Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Prediction of infectivity of SARS-CoV2: Mathematical model with analysis of docking simulation for spike proteins and angiotensin-converting enzyme 2

Yutaka Takaoka \(a,b,c,d,e,1,\ast\), Aki Sugano \(d,f,1\), Yoshitomo Morinaga \(g\), Mika Ohta \(a,b,d,e\), Kenji Miura \(b\), Haruyuki Kataguchi \(a,b\), Minoru Kumaoka \(b,d\), Shigemi Kimura \(d\), Yoshimasa Maniwa \(d\)

\(a\) Department of Computational Drug Design and Mathematical Medicine, Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama, Toyama 930-0194, Japan
\(b\) Data Science Center for Medicine and Hospital Management, Toyama University Hospital, Toyama 930-0194, Japan
\(c\) Center for Advanced Antibody Drug Development, University of Toyama, Toyama 930-0194, Japan
\(d\) Department of Medical Systems, Kobe University Graduate School of Medicine, Kobe, Hyogo 650-0017, Japan
\(e\) Life Science Institute, Kobe Tokiwa University, Kobe, Hyogo 653-0838, Japan
\(f\) Center for Clinical Research, Toyama University Hospital, Toyama 930-0194, Japan
\(g\) Department of Microbiology, Toyama University Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama, Toyama 930-0194, Japan

ARTICLE INFO

Keywords:
SARS-CoV-2
COVID-19
Infectivity
Spike protein
Binding affinity
Mathematical model

ABSTRACT

Objectives: Variants of a coronavirus (SARS-CoV-2) have been spreading in a global pandemic. Improved understanding of the infectivity of future new variants is important so that effective countermeasures against them can be quickly undertaken. In our research reported here, we aimed to predict the infectivity of SARS-CoV-2 by using a mathematical model with molecular simulation analysis, and we used phylogenetic analysis to determine the evolutionary distance of the spike protein gene (S gene) of SARS-CoV-2.

Methods: We subjected the six variants and the wild type of spike protein and human angiotensin-converting enzyme 2 (ACE2) to molecular docking simulation analyses to understand the binding affinity of spike protein and ACE2. We then utilized regression analysis of the correlation coefficient of the mathematical model and the infectivity of SARS-CoV-2 to predict infectivity.

Results: The evolutionary distance of the S gene correlated with the infectivity of SARS-CoV-2 variants. The calculated binding affinity for the mathematical model obtained with results of molecular docking simulation also correlated with the infectivity of SARS-CoV-2 variants. These results suggest that the data from the docking simulation for the receptor-binding domain of variant spike proteins and human ACE2 were valuable for prediction of SARS-CoV-2 infectivity.

Conclusion: We developed a mathematical model for prediction of SARS-CoV-2 variant infectivity by using binding affinity obtained via molecular docking and the evolutionary distance of the S gene.

1. Introduction

Variants of the novel coronavirus (SARS-CoV-2 [severe acute respiratory syndrome coronavirus 2]) that are responsible for the worldwide pandemic known as COVID-19 have led to difficulties in enacting countermeasures against infection, because many variants have occurred in succession (Bar-On et al., 2020; Cao et al., 2021; Tao et al.,...
The infectivity of SARS-CoV-2 variants has been rapidly changing, so understanding the infectivity of new variants is important for effective responses to the pandemic (Singh et al., 2021). These changes have resulted in higher infectivities of new SARS-CoV-2 variants compared with past variants (Allen et al., 2022; Araf et al., 2022; Chen and Wei, 2022; Davies et al., 2021; Faria et al., 2021; Radvak et al., 2021). These alterations in infectivity indicate that the infectivity of new coronavirus variants may be estimated by utilizing the evolutionary distance of the spike protein gene (S gene) between the wild type and the variants. We previously described a mathematical model in which we used docking simulation results to predict UDP-glucuronosyltransferase 1A1 conjugation capacity (Takaoka et al., 2019). By using the molecular docking simulation analyses, we found a plant leaf extract that inhibited the binding of SARS-CoV-2 spike protein to angiotensin-converting enzyme 2 (ACE2) (Hagiyama et al., 2022). Similarly, in silico docking data for a variant of SARS-CoV-2 spike protein and ACE2 may be used to estimate the infectivity of the new variant. In this research here, we choose these two approaches with mathematical models to achieve rapid and better understanding of the infectivity of new SARS-CoV-2 variants.

