The first report of the most important sequential differences between COVID-19 and MERS viruses by attribute weighting models, the importance of Nucleocapsid (N) protein

Mansour Ebrahimi (mansour@future.edu)
The University of Qom
Boris Novikov
National Research University Higher School of Economics
Esmaeil Ebrahimie
School of Agriculture and Veterinary Sciences, University of Adelaide, Adelaide, Australia
Alexey Spilman
National Research University Higher School of Economics
Reza Ahsan
Islamci Azad University
Mohammad Reza Tahsili
The Uniersity of Qom
Mojtaba Najafi
Peter the Great St. Petersburg Polytechnic University
Samaneh Navvabi
Peter the Great St. Petersburg Polytechnic University
Faridoddin Shariaty
Peter the Great St. Petersburg Polytechnic University

Research Article

Keywords: COVID-19, MERS, Pathogenicity, Attribute Weighting, Nucleocapsid protein

Posted Date: June 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-35367/v1

Abstract

COVID-19 and the Middle East respiratory syndrome-related coronavirus (MERS) viruses are from coronaviridae family; the former became a pandemic while the latter confined to a limited region. Their pathogenicity and infection rates are also different; the high mortality rate for MERS with low spreading capability. To investigate the possible structural changes at RNA sequences of both virus, 1621 and 125 sequences of COVID-19 and MERS downloaded and converted to polynomial datasets and seven attribute weighting (feature selection) approaches have been used for the analysis of genomic sequences of COVID-19 and MERS viruses. The end nucleotide sequences (from 29288 to the end genome positions) selected by the most attribute weighting models to be significantly different between two virus classes followed by smaller piece at 5700 and 1750 and 7600 nucleotide positions. These parts encode Nucleocapsid (N), Papin-like protease (NSP3) and NSP4 proteins of COVID-19. The finding for the first time reports the structural differences between two important viruses at the sequential level and paves the road to decipher new emerging COVID-19 virus high pathogenicity.

Introduction

The coronavirus 2019 (COVID-19) was first identified and reported in patients with severe respiratory disease in Wuhan, China. The virus was a novel member of the coronavirus family which scientifically named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1-3]. Since its discovery, more than four million cases have been infected, including nearly 400,000 who have died. The most worrisome features of COVID-19 are its apparent ability to spread readily, to cause severe disease in high-risk patients and older adults and to mutate and recombine with changes in the genomic sequence since it was first reported [4-6].

Coronaviruses members belong to the subfamily *Coronovirinae* within the family Coronaviridae and the order *Nidovirales* [7]. They are zoonotic pathogens that can be transmitted to human due to direct contact with animals; many scientific reports claimed COVID-19 originated in bats and transmitted to humans via intermediate host animals in the seafood market [8-10]. Coronaviruses genome is a single-stranded positive-sense RNA (+ssRNA) molecule with genomic size ranges between 27–32 kbp which contains at least six open reading frames (ORFs) [11-13]. The first ORFs (ORF1a/b) encodes a polyprotein1a,b (pp1a, pp1b) while other ORFs are located on 3' end encodes at least four structural proteins: envelop glycoprotein spike (S), responsible for recognizing host cell receptors, Membrane (M) proteins, responsible for shaping the virions, the Envelope (E) proteins, responsible for virions assembly and release and the Nucleocapsid (N) proteins are involved in packaging the RNA genome and in the virions and play roles in pathogenicity as an interferon (IFN) inhibitor [14, 15]. In addition to the four main structural proteins, there are structural and accessory proteins that are species-specific, such as HE protein, 3a/b protein, and 4a/b protein. Upon entrance of the viral genome into the cytoplasm of the target cell, the positive-sense RNA genome translates into two polyproteins 1a, b (pp1a, pp1b) and then

are processed into 16 Non-Structural Proteins (NSPs) to form a replication-transcription complex (RTC) that is involved in genome transcription and replication [16, 17].

