Prediction and Evolution of B Cell Epitopes of Surface Protein in SARS-CoV-2

Jerome R Lon[1], Yunmeng Bai[1], Bingxu Zhong[1], Fuqaing Cai[1], Hongli Du[1]*

[1] *School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China*

*Correspondence: hldu@scut.edu.cn; Tel.: +86-020-3938-0667

Running title: *Lon R et al/ Prediction and Evolution of B Cell Epitopes of Surface Protein in SARS-CoV-2*

Abstract

The discovery of epitopes is helpful to the development of SARS-CoV-2 vaccine. The sequences of the surface protein of SARS-CoV-2 and its proximal sequences were obtained by BLAST, the sequences of the whole genome of SARS-CoV-2 were obtained from the GenBank. Based on the NCBI Reference Sequence: NC_045512.2, the conformational and linear B cell epitopes of the surface protein were predicted separately by various prediction methods. Furthermore, the conservation of the epitopes, the adaptability and other evolutionary characteristics were also analyzed. 7 epitopes were predicted, including 5 linear epitopes and 2 conformational epitopes, one of the linear and one of the conformational were coincide. The epitope D mutated easily, but the other epitopes were very conservative and the epitope C was the most conservative. It is worth mentioning that all of the 6 dominated epitopes were absolutely conservative in nearly 1000 SARS-CoV-2 genomes, and they deserved further study. The findings would facilitate the vaccine development, had the potential to be directly applied on the treatment in this disease, but also have the potential to prevent the possible threats caused by other types of coronavirus.

Keywords SARS-CoV-2; Epitopes; Bioinformatics; Evolution

Introduction

In late December 2019, a novel coronavirus was officially named as SARS-CoV-2 by World Health Organization(WHO) and identified as the pathogen causing outbreaks of SARS-like and MERS-like illness in Chinese city of Wuhan, which was a zoonotic disease. As of March 13, 2020, the outbreak of SARS-CoV-2 has been reported in many areas of the world, with more than 130,000 people infected[1]. With an alarmingly human-to-human transmissibility, the reproductive number of SARS-CoV-2 has been computed to around 3.28 [2]. According to the data in NGDC(National Genomics Data Center), at 15:00(GMT+8) on March 13, 2020, 482 genomic variations of SARS-CoV-2 has been reported, which has aroused widespread concern.

The B cell epitope of viral surface protein can specifically bind to the host's B cell antigen receptor and induce the body to produce protective antibody and humoral immune response. The discovery of epitopes is helpful to the development of SARS-CoV-2 vaccine and the understanding of SARS-CoV-2's pathogenesis[3]. 3 proteins embedded in the virus envelope of SARS-CoV-2 have been identified, Spike(S) protein, Envelope(E) protein, Membrane(M) protein. At present, due to the lack of study of the crystal structure of surface protein of SARS-CoV-2, the study of epitopes is time-consuming, power-consuming, cost and difficult [4].

In this work, we analyzed the surface protein of SARS-CoV-2, predicted the structures with bioinformatics methods. On the basis, we predicted the linear and conformational B cell epitopes, analyzed the conservation of the epitopes, the adaptability and other evolutionary characteristics of the surface protein, which provided a theoretical basis for the vaccine development and prevention of SARS-CoV-2.

Results

Basic analysis of surface protein of SARS-CoV-2

The primary structure and physicochemical properties of the S/E/M protein were analyzed. The results revealed that the S protein has an average hydrophilic index of -

0.079**(Figure S1A)**. On the basis of hydrophilicity, it also showed amphoteric properties. There was an outside-in transmembrane helix in 23 residues from position 1214th to position 1236th at the N-terminal(Figure S2A). The protein instability index was 33.01, which revealed the S protein was stable. The E protein has an average hydrophilic index of 1.128(Figure S1B). It was hydrophobic. An inside-out transmembrane helix in 23 residues from position 12th to position 34th at the N-terminal was predicted(Figure S2B). The protein instability index was 38.68, which revealed the E protein was stable. The M protein has an average hydrophilic index of 0.446(Figure S1C). On the basis of hydrophobicity, it also showed amphoteric properties. There were two outside-in transmembrane helices, one was in 20 residues from position 20th to position 39th, the another one was in 23 residues from position 78th to position 100th, and an inside-out transmembrane helix in 20 residues from position 51st to position 73rd, at the N-terminal(Figure S2C). The protein instability index was 39.14, which revealed the M protein was stable.

Prediction of the 3D structure of surface protein of SARS-CoV-2

The optimal template for homology modeling of the S protein of SARS-CoV-2 was the S protein of SARS(PDB ID: 6acc.1), with the sequence identity of 76.47% and the GMQE score of 0.73. According to the evaluation of the structure by Ramachandran plot(**Figure 1A**), 99.3% of the residues were located in the most favoured regions and the allowed regions, 0.7% of the residues were located in the disallowed regions, the high-energy regions(**Table 1**), which was possibly due to some energy was spent in the protein processing to make these residues enter the high-energy regions [5]. The result generally showed that the structure was reliable. The structure of S protein(Figure 1B) of SARS-CoV-2 is a trimer, which can be divided into a tightly curled tail and a distributed head. The head is mainly composed of β-sheet, irregular curl and turn, which is exposed to the envelope of the virus, contributes to the formation of epitopes. The tail is mainly composed of several α-helices, part of which is embedded in the envelope, hinders the formation of epitopes.