2. Materials and methods

2.1. Determination of the evolutionary distance between wild-type and variant s genes and infectivities of the variants

We chose six variants of SARS-CoV-2—alpha (Claro et al., 2021), beta (Tegally et al., 2021), gamma (Loconsole et al., 2021), delta (Cherian et al., 2021), omicron BA.1 (Dejnirattisai et al., 2021), and omicron BA.2 (Christensen et al., 2022)—for this research. Table 1 shows the amino acid substitution of spike protein variants. We used the multiple sequence alignment method for the S gene of the SARS-CoV-2 strain Mutations in spike protein sequences and the template structure, we used Molecular Operating Environment software (Chemical Computing Group, Montreal, Quebec, Canada) to perform the homology modeling. The model structures were then subjected to structural optimization with molecular dynamics (MD) simulation, as described in our previous report (Takaoka et al., 2021). After 10,000-ps MD simulation of each spike protein, docking analysis with ACE2 was performed as follows: The ACE2 structure was obtained from the PDB (PDB ID: 6M0J), an X-ray crystal RBD and ACE2 complex with a high quality among the reported ones (other PDB ID: 6KV1, 6LZG). To reduce the search space for docking, a partial structure of the RBD in the “up” conformation was cut from our simulated structure of the SARS-CoV-2 spike protein trimer. Docking sites were defined on the basis of the crystal structure of SARS-CoV-2 spike protein RBD bound with ACE2 (PDB ID: 6M0J). Two thousand docking runs were performed for each variant by using ZDOCK software (University of Massachusetts Medical School, Worcester, MA) with the following parameters as the docking site: amino acid no 401, 430, 433, 437, 439, 440, 459, 470, 471, 473, 477, 480, 482, 484–486, 489 for RBD and amino acid no 24, 27, 28, 30, 31, 34, 35, 37, 38, 41, 42, 79, 82, 83, 330–355, 357, 393 for ACE2, respectively. To determine the stable complex among the docking runs, the ZDOCK scores of the resulting complexes were clustered by using the group average clustering algorithm. The upper tail rule Mojena (1977) was applied to determine the number of clusters. We used the most stable ZDOCK score in the docking runs and the average ZDOCK score of the most stable complex group except the one with the most stable score as the binding affinity to derive a mathematical model according to our previous research (Takaoka et al., 2019).

To verify our mathematical model, we used leave-one-out cross-validation for infectivity of SARS-CoV-2 as follows: We removed one of the SARS-CoV-2 strains, determined the constants of the mathematical model by using the rest of the five strains, and then performed calculations to predict the infectivity for the excluded strain. We then confirmed the correlation between the reported infectivity and the predicted infectivity.

Table 1

| SARS-CoV-2 strain | Mutations in spike protein |
|-------------------|---------------------------|
| Alpha (B.1.1.7)   | H69V, V706del, Y144del, N501Y, A570D, D614G, P664H, T716I, S982A, D1181L |
| Beta (B.1.351)    | D610A, D215G, L214A, A243del, K417N, E484K, N501Y, D614G, A701V |
| Gamma (P.1)       | L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F |
| Delta (B.1.617.2) | T19R, E156-F157del, R158G, L452R, T478K, T478K, D614G, H655Y, T1027I, V1176F |
| Omicron BA.1      | A67V, H69-V706del, T95I, G142D, V143-I145del, N211del, L212L, I214F, G393D, S371L, S373P, S375F, K417N, N440K, G465S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, G506S, D614G, H655Y, N679K, P681H, N684K, D766Y, N966K, L981F |
| Omicron BA.2      | T19L, L24P, S26del, A27N, G142D, V143l, G393D, S371F, S373P, S375F, T376A, D405N, R408K, K417N, N440K, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N684K, D766Y, N966K |

