence country. Each amino acid class considered a feature in this classification model. One number format and binary array format are the two methods formulated in labeling the predicted class (country). The six classifiers are carried out by Linear Regression (LR) [18], K-Nearest Neighbor (KNN) [19] using various numbers of neighbors, Support Vector Machine (SVM) [20] using different kernel functions, Naive Bayesian (NB) [21], Decision Tree (DT) [22], and Random Forest (RF) [23] using different numbers of estimators. The main objective of classifiers structure is to reduce the learning complexity for accurate examining of unknown samples.

The LR classifier is supported by the relationship between dependent and independent variables using a linear separating hyperplane. The KNN classifier is graded by a class of their closest neighbors based on K values. The KNN classifier focused on distance measurement by the density distribution that not related to decision boundary calculations. In this paper, the selected K values were 20, 50, 150, and 201 to attain the best K for results accuracy. The SVM classifier investigated decision boundary calculations by four types of kernel functions such as linear, sigmoid, polynomial, and radial basis function (RBF) to choose the effective function for discrimination. The NB classifier is based on probability theorem and maximum likelihood calculation. The DT classifier is based on data separation into groups according to low entropy measures. The RF classifier is based on large numbers of DT method and a high *correlation* between two trees.

3. Results And Discussion

The dataset is separated into 80% for training database and 20% for testing database to perform COVID 19 classification based country under Python computational environment. There are two methods of class labeling used for prediction of the COVID 19 country. The inclusive comparison between the machine learning approaches is investigated on COVID 19 protein sequences. For KNN classifier, the adjustable factor is the K value; for SVM classifier, it is the kernel function; for RF classifier, it is an estimator number. Table 6 discussed the classification results for one number labeling method.

Table 6: Performance comparison of different classifiers for One Number Labeling Method

Classifiers	Accuracy
Linear Regression (LR)	77.7%
K-Nearest Neighbor (KNN), k= 20	77.2%
K-Nearest Neighbor (KNN), k= 50	77.7%
K-Nearest Neighbor (KNN), k= 150	77.7%
K-Nearest Neighbor (KNN), k >= 201	77.7%
Support Vector Machine (SVM), Kernel: RBF	78.0%
Support Vector Machine (SVM), Kernel: linear	77.2%
Support Vector Machine (SVM), Kernel: sigmoid	74.4%
Support Vector Machine (SVM), Kernel: polynomial	78.2%
Naive Bayesian (NB)	13.2%
Decision Tree (DT)	**79.5%**
Random Forest (RF), estimators =10	79.0%

Table 7 discussed the results of COVID 19 protein sequences-assisted detection for the binary labeling method. The preferable results confirmed that the DT classifier achieved an accuracy of 79.5% for one number labeling method, whereas linear SVM, KNN, and LR classifiers achieved an accuracy of 100% for the binary array labeling method. For binary array method, the best classifier for SVM is that used the linear function and for KNN is that used the value of k >= 201. For one number labeling method, the worst classifier provided by NB that achieved an accuracy of 13.2%. For binary array labeling method, the worst classifier provided by KNN with K=20 that achieved an accuracy of 93.4%.

Additionally, it found that the best K value of >=50 is obtained by 77.7% of accuracy for one number labeling method. While the best K value of >=201 is obtained by 100% of accuracy for binary array labeling method. The choice of K value has a slight impact on the classification accuracy for one number labeling method but it has a robust impact on the accuracy classification accuracy for binary labeling method.

In the SVM classifier, the best kernel function was polynomial for one number labeling method that obtained 78.2% of accuracy. While the best kernel function was linear for binary array labeling method that obtained 100% accuracy. The accuracy of SVM classifier altered only by changing of kernel functions for binary labeling method.

It can be noted that the binary array labeling method is an effective method for the classification-based country of COVID 19 protein sequences.

