COVID

MDPI

Article

Predicting Transmissibility-Increasing Coronavirus (SARS-CoV-2) Mutations

Ege Çalışkan [1], Murat Işık [1], Cansu İlke Kuru [1,2] and Somenath Chakraborty [3,*]

[1] Buca Municipality Buca Science and Art Center, Buca 35380, İzmir, Turkey; egecaliskan06@gmail.com (E.Ç.); murat12881288@gmail.com (M.I.); cansuilke89@gmail.com (C.İ.K.)
[2] Biotechnology Department, Graduate School of Natural and Applied Sciences, Ege University, Bornova 35100, İzmir, Turkey
[3] Leonard C. Nelson College of Engineering and Sciences, Department of Computer Science and Information Systems, West Virginia University Institute of Technology, Beckley, WV 25801, USA
* Correspondence: somenath.chakraborty@mail.wvu.edu

Abstract: Advantageous variants of the SARS-CoV-2 virus have arisen through mutations, particularly on a single amino acid basis. These point mutations can cause changes in the structure of SARS-CoV-2 and affect the efficiency of interaction with the ACE2 protein. N501Y and E484K mutations affecting binding by ACE2 have been widely observed. This study aimed to predict SARS-CoV-2 mutations that could be as effective as N501Y and E484K and pose a danger due to their high contagiousness. Experimental data on SARS-CoV-2 and ACE2 binding and stability were associated with different amino acid properties and integrated into machine learning and computational biology techniques. As a result of the analyses made in algorithms, N501M, Q414A, N354K, Q498H and N460K have been predicted to be likely to have a dangerous effect. The N501W mutations are most likely to have dangerous effects on the spread of the coronavirus. We suggest that attention should be paid to the position 501 mutation since this position is repeated in the lists of mutations that the algorithm detected as dangerous. G446, G447, Y505, T500, Q493, Y473, and G476 were determined as the positions where dangerous variants could be seen as a result of the analyses of the multiple interaction data created with the ACE2 and RBD interaction data. The 13 dangerous positions and mutations have been detected to accurately describe the position of the mutations caused by the Omicron variant and were among the known dangerous mutations similar to those occurring at Q498, G446, Y505 and Q493 positions.

Keywords: coronavirus; COVID-19 pandemic; prediction algorithm; mutation; bioinformatics

Citation: Çalışkan, E.; Işık, M.; Kuru, C.İ.; Chakraborty, S. Predicting Transmissibility-Increasing Coronavirus (SARS-CoV-2) Mutations. *COVID* **2024**, *4*, 825–837. https://doi.org/10.3390/covid4060055

Academic Editor: Emanuele Pontali

Received: 11 May 2024
Revised: 11 June 2024
Accepted: 12 June 2024
Published: 19 June 2024

1. Introduction

Coronavirus 2 (SARS-CoV-2) is the virus responsible for the global epidemic that continues to infect millions of people worldwide [1–4]. SARS-CoV-2 is the seventh coronavirus identified as a human pathogen. While the HKU1, NL63, OC43, and 229E types of coronaviruses available to date may cause mild symptoms as a result of transmission to humans, SARS-CoV, MERS-CoV, and SARS-CoV-2 viruses cause serious diseases in humans. Coronavirus disease 2019 (COVID-19) caused by SARS-CoV-2 is currently the most important public health problem in the world [5–7]. COVID-19 disease has gone down in history as the biggest epidemic of the last 100 years, causing millions of cases and deaths [8].

SARS-CoV-2 belongs to the subfamily Coronavirinae in the Coronaviridae family of the Nidovirales order and has an enveloped and spherical structure with dimensions of 150–160 nanometers. The positive single-stranded RNA carrying the genetic material of SARS-CoV-2 is protected within a capsule of nucleoprotein. SARS-CoV-2's recognition of its target in the human host is made possible by the S-protein [9]. The Receptor Binding Domain (RBD) located on the upper part of the S-protein binds to the catalytic site of

the human angiotensin-converting enzyme 2 (ACE2) and initiates the infection. ACE2 is responsible for lowering blood pressure by forming angiotensin, a vasodilator, and is a type I membrane protein that can be found in most important organs such as lungs, arteries, heart, and kidney [10]. Due to the important role of RBD in the entry of the virus into the cell, RBD is the target of the most potent anti- -neutralizing antibodies identified to date and is used as the sole antigen in many promising vaccine studies [11].

By binding SARS-CoV-2 to ACE2, the virus will have found the stable environment it needs to enter the cell. Then, by transferring the genetic RNA material into the cell, the virus makes the host–cell produce a series of copies of itself. Newly formed viruses break down the cell, mix with the blood, and select new target cells to initiate the infection. Meanwhile, the immune system prepares an attack against the virus spread, and the antibodies it produces primarily recognize the virus's S-proteins [12]. S-proteins that are neutralized in this way are no longer able to bind to the ACE2 receptor [11]. In this case, the body successfully defeats the infection. However, mutations occur because the virus does not use DNA control mechanisms while making the cell produce its own parts. Since most of the mutations that occur are actually harmful to the virus, they cause the virus to lose its effectiveness. These SARS-CoV-2 variants are evolutionarily eliminated because these mutations are disadvantageous for the virus [13]. However, some types of mutations that happen by chance give the virus an advantage so that the virus continues its task of finding new cells to infect more efficiently. Some of these advantageous mutations cause the S-protein to be better adapted to ACE2 [11,14].

In this case, the rate of transmission of the virus will increase. N501Y ("Nelly") and E484K ("Eeek") mutations, which have been common in England, South Africa, and Brazil since the beginning of 2020, caused the virus's S-protein to bind more tightly to ACE2 in the aforementioned manner. It is also thought to cause some antibodies that previously recognized S-proteins to no longer work [15]. Therefore, the existence and understanding of these mutations is fundamental to drug and vaccine studies.