Bold, amino acids in RBD. The information of amino acid substitutions are obtained from the following sources: Alpha, Beta and Gamma, https://covdb.stanford.edu/variants/; Delta, https://covariants.org/variants/21ADelta; Omicron BA.1, https://www.cdc.gov/coronavirus/2019-ncov/science/science-brief/omicron-variant.html; Omicron BA.2, Perumal et al. (J Med Virol, DOI: 10.1002/jmv.27661, 2022).

2.2. Analysis of the three-dimensional structures of SARS-CoV-2 spike protein, analysis of docking of the receptor-binding domain (RBD) and ACE2 protein, and development of a mathematical model

The amino acid sequence of wild-type SARS-CoV-2 spike protein was obtained from UniProt (UniProt ID: P0DTC2) (UniProt Consortium, 2018). The three-dimensional structure of SARS-CoV-2 spike protein (Protein Data Bank [PDB] ID: 6ZGG) (Berman et al., 2000), which is an open state trimer, was used as the template for homology modeling of wild-type and variant SARS-CoV-2 spike proteins. With the amino acid sequences and the template structure, we used Molecular Operating Environment software (Chemical Computing Group, Montreal, Quebec, Canada) to perform the homology modeling. The model structures were then subjected to structural optimization with molecular dynamics (MD) simulation, as described in our previous report (Takaoka et al., 2021). After 10,000-ps MD simulation of each spike protein, docking analysis with ACE2 was performed as follows: The ACE2 structure was obtained from the PDB (PDB ID: 6M0J), an X-ray crystal RBD and ACE2 complex with a high quality among the reported ones (other PDB ID: 6KV1, 6LZG). To reduce the search space for docking, a partial structure of the RBD in the “up” conformation was cut from our simulated structure of the SARS-CoV-2 spike protein trimer. Docking sites were defined on the basis of the crystal structure of SARS-CoV-2 spike protein RBD bound with ACE2 (PDB ID: 6M0J). Two thousand docking runs were performed for each variant by using ZDOCK software (University of Massachusetts Medical School, Worcester, MA) with the following parameters as the docking site: amino acid no 401, 430, 433, 437, 439, 440, 459, 470, 471, 473, 477, 480, 482, 484–486, 489 for RBD and amino acid no 24, 27, 28, 30, 31, 34, 35, 37, 38, 41, 42, 79, 82, 83, 330–355, 357, 393 for ACE2, respectively. To determine the stable complex among the docking runs, the ZDOCK scores of the resulting complexes were clustered by using the group average clustering algorithm. The upper tail rule Mojena (1977) was applied to determine the number of clusters. We used the most stable ZDOCK score in the docking runs and the average ZDOCK score of the most stable complex group except the one with the most stable score as the binding affinity to derive a mathematical model according to our previous research (Takaoka et al., 2019).

To verify our mathematical model, we used leave-one-out cross-validation for infectivity of SARS-CoV-2 as follows: We removed one of the SARS-CoV-2 strains, determined the constants of the mathematical model by using the rest of the five strains, and then performed calculations to predict the infectivity for the excluded strain. We then confirmed the correlation between the reported infectivity and the predicted infectivity.
results of our cross-validation.

Fig. 1 provides a flow diagram of this research.

3. Results

3.1. Evolutionary distances for S gene variants and results of docking of RBD with ACE2 protein

Table 2 shows the evolutionary distances between the wild type and the variants, as well as the binding affinities of RBD with ACE2. After multiple alignments of wild-type and variant S genes by using the ClustalW program, FastTree was used to determine the phylogenetic tree (data not shown) and evolutionary distances. The binding affinity of the SARS-CoV-2 spike protein RBD and ACE2 was determined by using the ZDOCK score as an indicator. ZDOCK scores were utilized in cluster analysis to determine the average ZDOCK score of the most stable cluster and ZDOCK score of the most stable complex for each variant. We used these results to develop a mathematical model to predict infectivities of SARS-CoV-2 variants.

