Data Mining of Coronavirus: SARS-CoV-2, SARS-CoV and MERS-CoV

Jung Eun Huh
University of Oxford

Seunghee Han (seunghee991105@gmail.com)
University of Birmingham College of Medical and Dental Sciences https://orcid.org/0000-0001-6180-5329

Taeseon Yoon
Korea University - Seoul Campus: Korea University

Research note

Keywords: Coronavirus, SARS-CoV-2, SARS-CoV, MERS-CoV, BLAST, Apriori, Decision Tree, SVM

Posted Date: March 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-322281/v1

Version of Record: A version of this preprint was published at BMC Research Notes on April 20th, 2021.
See the published version at https://doi.org/10.1186/s13104-021-05561-4.

Abstract

Objectives: All three SARS-CoV-2, SARS-CoV and MERS-CoV belong to the Coronaviridae family. In this study we compare amino acid and codon sequence of SARS-CoV-2, SARS-CoV and MERS-CoV using different statistics programs to understand their characteristics. Specifically, we are interested in how differences in the amino acid and codon sequence lead to different incubation periods and outbreak periods.

Results: The initial question we had was to compare SARS-CoV-2 to different viruses in the coronavirus family to understand its characteristics. The result of experiments using BLAST, Apriori and Decision Tree has shown that SARS-CoV-2 had high similarity with SARS-CoV while having comparably low similarity with MERS-CoV. We decided to compare the codons of SARS-CoV-2 and MERS-CoV to see the difference. Though the viruses are very alike according to BLAST and Apriori experiments, SVM proved that they can be effectively classified using non-linear kernels. Decision Tree experiment has proved several remarkable properties of SARS-CoV-2 amino acid sequence that cannot be found in MERS-CoV amino acid sequence.

The consequential purpose of this paper is to minimize the damage on humanity from SARS-CoV-2. Hence, further studies can focus on the comparison of SARS-CoV-2 virus with other viruses that also can be transmitted during latent periods.

Introduction

All three SARS-CoV-2, SARS-CoV and MERS-CoV belong to the Coronaviridae family, Orthocoronavirinae subfamily and to Betacoronavirus genera. Betacoronavirus infect mammals and other known members are Bovine Coronavirus, Human coronavirus OC43, Tylonycteris bat coronavirus HKU4 and etc. Of six species of known human coronaviruses, seven including two different strains subdivided from one species, the three previously mentioned viruses are known to produce severe symptoms.

In this study we compare amino acid and codon sequence of SARS-CoV-2, SARS-CoV and MERS-CoV using different statistics programs to understand their characteristics. Specifically, we are interested in how differences in the amino acid and codon sequence lead to different incubation periods and outbreak periods. We also hope to provide insight on the solution of the current SARS-CoV-2 pandemic and suggest future research directions.

Main Text

Materials

SARS-CoV-2, SARS-CoV and MERS-CoV are all members of the coronavirus family. Thus, they share many microbiological similarities. Table 1 visually shows some of the similarities and differences among the viruses. The table was based on WHO website.

Table 1
Characteristics of SARS-CoV, MERS-CoV and SARS-CoV

	SARS-CoV-2	SARS-CoV	MERS-CoV
microbiology	Enveloped RNA virus	Enveloped RNA virus	Enveloped RNA virus
Outbreak period	2019-present	2002–2003	2012-present
Initial Site of isolation	Wuhan, China	Guangdong province, China	Saudi Arabia
Countries	214	29	27
No. of cases (mortality)	1,033,187 (2.9%)	8096 (9.6%)	2494 (~ 34%)
Reservoir (intermediate host)	Likely bats (pangolins)	Bats (palm civet)	Bats (dromedary camels)
Incubation period	2-5days (range, 2-14days)	2-7days (range, 2–21)	2–7 (range, 2–14 days)
Infectivity, R0	2.5-3	2.2–3.7 (range, 0.3–4.1)	0.3–1.3
Super spreaders	yes	Yes	Yes (Uncommon)
Transmission (including to HCP)	Droplet/direct, Airborne/Indirect	Droplet/direct, airborne/indirect?	Droplet/direct, airborne/indirect?
Treatment (PEP)	Dexamethasone, Remdesivir	Supportive (none)	Supportive (None)
Infection prevention	Droplet, contact, face shield	Droplet, contact, face shield	Droplet, contact, face shield