The global economic recession caused by the COVID-19 pandemic has been unusually severe, resulting in loss of livelihoods and incomes on a global scale. Under these circumstances, a large number of scientists and researchers are making an unprecedented effort to find vaccines and therapeutics to stop the SARS-CoV-2 epidemic [15]. The high mutation rate of the virus, which spreads rapidly as in other RNA viruses, is one of the biggest obstacles to the continued effectiveness of vaccines and drugs to be developed [16–19]. In this context, it is of great importance to elucidate all aspects of the structural and functional features of the genome of SARS-CoV-2.

The B.1.1.7 SARS-CoV-2 variant, which emerged in the UK at the beginning of 2020, has made the virus around 30–50% more contagious and widespread. B.1.1.7 is a change at position 501 in the S-protein to which many neutralizing antibodies bind [20]. Position 501, which was originally asparagine, was converted to the amino acid tyrosine (N501Y) as a result of this mutation. This position is in the RBD and improves S-protein binding to ACE2. The molecular mechanism of this improved binding is still unclear, requiring evaluation of its effects on current therapeutic antibodies [15,21,22]. Again, although we do not have as much information as N501Y about B.1.351, a second variant of SARS-CoV-2 that emerged in South Africa and Brazil at the beginning of 2020, this mutation also appears on the S-protein at position 484 in the RBD. Position 484, originally glutamic acid, was converted to lysine after mutation (E484K). Since its position is in the RBD, it is thought that E484K may show resistance to antibodies produced in people who have had COVID-19 [13].

In September 2020, a seminal article on SARS-CoV-2 was published by Starr et al. [11]. In this article, all possible mutations on the RBD region in the S-protein of SARS-CoV-2 were screened, and it was reported how these mutations affect the binding of SARS-CoV-2 to ACE2. According to this study, which screened 4221 RBD mutations, approximately 14% of these mutations improved S-protein binding to ACE2. The UK, Brazil, and South African variants N501Y and E484K mutations are also included in this 14% group. While there are 586 (14%) different mutations that can improve the binding of S-protein and ACE2,

the fact that only two mutations are common shows that not every mutation that may be advantageous for the virus has an equal chance. This study aimed to find SARS-CoV-2 mutations that will have similar advantages to N501Y and E484K mutations, taking into account the binding and S-protein stability data of Starr et al. In this context, it is aimed to introduce experimental and literature information to different machine learning grouping algorithms to find out which mutations will fall into the same class with N501Y and E484K mutations, and thus to predict the contagiousness levels of future mutations.

2. Methods

2.1. Experimental Data Sets

Experimental data sets used within the scope of the study are given in the link (https://github.com/SARS-CoV-2-Contagiousness/SARS-COV-2, accessed on 21 March 2021).

Experimental data sets were obtained from Starr et al.'s study in 2020 and https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/ (accessed on 13 November 2020). The original of this kit screened the effect of 4221 S-protein mutations onstability, and binding to ACE2. Among these mutations, 586 either did not affect or ameliorate the binding of the S-protein to ACE2. The study focused on this subset, which corresponds to approximately 14% of the main set. For this subset, mutation information, expr_avg, and bind_avg information were obtained from the data set linked above. expr_avg provides information about the stability of the protein, while bind_avg measures how strongly the S-protein binds to ACE2.

2.2. Characteristics of the Amino Acids

Within the scope of the study, the changes in amino acid mutations for the S-protein Receptor Binding Domain (RBD) in wild- and mutant-type structures were examined on the basis of 8 different features based on the experimental data set obtained from the article by Starr et al. These amino acid properties were determined as a result of the literature research on hydropathy, polarity, volume, molecular weight, ring number in amino acid structure, oxygen number, hydrogen number, and double bond number within the scope of their effects on bonding average and stability average.

The changes in the characteristics given in Table 1 for the mutations in the selected data set were calculated with the written Python codes. Finally, mutation and amino acid change information is tabulated as a single csv table. The grouping of the obtained table was provided by the Weka program [23].

Table 1. Characteristics of the amino acids.

Aminoacid Features	Purpose of Use in the Study	References
Volume	Classification based on determining the volume change, especially in side chains, due to mutation.	[24]
Hydropathy	Classification based on determining the change in the hydrophobicity of the side chains due to mutation.	[24]
Moleculer Mass	Classification based on determining the change in molecular mass of amino acids due to mutation.	-
Polarity	Classification based on determining the change in the charge state of amino acids due to mutation.	[23]
The number of cyclic structures in the amino acid structure	Classification based on determination of changes in polarity and hydropathy properties due to the aromatic ring structure and number in the R-group of the mutation-induced amino acid structure.	-
Oxygen number in amino acid structure	Classification based on the difference in the number of oxygen atoms due to the change in the amino acid structure due to mutation.	-

Table 1. *Cont.*

Aminoacid Features	Purpose of Use in the Study	References
Hydrogen number in amino acid structure	Classification based on the difference in the number of hydrogen atoms due to a change in the amino acid structure due to mutation. This property can also be associated with the number of cyclic structures. Since the carbon atoms at the ends are bonded to each other in ring structures, the number of hydrogen atoms the compound has naturally decreases. When the two ends of the compound are combined, one hydrogen atom will be removed from each end, so in cyclic structures with the same carbon number, two hydrogen atoms are missing compared to straight chains.	-
The number of double bonds in the amino acid structure	Classification based on the difference in the number of double bonds due to changes in amino acid structure due to mutation.	-

2.3. Programs and Algorithms

2.3.1. Machine Learning Data Analysis Program, WEKA

In the study, the machine learning program WEKA (Version 3.8.5.) was used for grouping. WEKA (Waikato Environment for Knowledge Analysis) is a program that can be used for data analysis and forecasting. In this study, K-means clustering and expectation maximization (EM) algorithms were used to observe the distribution of the data and extract meaning and information from the data.

2.3.2. K-Mean Clustering

K-means clustering, which is an algorithm within the Weka data analysis program, is an unsupervised machine learning algorithm that works with the "k" variable to be determined by the user. The "k" variable represents the number of clusters needed before starting the algorithm. This clustering method divides a data set consisting of N data objects into k clusters given as input parameters. The aim here is to ensure that the clusters obtained at the end of the partitioning process have maximum similarities within clusters and minimum similarities between clusters. K-means is one of the most commonly used and easy-to-implement clustering algorithms. This algorithm can cluster large-scale data quickly and effectively [23,25].