Table 7: Performance comparison of different classifiers for Binary Array Labeling Method

Classifiers	Accuracy
Linear Regression (LR)	**100.0%**
K-Nearest Neighbor (KNN), k= 20	93.4%
K-Nearest Neighbor (KNN), k= 50	95.7%
K-Nearest Neighbor (KNN), k= 150	99.8%
K-Nearest Neighbor (KNN), k >= 201	**100.0%**
Support Vector Machine (SVM), Kernal: RBF	99.6%
Support Vector Machine (SVM), Kernal: linear	**100.0%**
Support Vector Machine (SVM), Kernal: sigmoid	95.2%
Support Vector Machine (SVM), Kernal: polynomial	99.2%
Naive Bayesian (NB)	99.8%
Decision Tree (DT)	97.0%
Random Forest (RF), estimators =10	96.5%

Fig.5 revealed the relationship between the eight amino acid classes and frequency of amino acid classes in each COVID-19 protein sequences by using all samples found in the used dataset (9238 sequences) concerning the eight classes of amino acids. This figure confirmed that the COVID-19 virus has a high record of amino acids provided in the second class that consists of Isoleucine (I), Leucine (L), Phenylalanine (F), and Proline (P) amino acids.

4. Conclusion

In this paper, a dataset of 9238 COVID-19 protein sequences is used to evaluate the capability of the proposed model to predict the country of the protein sequences. The proposed model extracted suited features from the protein sequences by replacing the amino acids characters in each sequence by the eight amino acid classes normalized frequencies. After that, the model performed six different classifiers to predict the country of the virus protein sequence.

This model indicated that the classification using the binary array labeling method has more promising results than the one number labeling method. Hence, the preliminary proposed model produced a highest accuracy of 100% within an appropriate time using country binary array labeling method with LR, KNN, and SVM classifiers.

The better classification accuracy observed in USA protein sequences when comparing with different countries because of their high data records, around 7020 records from 9238 records (76% from all occurrences). It is interesting to record that unbalanced data has an inferior impact on the COVID-19 classification results. This leads to increase misclassifications on the testing set because COVID-19 sequences of all countries except USA occupied only 24 % from all sequences. Furthermore, the classifiers biased to USA records compared with the countries with a small number of occurrences.

Finally, this proposed model could be expand to large stored data of COVID-19 protein sequences and could then develop the prediction algorithms for unbalanced sequences within several countries.

Declarations

Conflict of interests

The authors declare that they have no conflict of interests

References

[1] S. KANNAN, P. SHAIK SYED ALI, A. SHEEZA, K. HEMALATHA, COVID-19 (Novel Coronavirus 2019) – recent trends, SARS, European Review for Medical and Pharmacological Sciences, 2020; 24: 2006-2011.

[2] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. Nature 2020.

[3] Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E. A novel coronavirus emerging in china - key questions for impact assessment. N Engl J Med. 2020 Jan 24. doi: 10.1056/NEJMp2000929.

[4] Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia[J]. N Engl J Med. 2020 Jan 29.

[5] Centers for Disease Control and Prevention, 2019 Novel Coronavirus (2019-nCoV), Wuhan, China (2019); https://www.cdc.gov/coronavirus/2019-nCoV/summary.html.

[6] Gurjit S. Randhawa, Maximillian P.M. Soltysiak, Hadi El Roz, Camila P.E. de Souza,

Kathleen A. Hill and Lila Kari, Machine learning-based analysis of genomes suggests associations between Wuhan 2019-nCoV and bat Betacoronaviruses, bioRxiv preprint doi: https://doi.org/10.1101/2020.02.03.932350

[7] NCBI virus: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202%20(SARS-CoV2),%20taxid:2697049

[8] Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y, Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1, J Proteome Res. 2020 Apr 3;19(4):1351-1360.

[9] Gurjit S. Randhawa, Maximillian P. M. Soltysiak, Hadi El Roz, Camila P. E. de Souza, Kathleen A. Hill, Lila Kari, Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study, PLOS ONE, 2020

[10] Xiao-Li Qiang, Peng Xu, Gang Fang, Wen-Bin Liu & Zheng Kou, Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus, Infectious Diseases of Poverty volume 9, Article number: 33 (2020)

[11] Zhou, Y., Hou, Y., Shen, J. *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* **6,** 14 (2020).