The COVID-19 or SARS-CoV-2 is the third novel coronavirus to cause a large-scale epidemic or as currently named by WHO as a pandemic in the recent century after the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) in 2003 [18] and the Middle East Respiratory Syndrome Coronavirus (MERS) in 2012 [19]. As a large group of viruses with large peplomers that make it look like a crown, they are common among many animals, coronaviruses can cause respiratory illnesses in humans and gastrointestinal illnesses in animals. Before SARS-CoV epidemy in 2003, this virus family was not considered a deadly virus in human and just caused mild symptoms in immunocompetent people with a chance of the lower respiratory illness like pneumonia and bronchitis; but after 2003 the first and second pandemics of the 21st century; the SARS-CoV and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) reported from China and Saudi Arabi, respectively [20-22]. MERS-CoV with a high fatality rate of 34.4% confirmed to infect approximately 2500 cases including 861 deaths [23].

In the present study, to search for the causes of different pathogenicity and fatality rates of two members of coronaviridae family, we convert the genomic sequences of the COVID-19 and MERS into the polynomial dataset (each nucleotide sat as a variable, therefore, about 30,000 variables generated for each virus RNA sequences) and seven attribute weighting (or feature selection) models applied on the dataset. As some attributes will be more important than others, each attribute weighting model gives each variable weight and normalize the figure into a digit between 0 and 1.0; higher weights reflect the importance of that variable regarding the virus type (a variable with two virus types of COVID-19 and MERS).

Materials And Methods

The flowchart of this study is presented in Figure 1. One thousand and two hundred and sixty-one (1261) COVID-19 sequences and 136 MERS sequences in Fasta format were downloaded from NCIB nucleotide site (https://www.ncbi.nlm.nih.gov/nuccore). The average sequences length for COVID-19 and MERS were 29500 and 29650, respectively. The sequences were converted to a polynomial dataset (single nucleotide treated as an attribute; dataset contained ~ 30,000 attributes or variable) and the target variable with two groups of virus types as label variable. The dataset imported into RapidMiner software (RapidMiner GmbH, Westfalendamm 8744141 Dortmund, Germany) and seven attribute weighting models applied on them as follows:

ATTRIBUTE WEIGHTING

To identify the most important features or attributes (or nucleotide position) that different between COVID-19 and MERS viruses, the following attribute weightings applied on the dataset:

1. **Weight by Information gain**

The Weight by Information Gain operator calculates the weight of attributes concerning the class attribute by using the information gain. The higher the weight of an attribute, the more relevant it is considered. Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect. A notable problem occurs when information gain is applied to attributes that can take on a large number of distinct values. For example, suppose some data that describes the customers of a business. When information gain is used to decide which of the attributes are the most relevant, the customer's credit card number may have high information gain. This attribute has a high information gain because it uniquely identifies each customer, but we may not want to assign high weights to such attributes.

2. **Weight by Information Gain ratio**

The Weight by Information Gain Ratio operator calculates the weight of attributes for the label attribute by using the information gain ratio. The higher the weight of an attribute, the more relevant it is considered. Information gain ratio is used because it solves the drawback of information gain. Although information gain is usually a good measure for deciding the relevance of an attribute, it is not perfect. A notable problem occurs when information gain is applied to attributes that can take on a large number of distinct values. For example, suppose some data that describes the customers of a business. When information gain is used to decide which of the attributes are the most relevant, the customer's credit card number may have high information gain. This attribute has a high information gain because it uniquely identifies each customer, but we may not want to assign high weights to such attributes. The Weight by Information Gain operator uses information gain for generating attribute weights.

3. **Weight by Rule**

The Weight by Rule operator calculates the weight of attributes to the label attribute by constructing a single rule for each attribute and calculating the errors. The higher the weight of an attribute, the more relevant it is considered.