2.3.3. Expectation Maximization (EM) Algorithm

The expectation maximization algorithm is an iterative search method used to find the largest likelihood or the largest aftereffects estimates of the parameters of statistical models based on unobservable hidden variables. It occurs by repeating the expectation (E) and maximization (M) steps in succession. The expectation step generates a log-likelihood expectation function using current estimates of the parameters. The maximization step updates the parameter values to maximize the log-likelihood expectation. That is, each of these two steps feeds on each other by calculating the input of the other. Expectation maximization steps are repeated until the amount of error in the estimation falls below a certain rate [23,25].

2.4. Interaction-Based Predictive Analytics

First of all, the SARS_CoV_2—ACE2 Multiple Interaction Data set was created for interaction-based predictive analysis. In the written Python code, the ACE2 and RBD single interaction data, which were created by calculating from the atomistic structure, were converted into multiple interaction data. (Github page, "Data_Edit.ipynb") Here, the features were rendered as the sum of the features of the ACE2 amino acids with which the

RBD interacts. The obtained data were analyzed by changing various algorithms and cluster numbers in the WEKA program. The algorithm giving optimal results was determined as the 6-cluster K-Means algorithm, and dangerous interactions were determined as a result of the analysis.

3. Results

Within the scope of the study, experimental data of Starr et al.'s mutation-induced change in S-protein stability and binding to ACE2 were obtained. This data set contains 4221 S-protein mutations. Using the typed Python code, data for 586 experimental mutations were extracted from this data set that either enhanced or did not affect the binding of the S-protein to ACE2. Then, these mutations were also coded in terms of changes in hydropathy, polarity, volume, molecular weight, ring number in amino acid structure, oxygen number, hydrogen number, and double bond number and correlated with the experimental data. The data set was created in this way and all the Python codes can be accessed in the link https://github.com/SARS-CoV-2-Contagiousness/SARS-COV-2 (accessed on 25 July 2021).

In total, 586 mutations, including N501Y and E484K variants, which were subsets within the above method, were introduced into two different machine learning grouping algorithms. Then, for different group (cluster) numbers, it was investigated which mutations would fall into the same cluster with N501Y and E484K mutations. The aim here was to weed out dangerous SARS-CoV-2 mutations that could cause similar changes to N501Y and E484K mutations. For this purpose, we tested our data set with various algorithms in the Weka program. In addition, a data set containing multiple interactions of RBD and ACE2 was created, and positions where dangerous mutations could occur were analyzed.

3.1. K-Means Clustering Algorithm Analysis Results

Using the k-means clustering algorithm, the whole data set was divided into different clusters with the number of clusters in the range of 2–12 (Table 2).

Table 2. K-means clustering algorithm analysis results.

Number of Clusters	Cluster Number Where the E484K Mutation Is Found	The Cluster Number Where the N501Y Mutation Is Located	The Number of Elements in the Set with the E484K Mutation	Number of Elements in the Set with the N501Y Mutation
2	1	1	445	445
3	1	1	299	299
4	3	3	164	164
5	3	3	164	164
6	5	3	118	66
7	5	3	119	65
8	1	1	57	57
9	8	8	56	56
10	**9**	**9**	**30**	**30**
11	9	9	30	30
12	9	9	29	29

When the clustering statistics given in Table 2 were examined, it was seen that the number of clusters containing the N501Y and E484K mutations did not change since the value of K = 10. Considering this, it was decided to divide our data set into 10 separate clusters using the K-means algorithm. In addition, it was observed that N501Y and E484K mutations fell into the same cluster for K = 2, 3, 4, 5, 8, 9, 10, 11, and 12 values in many clusters. As the algorithm was able to correctly put the contagious mutations in the same

cluster, we claim that it is able to categorize the mutations based on unseen relationships in the data; therefore, it is reliable.

This indicates the reliability of our clustering approach. As seen in Figures 1 and 2, the amino acid properties we used explain the S-protein binding and stability to ACE2 differently. This shows the necessity of using each feature. Looking at cluster 9 with 30 elements, the top five mutations with the highest stability were identified as follows:

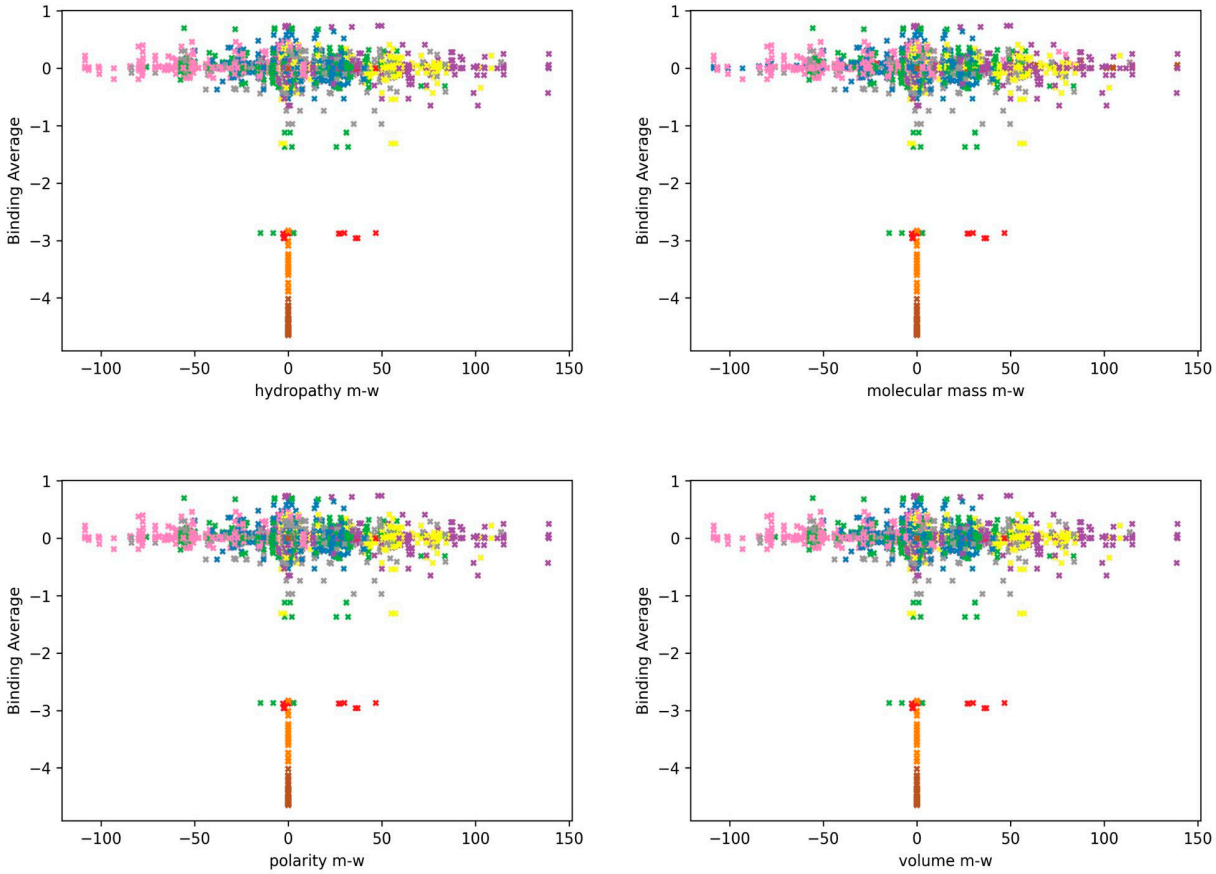

Figure 1. Analysis results of K-means clustering algorithm and K = 10 parameters to examine the binding effects of 4 different amino acid properties. Each graph has the bond strength on the y-axis (the higher the better) and on the x-axis from left to right: hydropathy, molecular weight, polarity, and volume, respectively. Data distributions are colored according to different groups.

N501M, Q414A, Q498H, N460K and **N501W (stability is sorted from most to least)**.

All mutations given in this list have been identified as the five most dangerous mutations suggested by K-clustering. The fact that the 501 position is listed twice in these mutations shows that care should be taken about the mutations that may occur at this position.

3.2. Expectation Maximization Clustering Algorithm Analysis Results

Using the expectation maximization clustering algorithm, the whole data set was clustered individually in multiple rounds. In each round, the algorithm was given a "K" value and asked to divide the data into "K" clusters. (Table 3).

At the clustering statistics given in Table 3, it is seen that the number of clusters containing N501Y and E484K mutations increased since K = 10 for this algorithm. Considering this, it was decided that dividing our data set into 10 separate clusters with the EM algorithm is the most appropriate parametric case. In addition, it was observed that the N501Y and E484K mutations fell into the same cluster in the groupings.

As seen in Figures 3 and 4, the amino acid properties we used explain the binding and stability of the S-protein to ACE2 in a different way. When we look at the eighth cluster

with 46 elements calculated for K = 10, the first five mutations with the highest stability were determined as follows:

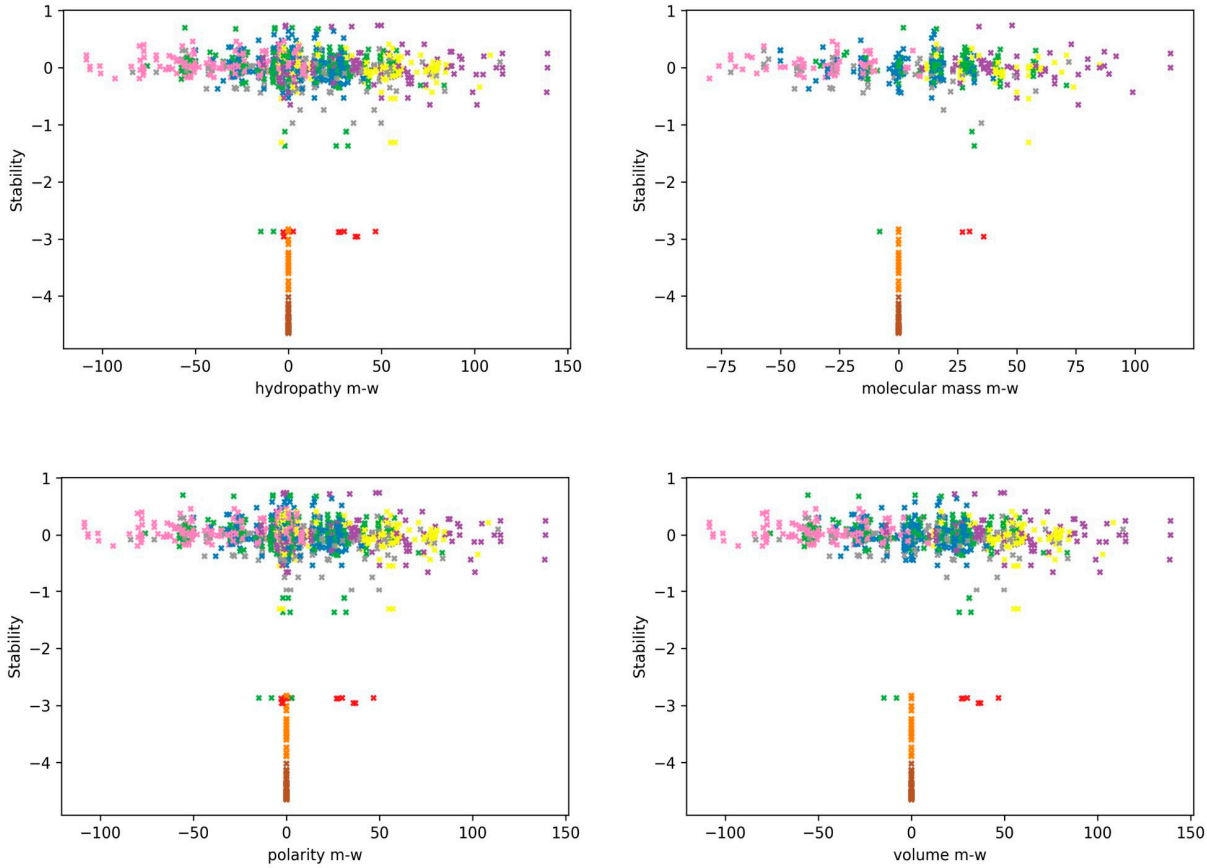

Figure 2. Analysis results of the K-means clustering algorithm and the K = 10 parameters to examine the effects of 4 different amino acid properties on the stability of the S-protein. Each graph has the bond strength on the y-axis (the higher the better) and on the x-axis from left to right: hydropathy, molecular weight, polarity, and volume, respectively. Data distributions are colored according to different groups.