[12] M. P. Girard, J. S. Tam, O. M. Assossou, M. P. Kieny, The 2009 A (H1N1) influenza virus pandemic: A review, *Vaccine,* **28** (2010), 4895–4902.

[13] S. Alguwaizani, B. Park, X. Zhou, D. S. Huang, K. Han, Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids, *J. Healthc. Eng.,***2018** (2018).

[14] E. Golemis, *Protein-protein interactions: A molecular cloning manual*, CSHL Press, (2005).

[15] Wiebe A, Longbottom J, Gleave K, Shearer FM, Sinka ME, Massey NC, et al. Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. Malar J 2017;16:1–10.

[16] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, “Predicting protein-protein interactions based only on sequences information,” Proceedings of the National Academy of Sciences of the USA (PNAS), 2007, pp. 4337-4341.

[17] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, et al. GAUSSIAN 03 (Gaussian, Pittsburgh, PA), Revision C.02 (2003).

[18] Stephan Dreiseitla and Lucila Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review” Journal of Biomedical Informatics 35 (2002) 352–359

[19] P. Cunningham, S.J. Delany, k-Nearest neighbour classifiers, in: Multiple Classifier System, 2007, pp. 1–17.

[20] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications, ACAI 1999: Machine Learning and Its Applications pp 249-257

[21] Rish, I. An empirical study of the naive bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence; IBM: New York, NY, USA, 2001; Volume 3, pp. 41–46.

[22] A. Gutierrez-Rodríguez, J.F. Martínez-Trinidad, M. García-Borroto, J. Carrasco- Ochoa, Mining patterns for clustering on numerical datasets using unsupervised decision trees, Knowl. Based Syst. 82 (2015) 70–79.

[23] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Figures

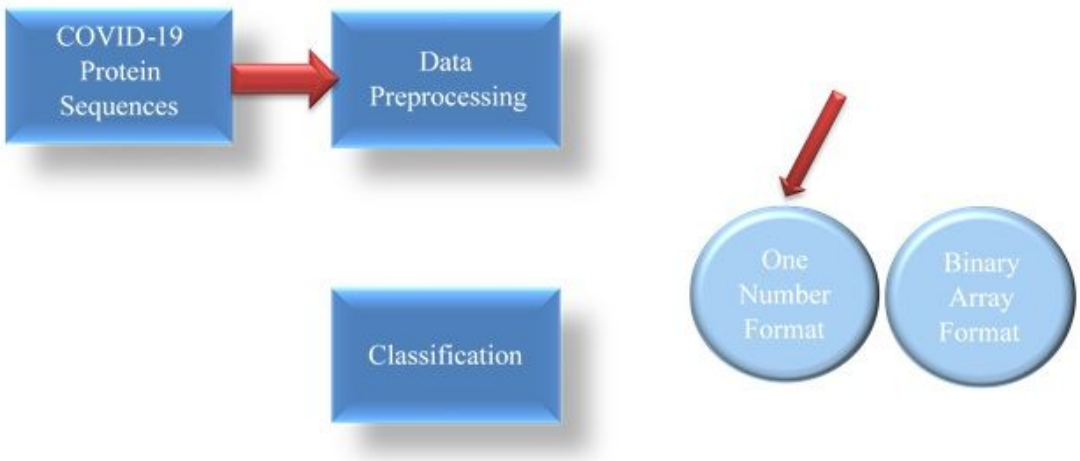
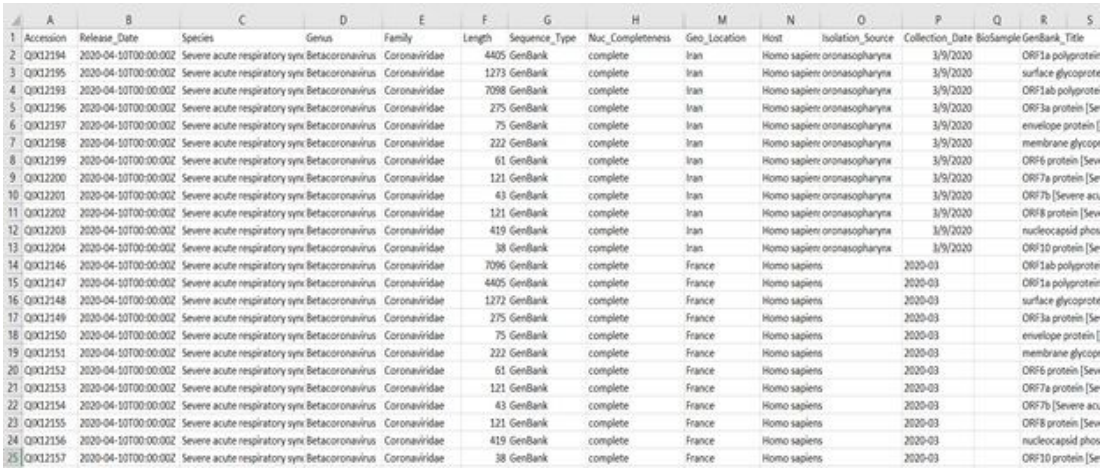