Table 2
Decision Tree on SARS-CoV-2, SARS-CoV and MERS-CoV

	Species	Rule (default class 1)	Rule (default class 2)	Rule (default class 3)
9window	SARS-CoV-2	pos2 = D & pos9 = Q	pos2 = K & pos7 = N	
	SARS-CoV			
	MERS-CoV	pos3 = W & pos6 = V	pos1 = M & pos4 = F	
		pos3 = S & pos9 = L	pos1 = P & pos7 = T	
		pos2 = F & pos6 = C	pos5 = G & pos7 = M	
		pos3 = I & pos6 = V	pos4 = H & pos7 = N	
		pos2 = L & pos6 = R	pos4 = D & pos7 = _	
		pos1 = Q & pos2 = G	pos5 = E & pos7 = M	
			pos2 = L & pos6 = R	
			pos2 = E & pos4 = N	
			pos2 = L & pos4 = Y	
			pos1 = Y &pos4 = F	
			pos1 = P & pos4 = K	
13window	SARS-CoV-2		pos12 = D & pos13 = N	
	SARS-CoV	pos1 = D & pos13 = V	pos12 = S & pos13 = N	

	Species	Rule (default class 1)	Rule (default class 2)	Rule (default class 3)
	MERS-CoV	pos1 = D & pos10 = I pos11 = I & pos13 = V pos12 L & pos13 = I pos12 = A & pos13 = P pos12 = V & pos13 = P pos6 = I & pos13 = _ pos7 = Y & pos13 = A pos11 = V & pos13 = L pos3 = A & pos13 = Q pos11 = F & pos 13 = V pos11 = I & pos13 = V pos5 = V & pos13 = P pos11 = H & pos13 = V	pos11 = H & pos13 = V pos5 = V & pos13 = P pos6 = L & pos13 = G pos11 = I & pos13 = E pos11 = I & pos13 = V	
19window	SARS-CoV-2	pos10 = T & pos12 = K		
	SARS-CoV	pos5 = L & pos10 = V pos4 = I & pos7 = K pos10 = I & pos13 = K		

	Species	Rule (default class 1)	Rule (default class 2)	Rule (default class 3)
	MERS-CoV	pos5 = Y & pos10 = V pos4 = L & pos7 = K pos7 = A & pos12 = Y pos7 = I & pos19 = T pos15 = G & pos16 = I pos13 = L & pos15 = K pos 13 = V & pos15 = K pos15 = V & pos16 = A pos15 = V & pos16 = P pos3 = S & pos15 = G pos12 = E & pos16 = G pos3 = S & pos6 = S pos7 = H & pos11 = I pos2 = S & pos15 = Q pos2 = E & pos15 = Q pos4 = T & pos10 = I pos3 = L & pos10 = L pos7 = S & pos15 = T pos15 = V & pos16 = S pos15 = V & pos16 = V		pos13 = V &pos15 = K

Methods

Window

In a peptide sequence, window is a region of a regularly divided peptide sequence. Appropriate window size is important to eliminate variability and to ensure reliable patterns.

FASTA Format

FASTA format converts nucleotide sequences or peptide sequences in a single letter code. This is useful in bioinformatics as nucleotide information can be inserted into text processing tools.

BLAST

BLAST is a program provided by NCBI that is used to compare the biological sequence information. Among several different BLAST programs, we chose Nucleotide-nucleotide BLAST(blastn) which finds DNA sequences that are mostly similar to the query DNA from NCBI DNA database.

Apriori Algorithm

Apriori is an algorithm that finds the frequency of individual items and identifies the relationships among them. Given databases containing itemsets, Apriori algorithm shows the itemsets that are over given threshold.

SVM

SVM is used in classifying, predicting and regressing problems. It classifies samples into categories. It is originally based on Statistical Learning Theory. Each sample is plotted on a n-dimensional space.

Decision Tree

Decision Tree display decisions and their possible consequences. We use this algorithm to specify the difference between two viruses. Among two types of decision tree algorithm usage, classification and regression, our usage is to classify the cases by choosing the right path at each node starting from the root, so that the case reaches a single leaf after satisfying all the conditions of the path.