4. **Weight by Chi-squared statistic:**

The Weight by Chi-Squared Statistic operator calculates the weight of attributes concerning the class attribute by using the chi-squared statistic. The higher the weight of an attribute, the more relevant it is considered. Please note that the chi-squared statistic can only be calculated for nominal labels. The chi-square statistic is a nonparametric statistical technique used to determine if a distribution of observed frequencies differs from the theoretically expected frequencies. Chi-square statistics use nominal data, thus instead of using means and variances, this test uses frequencies. The value of the chi-square statistic is given by

X2 = Sigma [(O-E)2/ E]

where X2 is the chi-square statistic, O is the observed frequency and E is the expected frequency. Generally, the chi-squared statistic summarizes the discrepancies between the expected number of times

each outcome occurs (assuming that the model is true) and the observed number of times each outcome occurs, by summing the squares of the discrepancies, normalized by the expected numbers, overall the categories.

5. **Weight by Gini index**

The Weight by Gini Index operator calculates the weight of attributes with respect to the label attribute by computing the Gini index of the class distribution, if the given example set would have been split according to the attribute. The higher the weight of an attribute, the more relevant it is considered. Please note that this operator can be only applied on ExampleSets with the nominal label.

6. **Weight by Uncertainty**

The Weight by Uncertainty operator calculates the weight of attributes with respect to the label attribute by measuring the symmetrical uncertainty with respect to the class. The higher the weight of an attribute, the more relevant it is considered. The relevance is calculated by the following formula:

*relevance = 2 * (P(Class) - P(Class | Attribute)) / P(Class) + P(Attribute)*

7. **Weight by Relief**

Relief is considered one of the most successful algorithms for assessing the quality of features due to its simplicity and effectiveness. The key idea of Relief is to estimate the quality of features according to how well their values distinguish between the instances of the same and different classes that are near each other. Relief measures the relevance of features by sampling examples and comparing the value of the current feature for the nearest example of the same and of a different class. The resulting weights are normalized into the interval between 0 and 1 if the normalize weights parameter is set to true.

Results And Discussion

Seven attribute weighting (or feature selection) models applied on the dataset of 1261 and 136 COVID-19 and MERS sequences. Each attribute weighting computes a weight for each variable (here about 30,000 variables of viruses' sequences generated) in regard to the target or label variable (a column containing two groups of viruses: COVID-19 and MERS). As stated in Materials and Methods, the computed weights are normalized into the interval between 0 and 1.0; higher weight close to 1.0 indicates the target variables of two viruses can be separated with the important variable. Variables (nucleotides) are similar in the same positions for both viruses gained weights near to 0.

One hundred twenty-six variables weighed higher than 0.5 with 6 or 85% attribute weighting models, five nucleotides positions (29579, 29598, 29621, 29652 and 29662) received weights higher than 0.75 and just position (29617) had weights equal to 1 by 60% of attribute weighting models. These findings indicate that around the nucleotide position of 29600, clear structural differences can be traced between COVID-19 and MERS. In COVID-19, this region of sequence encodes for Nucleocapsid proteins (N) that

pack the RNA genome and, in the virions, and play roles in pathogenicity as an interferon (IFN) inhibitor. Although it is a structural protein, in an unknown way it functions in viral replication and localizes to the viral replication-transcription complexes (RTCs). In fact, the nucleocapsid protein packages the viral genomic RNA to form the helical nucleocapsid that is incorporated into the budding particle but also fulfils additional roles during the viral infection. It has been shown to function as an RNA chaperone (33), to facilitate viral RNA synthesis (2, 5, 16) and contributes to the perturbation of several host cellular processes (reviewed in reference 27).

The results of attribute weighting models also suggested three other regions in genomic sequences that are significantly different between COVID-19 and MERS viruses (weights higher than 0.5 by at least 5 attributes weighting out of 7, more than 71% of the models). Those regions code for Papin-like protease (NSP3) and NSP4 proteins.

The results of this study for the first time report the structural differences between two important viruses of coronaviridae at the genomic level and paves the road to decipher new emerging COVID-19 virus high pathogenicity.