Table 3. Expectation maximization clustering algorithm results.

Number of Clusters (K)	Cluster Number Where the E484K Mutation Is Found	The Cluster Number Where the N501Y Mutation Is Located	The Number of Elements in the Set with the E484K Mutation	Number of Elements in the Set with the N501Y Mutation
2	1	1	449	449
3	0	0	342	342
4	0	0	175	175
5	0	0	63	63
6	2	2	57	57
7	2	2	61	61
8	4	4	56	56
9	4	4	51	51
10	**8**	**8**	**46**	**46**
11	3	3	49	49
12	3	3	49	49

N501M, N354K, Q498H, N460K and **N501W (stability is sorted from most to least)**.

In this algorithm, similar to the K-means algorithm, the same mutations belonging to the 501 position were selected in the list of the five most dangerous mutations.

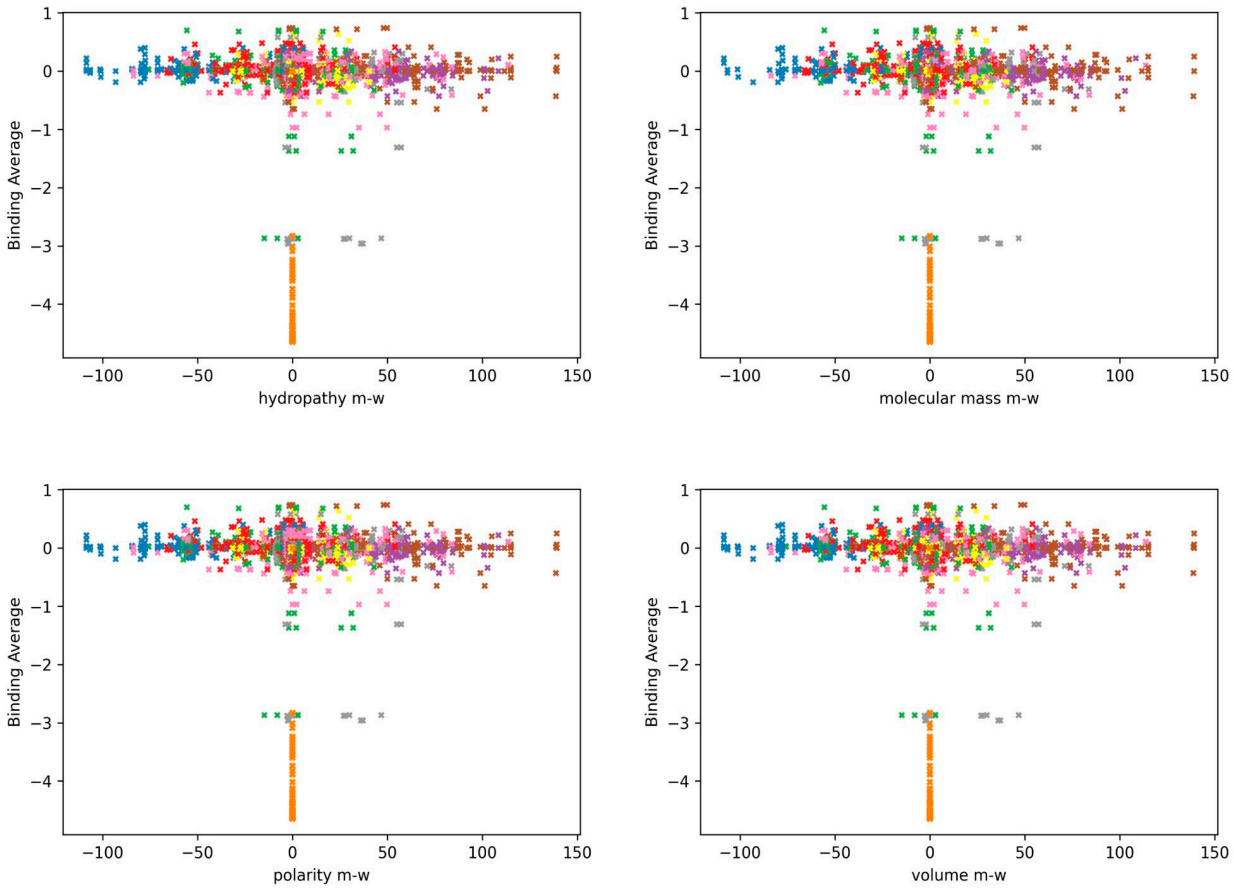

Figure 3. Analysis results of the expectation maximization clustering algorithm to examine the effect of S-protein binding to ACE2 of 4 different amino acid properties as a result of the K = 10 parameters. Each graph has the bond strength on the y-axis (the higher the better) and on the x-axis from left to right: hydropathy, molecular weight, polarity, and volume, respectively. Data distributions are colored according to different groups. Labels of different groups identified by the algorithm are omitted, as the study focuses only on a single cluster, which will be detailed in the rest of the paper.

3.3. Common Analysis Results of EM and K-Means Clustering Algorithms

As a result of the analysis of the clusters obtained from the two algorithms above, it is seen that there are common mutations in the list of the five most dangerous mutations. With the evaluation of these common mutations, as a consensus, six mutations were identified as mutations with a high probability of having dangerous effects in the spread of the coronavirus. The mutations are as follows:

N501M, Q414A, N354K, Q498H, N460K and **N501W**

It was particularly suggested that special attention should be paid to the 501st position mutation seen in one of the common variants because of the repetition of position 501 in this list.