Experiment Design

We conducted a data analysis on the protein sequence of SARS-CoV-2, SARS-CoV and MERS-CoV using three algorithms: BLAST, Apriori and Decision Tree. Considering the results of those experiments on protein sequences, we concluded that MERS-CoV is remarkably different from SARS-CoV-2 and SARS-CoV. Hence, we decided to conduct further analysis using BLAST, Apriori, SVM and Decision Tree to compare SARS-CoV-2 and MERS-CoV, but this time comparing the codon sequences of the viruses rather than the protein sequences. We expected to earn more accurate and useful results from such experiments since codon sequence is a form of DNA sequence which is more related to actual properties of a virus.

Result of Experiment 1: SARS-CoV-2, SARS-CoV and MERS-CoV

BLAST

First, we briefly experimented BLAST on the three Coronaviruses: SARS-CoV-2, SARS-CoV and MERS-CoV. The result shows that SARS-CoV-2 is almost identical to SARS-CoV while MERS-CoV shows substantial difference in amino acid sequence.

We have experimented the virus in pairs. The BLAST experiment on SARS-CoV-2 and SARS-CoV shows 92% identities, 96% positives and 0% gaps which indicates high similarity. The BLAST experiment on SARS-CoV-2 and MERS-CoV shows 51% identities, 66% positives and 3% gaps which indicates relatively

low similarity. To add, the BLAST experiment on SARS-CoV and MERS-CoV shows 56% identities, 72% positives and 1% gaps.

Apriori

We firstly analysed the genome of SARS-CoV-2, SARS-CoV and MERS-CoV using Apriori algorithm in 9, 13, 19windows. For each window, we set the minimum support as 0.1, so that only the associations appearing more than 10% of the whole instances are regarded as best rules. We define the rule as the tendency of an amino acid A to appear in position N of window, written posN = A. For accurate analysis, we set the minimum metric confidence level as 0.9 and performed the experiment for 18 cycles.

Apriori in 9window The results showed that the most rules involve Leucine in position 5 with large instances in all three genomes. Additionally, in MERS-CoV, Valine appeared frequently in position 4 and 6.

Apriori in 13window The results showed that all three genomes involve Valine in position 1 and Leucine in position 2 with large instances in both genomes. Additionally, in MERS-CoV, Valine appeared frequently in position 2.

Apriori in 19window All three genomes involve Leucine in some positions as one of the best rules with large instances. Additionally, both SARS-CoV-2 and MERS-CoV involve Valine. In SARS-CoV-2, Valine appears frequently in position 4 and in MERS-CoV, Valine is more dominant than Leucine, appearing frequently in position 4, 6, 9, 11, and 13. SARS-CoV only had one best rule, having Leucine in position 1.

These results suggest that Leucine is a commonly significant amino acid in the entire genome of all three genomes. To add, the experiment also suggests that Valine is also a commonly essential amino acid in SARS-CoV-2 and MERS-CoV, especially in MERS-CoV.

Decision Tree

We defined SARS-CoV-2 as class 1, SARS-CoV as class 2 and MERS-CoV as class 3. We compared the data from the start codon to the stop codon. The characteristics written down are rules that had the probability of at least 0.800. This value is high enough to conclude that the species possess a distinguishable trait to the default class.

Decision tree in 9window. The results show that SARS-CoV-2 and MERS-CoV have their unique characteristics that can distinguish them from SARS-CoV-2 and SARS-CoV. However, there weren't any unique characteristics that can differentiate them from MERS-CoV. SARS-CoV does not have a distinct amino acid sequence characteristics compared to the other two viruses. The results show that there are few unique characteristics to distinguish SARS-CoV-2 and MERS-CoV but that SARS-CoV are more similar to the other two viruses. Also, the results show that there are no unique characteristics to distinguish the three viruses from default class 3. This means that all three viruses are similar to default class 3.

Decision tree in 13window. The results show that SARS-CoV-2 has one unique characteristic that can distinguish them from the default 2. SARS-CoV has one distinct characteristic each to default class 1 and

2. MERS-CoV has few unique characteristics that can distinguish them from default class 1 and 2. The results show that there are no unique characteristics to distinguish the three viruses from default class 3. This means that all three viruses are similar to default class 3.