3.2. Development of a mathematical model to estimate infectivities of SARS-CoV-2 wild type and variants

We developed a mathematical model to predict the infectivities of SARS-CoV-2 wild type and variants according to our previous research (Takaoka et al., 2019). According to Fig. 2, the relationship between the evolutionary distance and infectivity of each variant can be represented by the following equation:

\[ F = px + q \]

where \( x \) is the evolutionary distance between the wild-type (wild) and a mutant SARS-CoV-2 spike protein. Constant values \( p \) and \( q \) were estimated by minimizing the sum of squared error between the calculated infectivity and the reported infectivity (ratio per wild type) (Allen et al., 2022; Araf et al., 2022; Chen and Wei, 2022; Davies et al., 2021; Faria et al., 2021; Radvak et al., 2021). We then derived the mathematical model for the binding affinity of RBD and ACE2 by using the results of docking analyses. The binding affinity for the mathematical model was derived from the following equation:

\[ B \frac{\log(A)}{A_{\text{wild}}} + t \]

where \( A \) is the average ZDOCK score of the most stable cluster and \( B \) is the ZDOCK score of the most stable complex. The relationship between the binding affinity and the reported infectivity (Fig. 3) can be represented by the following equation:

\[ G = s \times B \frac{\log(A)}{A_{\text{wild}}} + t \]

Constant values \( s \) and \( t \) were estimated by minimizing the sum of squared error between the calculated infectivity and the reported infectivity (ratio per wild type) (Allen et al., 2022; Araf et al., 2022; Chen and Wei, 2022; Davies et al., 2021; Faria et al., 2021; Radvak et al., 2021). The infectivity of SARS-CoV-2 (\( P \)) is defined by the evolutionary distance and the docking result with ACE2:

\[ P = uF + vG + w = u(px + q) + v \left( s \times B \frac{\log(A)}{A_{\text{wild}}} + t \right) + w \]

Constant values \( u, v, \) and \( w \) were estimated by minimizing the sum of squared error between the calculated infectivity \( P \) and the reported infectivity of each SARS-CoV-2 variant \( V \):

\[ \langle u, v, w \rangle = \text{argmin} \left\{ \sum_{M} (P - V)^2 \right\} \]

where set \( M \) represents the SARS-CoV-2 variants whose infectivities are reported. These results suggest that our mathematical model can predict the infectivities of SARS-CoV-2 wild type and variants. Finally, the infectivity of SARS-CoV-2 variants is defined by the following equation:

\[ P = 0.3074869 \times B \frac{\log(A)}{A_{\text{wild}}} + 1.18598 \times x + 1.3879845 \]

In addition, the result of the cross-validation of our mathematical model was indicated by dotted line as an equation of "\( y = 201.06x + 0.5784 \)."

![Figure 2](image-url) - Correlation between evolutionary distance and infectivity of SARS-CoV-2 variants. Collinearity between the evolutionary distance and infectivity was noted. The regression line was indicated by dotted line as an equation of "\( y = 1.9455\ln(x) + 1.20299 \)."

![Figure 3](image-url) - Correlation between the biding affinity based on our docking simulation analyses and infectivity of SARS-CoV-2 variants. The binding affinity for our mathematical model which is based on the molecular simulation data can be used to predict the infectivity of SARS-CoV-2 variants. The regression curve was indicated by dotted line as an equation of "\( y = 1.9455\ln(x) + 1.20299 \)."
model showed that the predicted infectivity of each variant, which was calculated by our model using the data of the other five variants was strongly correlated with the reported infectivity (Fig. 4).

4. Discussion

Reported infectivities of SARS-CoV-2 variants were correlated with both evolutionary distance of the S gene and the binding affinity for the mathematical model according to our in silico docking data (Figs. 2 and 3). In addition, our mathematical model including these two factors reproduced the reported infectivity (Fig. 4).