3.4. Positions Where Dangerous SARS-CoV-2 Variants Can Be Seen as a Result of Interaction-Based Predictive Analysis

At the end of the analyses made on the multiple interaction data we created with the ACE2 AND RBD interaction data, the positions that fall into the same cluster as K417 are G446 and G447; positions in the same cluster as N501 are Y505, T500, and Q493; the positions falling in the same cluster as E484 were determined as Y473 and G476.

As can be seen in Table 4, this number was chosen as the optimal number of clusters, since the variants identified as dangerous were separated into six clusters. Next, it was investigated which variants would fall in the same cluster as N501, K417, and E484K variants.

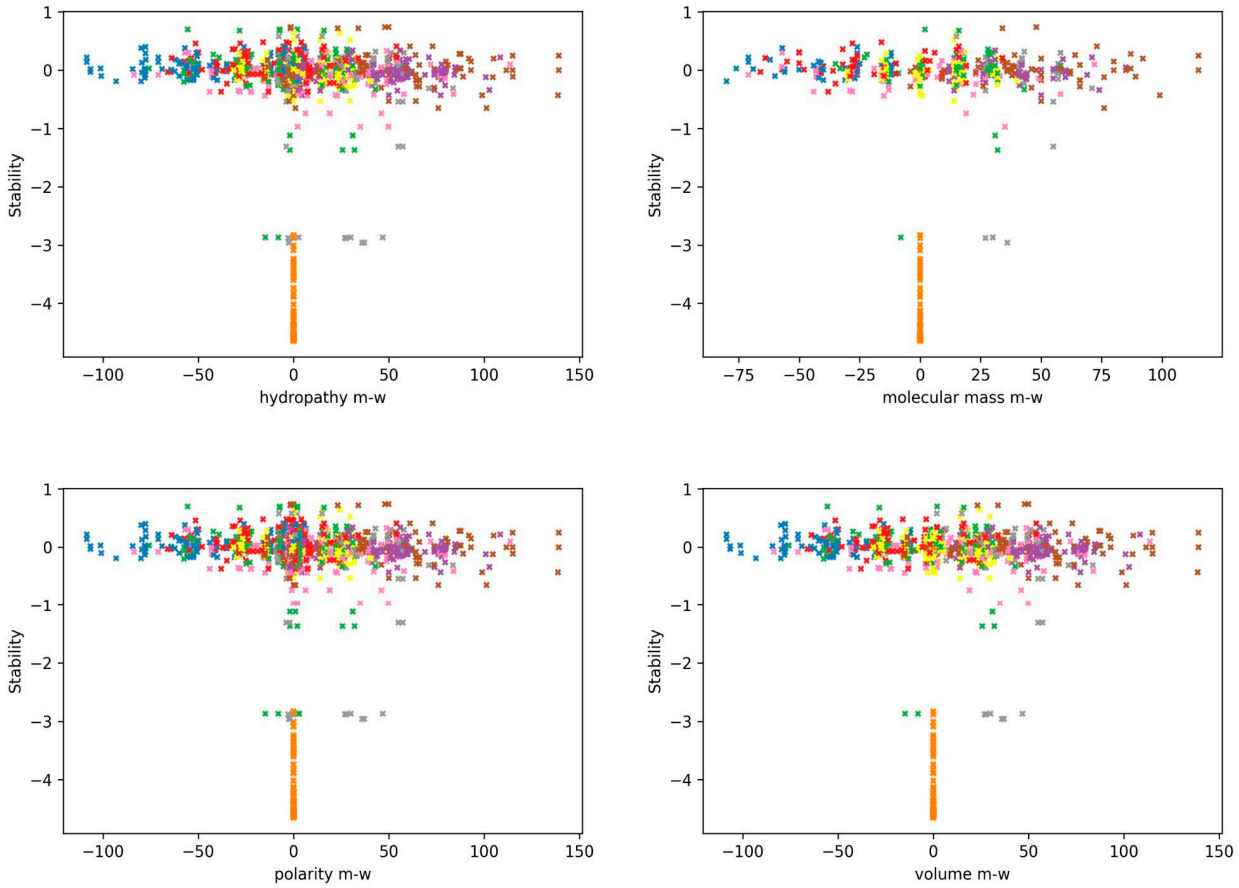

Figure 4. Analysis results of the expectation maximization clustering algorithm to examine the effect of 4 different amino acid properties on the stability of the S-protein as a result of the K = 10 parameter. Each graph has the bond strength on the y-axis (the higher the better) and on the x-axis from left to right: hydropathy, molecular weight, polarity, and volume, respectively. Data distributions are colored according to different groups.

Table 4. K-Means analysis results in ACE2-RBD data set.

Number of Clusters	Number of Elements in Cluster 0	Number of Elements in Cluster 1	Number of Elements in Cluster 2	Number of Elements in Cluster 3	Number of Elements in Cluster 4	Number of Elements in Cluster 5	Number of Elements in Cluster 6
2	11 (%52 K417 and K484)	10 (%48 Y501)	0	0	0	0	0
3	7 (%33 K417 and K484)	6 (29)	8 (%38,Y501)	0	0	0	0
4	5 (%24 484K)	3 (%14)	8 (%38, Y501)	8 (%38,Y501)	0	0	0
5	3 (%14 K484)	5(%24)	2(%10)	3 (%14 K417)	8 (%38, Y501)	0	0
6	3 (%14 K484)	5(%24)	2(%10)	3 (%14 K417)	4 (%19)	4 (%19 501Y)	0
7	3 (%14 K484)	5(%24)	2(%10)	2 (%10)	4(%19)	4 (%19 501Y)	1 (%5 K417)

3.5. Prediction Accuracy

Of the 13 detected dangerous positions, position Q498, position G446, position Y505, and position Q493 accurately identify mutation positions caused by the Omicron variant, which has been declared a variant of concern by the World Health Organization. One of these determinations is obtained from our first analysis and three from our second analysis. Among the detected dangerous mutation positions, Q498 position, G446 position, Y505 position, and Q493 position mutations are observed in the Omicron variant.