Decision tree in 19window. The results show that SARS-CoV-2 has one unique characteristic that can distinguish them from the default 2. SARS-CoV has three distinct characteristics to default class 1. MERS-CoV has few unique characteristics that can distinguish them from default class 1 and one unique characteristic to default class 2. The results show that there are no unique characteristics to distinguish the three viruses from default class 2. This means that all three viruses are similar to default class 2.

Result of Experiment 2: SARS-CoV-2 and MERS-CoV

Blast

BLASTN program of NCBI is used to analyze the identicality of SARS-CoV-2 and MERS-CoV. The result shows 59% identity and we could see the distribution of top 8 blast hits on the subject sequence.

Therefore, using the remaining three methods, we compared the two DNA sequences and figure out appreciable similarities and differences. Throughout following experiments, we chose to compare orf1ab, the first and the longest ORF, of SARS-CoV-2 and MERS-CoV since it presents the most remarkable difference between two viruses among several ORFs with the same position.

Apriori Algorithm

We firstly analysed the genome of SARS-CoV-2 and MERS-CoV using the Apriori algorithm in 9, 13, 19 windows. Other settings were identical to the previous experiment.

Apriori Algorithm in 9window. Most rules involved Leucine in most positions with large instances in both genomes. Additionally, in MERS-CoV, Valine appeared frequently in position 1, 3, 4, and 8.

Apriori Algorithm in 13window. Most rules involved Leucine in almost all positions with large instances in both genomes. Additionally, in SARS-CoV-2, Valine appeared frequently in position 4. Also, in MERS-CoV, Valine appeared frequently in position 3, 6, 7, and 13.

Apriori Algorithm in 19window. Most rules involve Leucine in almost all positions with large instances in both genomes. Additionally, in SARS-CoV-2, Valine appeared frequently in position 12 and 16; and Threonine also appeared frequently in position 17. Also, in MERS-CoV, Valine appeared frequently in position 2, 13, 14, and 16; Threonine appeared frequently in position 13; and Serine also appeared frequently in position 19.

These results suggest that Leucine is a significant amino acid in the entire genome of both genomes. To add, Valine and Threonine are also essential amino acids in certain positions of both genomes, with MERS-CoV having more Valine as well as Serine.

SVM

The result of Apriori experiment suggests that the DNA sequence of SARS-CoV-2 and MERS-CoV are very similar, having Leucine as their main amino acid. However, the slight difference such as frequency of Valine and Threonine is not neglectable, so for more accurate results SVM algorithm is utilized. The SVM experiment is conducted in 9window, 13window, and 19window with four types of functions: normal, polynomial, RBF, and sigmoid. The experiment method was 10 fold cross validation.

During the experiment, we made data types of < SARS-CoV-2 and MERS-CoV>. Normal SVM experiments have accuracy rates slightly over 50%, which is quite low. This implies that there is some difference between SARS-CoV-2 and MERS-CoV. Polynomial SVM experiment and sigmoid SVM experiment show low accuracy rates. These results support that SARS-CoV-2 and MERS-CoV are difficult to differentiate using linear classifying processes. However, the accuracy rate of RBF, a non-linear kernel, is remarkably high, implying that it is the best chance of classifying the data set.

Decision Tree

We defined SARS-CoV-2 as class 1 and MERS-CoV as class 2. We compared the data from the start codon to the stop codon. Rules that had the probability of at least 0.850 were selected as distinguishable trait. Table 3 shows that SARS-CoV-2 and MERS-CoV have their unique characteristics in all 9, 13, and 19 window. The results show that there are many unique characteristics to distinguish the two viruses.