References

1. Antonelli G, Capobianchi MR, Riva E. The SARS-CoV-2 epidemic: how the Italian public is being informed. Clin Microbiol Infect. 2020;26(6):791-2. Epub 2020/04/05. doi: 10.1016/j.cmi.2020.03.037. PubMed PMID: 32246996; PubMed Central PMCID: PMCPMC7195308.
2. Bar-On YM, Flamholz A, Phillips R, Milo R. SARS-CoV-2 (COVID-19) by the numbers. Elife. 2020;9. Epub 2020/04/02. doi: 10.7554/eLife.57309. PubMed PMID: 32228860; PubMed Central PMCID: PMCPMC7224694.
3. Biondi Zoccai G, Landoni G, Carnevale R, Cavarretta E, Sciarretta S, Frati G. SARS-CoV-2 and COVID-19: facing the pandemic together as citizens and cardiovascular practitioners. Minerva Cardioangiol. 2020;68(2):61-4. Epub 2020/03/10. doi: 10.23736/S0026-4725.20.05250-0. PubMed PMID: 32150358.
4. Ferioli M, Cisternino C, Leo V, Pisani L, Palange P, Nava S. Protecting healthcare workers from SARS-CoV-2 infection: practical indications. Eur Respir Rev. 2020;29(155). Epub 2020/04/06. doi: 10.1183/16000617.0068-2020. PubMed PMID: 32248146; PubMed Central PMCID: PMCPMC7134482 C. Cisternino has nothing to disclose. Conflict of interest: V. Leo has nothing to disclose. Conflict of interest: L. Pisani has nothing to disclose. Conflict of interest: P. Palange has nothing to disclose. Conflict of interest: S. Nava has nothing to disclose.
5. Hajifathalian K, Mahadev S, Schwartz RE, Shah S, Sampath K, Schnoll-Sussman F, et al. SARS-COV-2 infection (coronavirus disease 2019) for the gastrointestinal consultant. World J Gastroenterol. 2020;26(14):1546-53. Epub 2020/04/25. doi: 10.3748/wjg.v26.i14.1546. PubMed PMID: 32327904; PubMed Central PMCID: PMCPMC7167410.

6. Yu AY, Tu R, Shao X, Pan A, Zhou K, Huang J. A comprehensive Chinese experience against SARS-CoV-2 in ophthalmology. Eye Vis (Lond). 2020;7:19. Epub 2020/04/15. doi: 10.1186/s40662-020-00187-2. PubMed PMID: 32289038; PubMed Central PMCID: PMCPMC7136699.
7. Carneiro Leao J, Paula de Lima Gusmao T, Machado Zarzar A, Leao Filho JC, Barkokebas Santos de Faria A, Morais Silva IH, et al. Coronaviridae - old friends, new enemy! Oral Dis. 2020. Epub 2020/06/01. doi: 10.1111/odi.13447. PubMed PMID: 32475006.
8. Adhikari SP, Meng S, Wu YJ, Mao YP, Ye RX, Wang QZ, et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Infect Dis Poverty. 2020;9(1):29. Epub 2020/03/19. doi: 10.1186/s40249-020-00646-x. PubMed PMID: 32183901; PubMed Central PMCID: PMCPMC7079521.
9. Al-Mandhari A, Samhouri D, Abubakar A, Brennan R. Coronavirus Disease 2019 outbreak: preparedness and readiness of countries in the Eastern Mediterranean Region. East Mediterr Health J. 2020;26(2):136-7. Epub 2020/03/07. doi: 10.26719/2020.26.2.136. PubMed PMID: 32141588.
10. Barry M, Al Amri M, Memish ZA. COVID-19 in the Shadows of MERS-CoV in the Kingdom of Saudi Arabia. J Epidemiol Glob Health. 2020;10(1):1-3. Epub 2020/03/17. doi: 10.2991/jegh.k.200218.003. PubMed PMID: 32175703.
11. Cagliani R, Forni D, Clerici M, Sironi M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. Infect Genet Evol. 2020;83:104353. Epub 2020/05/11. doi: 10.1016/j.meegid.2020.104353. PubMed PMID: 32387562; PubMed Central PMCID: PMCPMC7199688.
12. Cagliani R, Forni D, Clerici M, Sironi M. Computational Inference of Selection Underlying the Evolution of the Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2. J Virol. 2020;94(12). Epub 2020/04/03. doi: 10.1128/JVI.00411-20. PubMed PMID: 32238584.
13. Devendran R, Kumar M, Chakraborty S. Genome analysis of SARS-CoV-2 isolates occurring in India: Present scenario. Indian J Public Health. 2020;64(Supplement):S147-S55. Epub 2020/06/05. doi: 10.4103/ijph.IJPH_506_20. PubMed PMID: 32496247.
14. Chen YW, Yiu CB, Wong KY. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL (pro)) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Res. 2020;9:129. Epub 2020/04/22. doi: 10.12688/f1000research.22457.2. PubMed PMID: 32194944; PubMed Central PMCID: PMCPMC7062204.2.
15. Kannan S, Shaik Syed Ali P, Sheeza A, Hemalatha K. COVID-19 (Novel Coronavirus 2019) - recent trends. Eur Rev Med Pharmacol Sci. 2020;24(4):2006-11. Epub 2020/03/07. doi: 10.26355/eurrev_202002_20378. PubMed PMID: 32141569.
16. Bao L, Deng W, Huang B, Gao H, Liu J, Ren L, et al. The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice. Nature. 2020. Epub 2020/05/08. doi: 10.1038/s41586-020-2312-y. PubMed PMID: 32380511.