3.6. Additional Analysis of RBD and ACE2 Interaction Data

In the second analysis, we examined interactions instead of mutations. We converted the RBD-ACE2 binary interaction data generated by Star et al. into multiple interaction data by processing it in Python code. In this analysis, we approached the problem from a different angle and made a more holistic analysis by examining not only RBD but also RBD and ACE2 together.

In our Python code, we converted ACE2 and RBD single interaction data, which we calculated ourselves from the atomistic structure, into multiple interaction data. Details about this can be found in the "Data_Edit.ipynb" file on our github page.

Our new data are based on interactions, so it gives us a different perspective. Here, we have processed the features as the sum of the features of the ACE2 amino acids with which RBD interacts. Details about this can be found in the "Data_Edit.ipynb" file as well.

We tried various algorithms and cluster numbers on our new data in the Weka program. We determined the most optimal algorithm as the 6-cluster K-Means algorithm. We identified dangerous interactions as a result of our analysis using this algorithm:

Positions falling in the same cluster as K417: G446 and G447
Positions falling in the same cluster as N501: Y505, T500, and Q493
Positions falling in the same cluster as E484: Y473 and G476

4. Discussion

The global epidemic caused by a new type of human coronavirus, SARS-CoV-2, is a concern for all humanity. The COVID-19 infection pandemic created by the SARS-CoV-2 virus appears to have become a major public health problem, given its scale and rapid spread. Considering the worldwide size of the pandemic we are living in, collaboration among expert researchers in different fields has been maximized in the race against time in the fight against coronavirus. In addition, it has become crucial to follow approaches that combine different disciplines and prioritize collective thinking in order to find effective solutions to the common threat immediately. In this context, the research results to be obtained in the field of computational biology and machine learning will provide preliminary information on detailed laboratory studies and will help in experimental planning, interpretation of analyses, and taking various treatment and health measures.

The study's objective was to identify and eliminate harmful SARS-CoV-2 mutations that might alter N501Y and E484K alterations similarly. We used the Weka application to test our data set using different techniques for this purpose. Furthermore, sites where potentially harmful mutations could arise were examined, and a data set comprising many interactions between RBD and ACE2 was generated. Based on the acquired results, the K-means and expectation maximization algorithms identified the five most harmful mutations, which are listed in this list. The fact that these mutations list position 501 twice indicates that the mutations at this place should be taken into consideration. The positions that fall in the same cluster as K417 were G446 and G447; positions in the same cluster as N501 were Y505, T500, and Q493; positions in the same cluster as E484 were identified as Y473 and G476. These findings are the consequence of the analysis we conducted on the multiple interaction data we created with ACE2 AND RBD interaction data. The variants that belong to the same cluster as the N501, K417, and E484K variants were then examined. Locations Q498, G446, Y505, and Q493 accurately identified the mutation locations caused by the Omicron variation, which has been designated as a variant of concern by the World

Health Organization, out of the 13 risky positions detected. In the second analysis, we used a different approach and conducted a more comprehensive analysis by looking at interactions rather than mutations by analyzing not just RBD but also RBD and ACE2 jointly. Using the Python code, we processed the RBD-ACE2 paired interaction data produced by Star et al. to create multiple interaction data. Using our new data, we experimented with different cluster counts and algorithms in the Weka software. The 6-cluster K-Means algorithm was found to be the most appropriate algorithm. Using this approach, we conducted analysis and found potentially harmful interactions as follows: Y505, T500, Q493, and E484 positions in the same cluster as Y473 and G476; K417 positions in the same cluster as G446 and G447; positions in the same cluster as N501.

For this purpose, this study aimed to detect SARS-CoV-2 mutations, which can cause similar changes to N501Y and E484K mutations, which are commonly observed in England, South Africa, and Brazil and cause concern all over the world. These mutations may pose a danger due to their high contagiousness, and the estimation of interaction positions was performed successfully.

5. Conclusions

The scientific world is experiencing a dynamic pandemic process where any work to be carried out to determine and predict the transmission rate and method of the virus will both provide serious benefits and undergo testing. In this area, the development of the possibility of accelerating the transmission of SARS-CoV-2 mutations with prediction algorithms provides strong evidence that richer and more productive solutions will be obtained in the fight against the virus. Therefore, such studies will guide the spread and progression of the virus.

6. Future Perspectives

In future studies, homology modeling and MD simulations should be performed. Homology modeling is a technique employed to construct three-dimensional protein structures based on the primary sequences of proteins, drawing from previous insights derived from structural parallels with other proteins. The homology modeling procedure involves a series of systematic stages, encompassing the optimization of sequence/structure alignment, followed by the construction of a backbone, and subsequent addition of side chains. After modeling the low-homology loops, the comprehensive three-dimensional structure is then fine-tuned and authenticated. In this concept, spike protein or only the RBD should be evaluated with the homology models with the newly founded amino acids.

Also, the mutated residues' exposition to the solvent and the ACE2 receptor should be evaluated by molecular dynamics (MD) simulations that utilize a universal model of the physical laws dictating interatomic relationships to forecast the trajectory of each atom within a protein or another molecular system as it evolves through time. These types of studies are complex and require well-experienced personnel, high-performance computational resources, and time. The implementation of these strategies by providing resources can be considered as a future perspective.