Figure 1

Block diagram of classification based country for COVID-19 protein sequences



	A	B	C	D	E	F	G	H	M	N	O	P	Q	R	S
1	Accession	Release_Date	Species	Genus	Family	Length	Sequence_Type	Nuc_Completeness	Geo_Location	Host	Isolation_Source	Collection_Date	BioSample	GenBank_Title	
2	QIX12194	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	4405	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF1a polyprotei	
3	QIX12195	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	1273	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		surface glycoprote	
4	QIX12193	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	7098	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF1ab polyprote	
5	QIX12196	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	275	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF3a protein [Se	
6	QIX12197	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	75	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		envelope protein [	
7	QIX12198	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	222	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		membrane glycop	
8	QIX12199	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	61	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF6 protein [Sev	
9	QIX12200	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	121	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF7a protein [Se	
10	QIX12201	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	43	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF7b [Severe acu	
11	QIX12202	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	121	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF8 protein [Sev	
12	QIX12203	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	419	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		nucleocapsid phos	
13	QIX12204	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	38	GenBank	complete	Iran	Homo sapien	oronasopharynx	3/9/2020		ORF10 protein [Se	
14	QIX12146	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	7096	GenBank	complete	France	Homo sapiens		2020-03		ORF1ab polyprote	
15	QIX12147	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	4405	GenBank	complete	France	Homo sapiens		2020-03		ORF1a polyprotei	
16	QIX12148	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	1272	GenBank	complete	France	Homo sapiens		2020-03		surface glycoprote	
17	QIX12149	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	275	GenBank	complete	France	Homo sapiens		2020-03		ORF3a protein [Se	
18	QIX12150	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	75	GenBank	complete	France	Homo sapiens		2020-03		envelope protein [	
19	QIX12151	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	222	GenBank	complete	France	Homo sapiens		2020-03		membrane glycop	
20	QIX12152	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	61	GenBank	complete	France	Homo sapiens		2020-03		ORF6 protein [Sev	
21	QIX12153	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	121	GenBank	complete	France	Homo sapiens		2020-03		ORF7a protein [Se	
22	QIX12154	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	43	GenBank	complete	France	Homo sapiens		2020-03		ORF7b [Severe acu	
23	QIX12155	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	121	GenBank	complete	France	Homo sapiens		2020-03		ORF8 protein [Sev	
24	QIX12156	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	419	GenBank	complete	France	Homo sapiens		2020-03		nucleocapsid phos	
25	QIX12157	2020-04-10T00:00:00Z	Severe acute respiratory syn	Betacoronavirus	Coronaviridae	38	GenBank	complete	France	Homo sapiens		2020-03		ORF10 protein [Se	