17. Buonaguro L, Tagliamonte M, Tornesello ML, Buonaguro FM. SARS-CoV-2 RNA polymerase as target for antiviral therapy. J Transl Med. 2020;18(1):185. Epub 2020/05/07. doi: 10.1186/s12967-020-02355-3. PubMed PMID: 32370758; PubMed Central PMCID: PMCPMC7200052.
18. Fagone P, Ciurleo R, Lombardo SD, Iacobello C, Palermo CI, Shoenfeld Y, et al. Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies. Autoimmun Rev. 2020;19(7):102571. Epub 2020/05/08. doi: 10.1016/j.autrev.2020.102571. PubMed PMID: 32376402; PubMed Central PMCID: PMCPMC7252184.
19. Abd El-Aziz TM, Stockand JD. Recent progress and challenges in drug development against COVID-19 coronavirus (SARS-CoV-2) - an update on the status. Infect Genet Evol. 2020;83:104327. Epub 2020/04/23. doi: 10.1016/j.meegid.2020.104327. PubMed PMID: 32320825; PubMed Central PMCID: PMCPMC7166307.
20. Abboud H, Abboud FZ, Kharbouch H, Arkha Y, Abbadi NE, Ouahabi AE. COVID-19 and SARS-Cov-2 Infection: Pathophysiology and Clinical Effects on the Nervous System. World Neurosurg. 2020. Epub 2020/06/01. doi: 10.1016/j.wneu.2020.05.193. PubMed PMID: 32474093; PubMed Central PMCID: PMCPMC7255736.
21. Abduljalil JM, Abduljalil BM. Epidemiology, genome, and clinical features of the pandemic SARS-CoV-2: a recent view. New Microbes New Infect. 2020;35:100672. Epub 2020/04/24. doi: 10.1016/j.nmni.2020.100672. PubMed PMID: 32322400; PubMed Central PMCID: PMCPMC7171182.
22. Alanagreh L, Alzoughool F, Atoum M. The Human Coronavirus Disease COVID-19: Its Origin, Characteristics, and Insights into Potential Drugs and Its Mechanisms. Pathogens. 2020;9(5). Epub 2020/05/06. doi: 10.3390/pathogens9050331. PubMed PMID: 32365466.
23. Ammad Ud Din M, Boppana LKT. An update on the 2019-nCoV outbreak. Am J Infect Control. 2020;48(6):713. Epub 2020/03/17. doi: 10.1016/j.ajic.2020.01.023. PubMed PMID: 32171622; PubMed Central PMCID: PMCPMC7102631.

Declarations

The authors declare no competing interests.

Figures

Figure 1. The genomic structure of COVID-19 virus, showing the location of Nucleocapsid (N) protein (*adapted from https://doi.org/10.3390/pathogens9050331*).