In previous research, only the RBD domain of the spike protein was used in the analysis of binding affinity with ACE2: no MD analysis of the structure optimization (Celik et al., 2022; Gan et al., 2021; Shah and Woo, 2021; Sithiyothy and Chunstriviroj, 2021; Tragni et al., 2022) and/or local optimization of only the RBD domain of the spike protein (Celik et al., 2022; Gan et al., 2021; Kumar et al., 2022; Shah and Woo, 2021; Sithiyothy and Chunstriviroj, 2021; Tragni et al., 2022; Xue et al., 2021). However, we chose the three-dimensional structure of the complete trimeric spike protein (PDB ID: 6ZGG) in homology modeling in this research. After MD analysis, we confirmed the trajectory of the root mean square deviation and the Ramachandran plot for proper quality of each variant spike protein structure. The RBD domain of the variant trimeric spike protein structures was then provided for the docking analyses with ACE2. Therefore, our experimental strategy achieved high accuracy because we used a molecular simulation method which possibly imitates the behavior of the biomolecules. In addition, our previous research suggested the value of this prediction that utilized our molecular simulation methods (Nakamura et al., 2017; Ohta et al., 2018; Qiang et al., 2017; Sakaeda et al., 2018; Sugawara et al., 2019; Yamamura et al., 2020).

5. Conclusion

We developed a mathematical model for prediction of SARS-CoV-2 variant infectivity. Additional analyses of binding affinities of other spike proteins of SARS-CoV-2 variants with ACE2 may provide greater accuracy, and such experiments are now in progress.

Acknowledgments

Not applicable.

References

Adil, M.T., Rahman, R., Whitehead, D., Jain, V., Al-Taan, O., Rashid, F., et al., 2021. SARS-CoV-2 and the pandemic of COVID-19. Postgrad. Med. J. 97 (1144), 110–116.
Allen, H., Vutrikala, A., Flannagan, J., Twohig, K.A., Zaidi, A., Chudasama, D., et al., 2022. Household transmission of COVID-19 cases associated with SARS-CoV-2 delta variant (B.1.617.2): national case-control study. Lancet Respir. Med. 12, 100252.
Araf, Y., Akter, F., Tang, Y.D., Fatemi, R., Parvez, M.S.A., Zheng, C., et al., 2022. Omicron variant of SARS-CoV-2: genomics, transmissibility, and responses to current COVID-19 vaccines. J. Med. Virol. 94 (5), 1825–1832.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al. 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242. Available from: https://doi.org/10.1002/10.1136/postgradmedj-2020-138386.
Cao, W., Dong, C., Kim, S., Hou, D., Tai, W., Du, L., et al., 2021. Biomechanical characterization of SARS-CoV-2 spike RBD and human ACE2 protein-protein interaction. Biophys. J. 120 (6), 1011–1019. https://doi.org/10.1016/j.bpj.2021.02.007.
Celik, I., Khan, A., Devivany, F.M., Fatimawali, Wei DQ, Tallei, T.E, 2022. Computational characterization of SARS-CoV-2 spike RBD and human ACE2 protein-protein interaction. Biophys. J. 120 (6), 1011–1019. https://doi.org/10.1016/j.bpj.2021.02.007.
Che, T., Wei, C., Yau, L., 2022. Omicron BA.2 (B.1.1.529.2): high potential to becoming the next dominating variant. J. Phys. Chem. Lett. 13, 3840–3849. https://doi.org/10.1021/acs.jpclett.2c03649.
Cherian, S., Potdar, V., Jadhav, S., Yadav, P., Gupta, N., Das, M., et al., 2021. SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. Microorganisms 9 (7). https://doi.org/10.3390/microorganisms90701452.
Christensen, P.A., Olsen, R.J., Long, S.W., Snelak, R., Davis, J.J., Ojeda Saavedra, M., et al., 2022. Signals of significantly increased vaccine breakthrough, decreased hospitalization rates, and less severe disease in patients with coronavirus disease 2019 caused by the omicron variant of severe acute respiratory syndrome...