Figure 2

CSV Data Samples

```
>QIX12196 |ORF3a protein [Severe acute respiratory syndrome coronavirus 2]
MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSAS
KIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFLYLYALVYFLQSINF
VRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDYCIPYNSVTSSIVITSGDGTTSPIS
EHDYQIGGYTEKWESGVKDCVVLHSYFTSDYYQLYSTQLSTDTGVEHVTFFIYNKIVDEP
EEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL
>QIX12197 |envelope protein [Severe acute respiratory syndrome coronavirus 2]
MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS
RVKNLNSSRVPDLLV
>QIX12198 |membrane glycoprotein [Severe acute respiratory syndrome coronavirus 2]
MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPV
TLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILL
NVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYK
LGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ
>QIX12199 |ORF6 protein [Severe acute respiratory syndrome coronavirus 2]
MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEI
D
>QIX12200 |ORF7a protein [Severe acute respiratory syndrome coronavirus 2]
MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNSPFHPLADNKFALTCFS
TQFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKT
E
>QIX12201 |ORF7b [Severe acute respiratory syndrome coronavirus 2]
MIELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHNETCHA
>QIX12202 |ORF8 protein [Severe acute respiratory syndrome coronavirus 2]
MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL
CVDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF
I
>QIX12203 |nucleocapsid phosphoprotein [Severe acute respiratory syndrome coronavirus 2]
```

Figure 3

FASTA Data Samples

Key	Type	Size	Value
QIV64999	list	7096	[3, 6, 3, 2, 1, 2, 1, 2, 4, 6, ...]
QIV65000	list	1273	[3, 2, 1, 2, 2, 1, 2, 2, 2, 2, ...]
QIV65002	list	75	[3, 3, 3, 2, 1, 3, 6, 6, 3, 1, ...]
QIV65005	list	121	[3, 5, 2, 2, 2, 2, 2, 1, 2, 2, ...]
QIV65007	list	419	[3, 3, 6, 4, 1, 2, 4, 4, 4, 5, ...]
QIV65009	list	43	[3, 2, 6, 2, 3, 2, 2, 6, 2, 3, ...]
QIV65010	list	7096	[3, 6, 3, 2, 1, 2, 1, 2, 4, 6, ...]
QIV65012	list	275	[3, 6, 2, 2, 3, 5, 2, 2, 3, 2, ...]
QIV65013	list	75	[3, 3, 3, 2, 1, 3, 6, 6, 3, 1, ...]
QIV65016	list	121	[3, 5, 2, 2, 2, 2, 2, 1, 2, 2, ...]
QIV65017	list	121	[3, 5, 2, 2, 1, 2, 2, 1, 2, 2, ...]
QIV65020	list	43	[3, 2, 6, 2, 3, 2, 2, 6, 2, 3, ...]

Figure 4

Protein sequences after conversion using eight amino acids classes

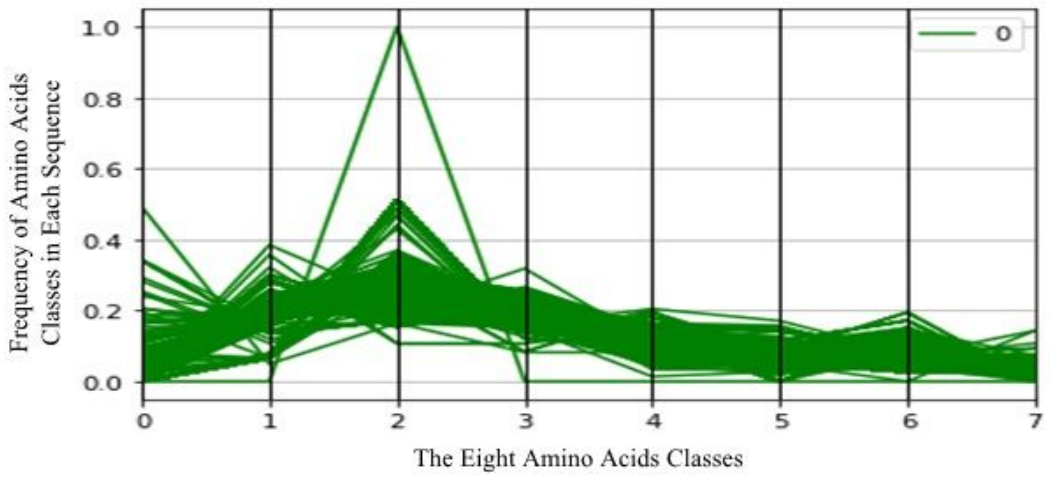


Figure 5

COVID-19 Samples with respect to amino acids classes