Author Contributions: Conceptualization, E.Ç., M.I. and C.İ.K.; methodology, E.Ç., M.I. and C.İ.K.; software, E.Ç. and M.I.; validation, E.Ç., M.I. and C.İ.K.; formal analysis, E.Ç. and M.I.; investigation, E.Ç., M.I. and C.İ.K.; resources, E.Ç., M.I. and C.İ.K.; data curation, E.Ç. and M.I.; writing—original draft preparation E.Ç., M.I., C.İ.K. and S.C.; writing—review and editing, C.İ.K. and S.C.; visualization, E.Ç., M.I. and C.İ.K.; supervision, C.İ.K.; project administration, C.İ.K.; funding acquisition, C.İ.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank Ezgi Karaca and Burcu Özden from İzmir Biomedicine and Genome Center for their support during the research process.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rodrigues, J.P.; Barrera-Vilarmau, S.; Mc Teixeira, J.; Sorokina, M.; Seckel, E.; Kastritis, P.L.; Levitt, M. Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLoS Comput. Biol.* **2020**, *16*, e1008449. [CrossRef] [PubMed]
2. Bhaskar, L.V.K.S.; Roshan, B.; Nasri, H. The fuzzy connection between SARS-CoV-2 infection and loss of renal function. *Am. J. Nephrol.* **2020**, *51*, 572–573. [CrossRef] [PubMed]
3. Cucinotta, D.; Vanelli, M. WHO declares COVID-19 a pandemic. *Acta Bio Med. Atenei Parm.* **2020**, *91*, 157. [CrossRef] [PubMed]
4. Hoffmann, M.; Kleine-Weber, H.; Pöhlmann, S. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell* **2020**, *78*, 779–784. [CrossRef] [PubMed]
5. Mahase, E. Covid-19: Death rate is 0.66% and increases with age, study estimates. *BMJ Br. Medi. J.* **2020**, *369*, m1327. [CrossRef] [PubMed]
6. Verma, H.K.; Merchant, N.; Verma, M.K.; Kuru, C.İ.; Singh, A.N.; Ulucan, F.; Verma, P.; Bhattacharya, A.; Bhaskar, L.V.K.S. Current updates on the European and WHO registered clinical trials of coronavirus disease 2019 (COVID-19). *Biomed. J.* **2020**, *43*, 424–433. [CrossRef] [PubMed]
7. Walls, A.C.; Park, Y.J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **2020**, *181*, 281–292. [CrossRef] [PubMed]
8. World Health Organization. COVID-19 Weekly Epidemiological Update. Available online: https://data.who.int/dashboards/covid19/cases (accessed on 25 May 2024).
9. Kannan, S.; Ali, P.S.S.; Sheeza, A.; Hemalatha, K. COVID-19 (Novel Coronavirus 2019) recent trends. *Eur. Rev. Med. Pharmacol. Sci.* **2020**, *24*, 2006–2011. [PubMed]
10. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **2020**, *182*, 812–827. [CrossRef]
11. Starr, T.N.; Greaney, A.J.; Hilton, S.K.; Ellis, D.; Crawford, K.H.; Dingens, A.S.; Navarro, M.J.; Bowen, J.E.; Tortorici, M.A.; Walls, A.C.; et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **2020**, *182*, 1295–1310. [CrossRef]
12. Li, G.; Pahari, S.; Murthy, A.K.; Liang, S.; Fragoza, R.; Yu, H.; Alexov, E. SAAMBE-SEQ: A sequence-based method for predicting mutation effect on protein-protein binding affinity. *Bioinformatics* **2021**, *37*, 992–999. [CrossRef]
13. Nelson, G.; Buzko, O.; Spilman, P.; Niazi, K.; Rabizadeh, S.; Soon-Shiong, P. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y. V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *bioRxiv* **2021**. [CrossRef]
14. Seah, I.; Su, X.; Lingam, G. Revisiting the dangers of the coronavirus in the ophthalmology practice. *Eye* **2020**, *34*, 1155–1157. [CrossRef] [PubMed]
15. Luan, B.; Wang, H.; Huynh, T. Molecular Mechanism of the N501Y Mutation for Enhanced Binding between SARS-CoV-2's Spike Protein and Human ACE2 Receptor. *bioRxiv* **2021**. [CrossRef]
16. Van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.; Ford, C.; Ormond, L.; Owen, C.; Pang, J.; Tan, C.; Boshier, F.; et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *83*, 104351. [CrossRef]
17. Pachetti, M.; Marini, B.; Benedetti, F.; Giudici, F.; Mauro, E.; Storici, P.; Masciovecchio, C.; Angeletti, S.; Ciccozzi, M.; Gallo, R.; et al. Emerging SARS CoV-2 mutation hot spots ınclude a novel RNA dependent-RNA polymerase variant. *J. Transl. Med.* **2020**, *18*, 179. [CrossRef] [PubMed]
18. Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *81*, 104260. [CrossRef] [PubMed]
19. Rajgor, D.D.; Lee, M.H.; Archuleta, S.; Bagdasarian, N.; Quek, S.C. The many estimates of the COVID-19 case fatality rate. *Lancet Infect. Dis.* **2020**, *20*, 776–777. [CrossRef]
20. Gu, H.; Chen, Q.; Yang, G.; He, L.; Fan, H.; Deng, Y.Q.; Wang, Y.; Teng, Y.; Zhao, Z.; Cui, Y.; et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **2020**, *369*, 1603–1607. [CrossRef]
21. Jiang, F.; Deng, L.; Zhang, L.; Cai, Y.; Cheung, C.W.; Xia, Z. Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). *J. Gen. Intern. Med.* **2020**, *35*, 1545–1549. [CrossRef]
22. Jia, Y.; Shen, G.; Nguyen, S.; Zhang, Y.; Huang, K.S.; Ho, H.Y.; Hor, W.S.; Yang, C.H.; Bruning, J.B.; Li, C.; et al. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv* **2021**. [CrossRef]
23. Sun, T.; Zhou, B.; Lai, L.; Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinform.* **2017**, *18*, 277. [CrossRef] [PubMed]

24. Amengual-Rigo, P.; Fernández-Recio, J.; Guallar, V. UEP: An open-source and fast classifier for predicting the impact of mutations in protein-protein complexes. *Bioinformatics* **2021**, *37*, 334–341. [CrossRef] [PubMed]
25. Stamp, M. A survey of machine learning algorithms and their application in information security. In *Guide to Vulnerability Analysis for Computer Networks and Systems*; Springer: Cham, Switzerland, 2018; pp. 33–55. [CrossRef]