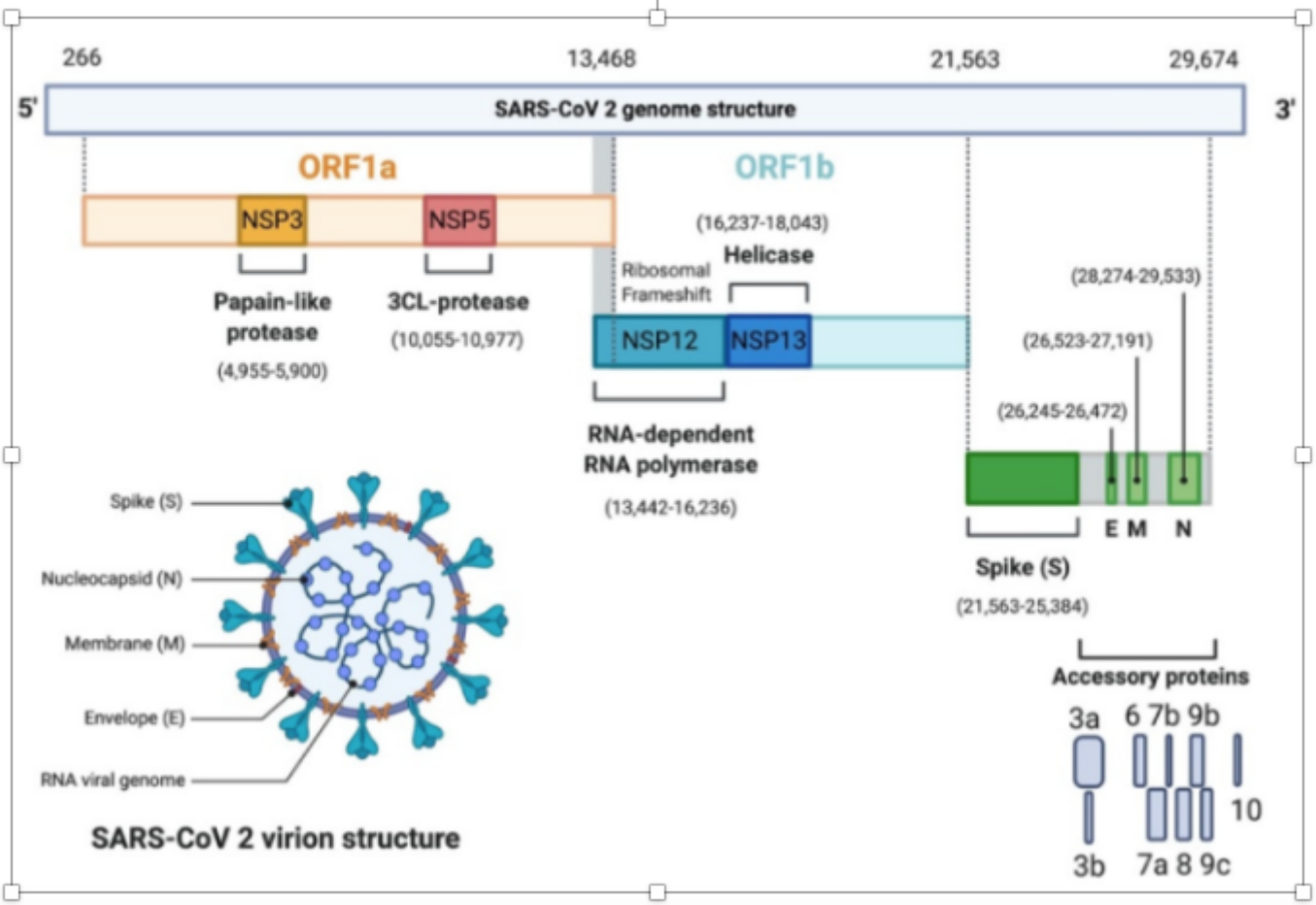

Figure 1

[See figure]