Table 3
Decision Tree on SARS-CoV and MERS-CoV

Species	Rules in 9window	Rules in 13window	Rules in 19window
SARS-CoV-2	pos3 = L & pos5 = P	pos1 = T & pos10 = G	pos17 = N & pos19 = L
	pos3 = N & pos8 = I	pos5 = L & pos11 = I	pos14 = K & pos18 = L
	pos1 = G & pos3 = V	pos6 = T & pos11 = A	pos12 = T & pos17 = V
		pos2 = R & pos6 = M	pos17 = H
		pos10 = L & pos12 = I	
MERS-CoV	pos1 = Y & pos3 = V	pos10 = Q & pos13 = L	pos4 = V & pos12 = G
	pos1 = V & pos3 = P	pos3 = A & pos10 = T	pos12 = S & pos17 = V
	pos3 = S & pos9 = V	pos6 = C & pos11 = A	pos17 = L & pos18 = V
	pos1 = M & pos3 = V	pos11 = W	
	pos2 = D & pos3 = L	pos5 = S & pos11 = I	
	pos1 = Y & pos3 = V	pos2 = T & pos13 = I	
	pos2 = L & pos3 = Q	pos5 = V & pos11 = D	
	pos1 = Q & pos3 = V	pos6 = V & pos11 = A	
		pos2 = Y & pos4 = S	

Conclusion

Our research is composed of three experiments on SARS-CoV-2, SARS-CoV, and MERS-CoV using three algorithms (BLAST, Apriori, and Decision Tree, and SVM) followed by four further experiments on SARS-CoV-2 and MERS-CoV.

Comparing SARS-CoV-2, SARS-CoV and MERS-CoV, the result of BLAST has shown that SARS-CoV-2 and SARS-CoV had remarkable gap with MERS-CoV. The Apriori experiment specifies that SARS-CoV-2 and SARS-CoV have almost the same distribution of amino acids, having Leucine as an unrivaled main amino acid, while MERS-CoV has high frequency of Valine as well. In Decision tree experiment, all three viruses are similar to MERS-CoV in 9 and 11window. The three viruses are similar to SARS-CoV in 19window.

These experiments showing high similarity as well as remarkable difference between SARS-CoV-2 and MERS-CoV led us to conduct further experiments on those two viruses, this time using the codon sequence of the viruses instead of protein sequence, as codon sequence is more related to the actual properties of the virus.

In further experiments on the codon sequence of SARS-CoV-2 and MERS-CoV, the result of BLAST has shown 59% similarity. The Apriori experiment specifies that the viruses are similar in having Leucine and Valine as their main amino acid, as well as having Threonine frequently appearing. However, the SVM result shows that though the viruses are very alike they can be effectively classified using non-linear kernels such as RBF. Decision Tree experiment has proved several remarkable properties of SARS-CoV-2 amino acid sequence that cannot be found in MERS-CoV amino acid sequence: each 9window, 13window and 19window result has shown characteristic rules of both MERS-CoV and SARS-CoV-2.

Limitations

Not applicable

Abbreviations

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

SARS-CoV: Severe acute respiratory syndrome coronavirus

MERS-CoV: Middle East respiratory syndrome coronavirus

BLAST: Basic Local Alignment Search Tool

SVM: Support Vector Machine

Declarations

(1) Ethics approval and consent to participate: N/A

(2) Consent for publication: Yes

(3) Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. [Email] seunghee991105@gmail.com

(4) Competing interests: The authors declare that they have no competing interests.

(5) Funding: N/A

(6) Authors' contributions: Jung Eun Huh and Seunghee Han has equally contributed to this work. Taeseon Yoon was a supervisor.

(7) Acknowledgements: N/A

References

1. Han, S. and Huh, J., 2017. Data Mining of Influenza A: H3N8, H7N3, And H7N7 - WCSE 2017 - WCSE. [online] Wcse.org. Available at: <http://www.wcse.org/content-14-357-1.html>.
2. Jang S, Lee S, Choi S et al. Comparison between SARS CoV and MERS CoV Using Apriori Algorithm, Decision Tree, SVM. MATEC Web of Conferences 2016;49:08001. doi:10.1051/matecconf/20164908001.
3. Gusnanto, A., C. C. Taylor, I. Nafisah, H. M. Wood, P. Rabbitts, and S. Berri. "Estimating Optimal Window Size for Analysis of Low-coverage Next-generation Sequence Data." Bioinformatics 30.13 (2014): 1823-829.
4. Xu J, Zhao S, Teng T et al. Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV. Viruses 2020;12:244. doi:10.3390/v12020244.