The optimal template for homology modeling of the E protein of SARS-CoV-2 was the E protein of SARS(PDB ID: 5x29.1), with the sequence identity of 91.38% and

the GMQE score of 0.73. According to the evaluation of the structure by Ramachandran plot(Figure 1C), 100% of the residues were located in the most favoured regions(Table 1), indicating that the structure was reliable. The E protein of SARS-CoV-2 is a pentamer(Figure 1D), which can be divided into the concentrated transmembrane part and the head located outside the envelope. The head is mainly composed of α-helix, irregular curl and turn, which is exposed to the envelope, contributes to the formation of epitopes. The tail is mainly composed of long α-helix, most of which are embedded in the envelope, hinders the formation of epitopes.

The optimal template for homology modeling of the M protein of SARS-CoV-2 was the effector protein Zt-KP6-1(PDB ID: 6qpk. 1. A), with the sequence identity of 20.00% and the GMQE score of 0.06. The sequence identity between the optimal template and the M protein of SARS-CoV-2 and the GMQE score are too low, so that the template is not suitable for homology modeling.

Prediction of linear B cell epitopes

All linear B cell epitopes of the surface protein were filtered according to the following criteria: (1) region with high surface probability(≥0.75), strong antigenicity(≥0) and high flexibility; (2) excluding the region with α-helix, β-sheet and glycosylation site(Figure 2); (3) in line with the prediction by BepiPred 2.0(cut off to 0.35)**(Table S1)** and ABCpred(cut off to 0.51)(Table S2). Based on the results obtained with these methods and artificial optimization, 4 potential linear B cell epitopes of the S protein were predicted(Table 2, Figure 3A), including 601-605 aa, 656-660 aa, 676-682 aa, 808-813 aa, and they were named as the epitope A, B, C, D, respectively; 1 epitope of the E protein was selected(60-65 aa) and named as the epitope F(Table 2, Figure 3C); 1 epitope of the M protein was selected (211-215 aa) and named as the epitope H(Table 2).

Prediction of conformational B cell epitopes

The conformational B cell epitopes of surface protein were predicted with Ellipro(Table S3) and SEPPA 3.0(Table S4) with the threshold of 0.063 and 0.5, respectively. After the artificial optimization, one conformational B-cell epitope (403-405,416,445,446,455,500 aa) of S protein was predicted(Table 2). It is obvious that the

region located on the head of the S protein(Figure 3B), which is the outside of SARS-CoV-2, making it easy to form an epitope. We selected it as a dominant conformational epitope and named it as the epitope E. Additionally, one conformational B-cell epitope(60-65 aa) of E protein was predicted(Table 2), which is consistent with the epitope F of the E protein. Similarly, this region located on the outside(Figure 3C), we selected it as a dominant conformational epitope and named G. However, the conformational epitope of the M protein could not be predicted due to the failure of credible homology modeling.

Analysis of epitope conservation

The Consurf Server was used to predict epitope conservative sites with the structure of surface proteins and the alignment results in different datasets (Table S5). All the epitopes of the S, E, M protein were absolute conservative among all SARS-CoV-2 sequences(Table 3A, Figure S3A-G). To further calculate the conservation of the epitopes in different coronavirus datasets, the representative sequences from SARS-CoV-2 were selected to participate in the human coronavirus dataset and the coronavirus dataset, due to amino acid sequences of some S or E or M protein were absolute conservative in SARS-CoV-2. The conservation was a little lower in human coronavirus than those of in SARS-CoV-2(Table 3B, Figure S4A-G), and the epitope D was easy to mutate. The other epitopes were conservative and the epitope F/G was the most conservative one. As for the coronavirus(Table 3C, Figure S5A-G), in the 5 epitopes of S protein, 4 of which obtained the conservative score less than 1, ranging from -0.854 to 0.256. It showed that the epitope C with the minimum score is the most stable and not easy to mutate. Besides, the score of the epitope D was 1.247, showing the relatively high possibility to mutate. The epitopes of the E protein were stable with the conservative score less than 1. The site conservation of the M protein could not be predicted due to the failure of credible homology modeling.

Discussion

SARS-CoV-2 caused huge impact to human production, living and even life, has

become a major challenge confronting the whole world. Development of vaccine is one of the effective means of prevention and treatment of the virus long-term. Epitope vaccine is the trend of development of vaccine due to the advantages of strong pertinence, less toxic and side effects and easy to transportation and storage [6]. The determination of epitopes is the basis of the development and application of vaccine, and the clinical diagnosis and treatment. Currently, the methods which were mainly used are X- crystal diffraction method, immune experiment method and bioinformatics method. The first two are time-consuming and laborious, the bioinformatics method is gaining more and more credibility among researchers [3,6,7]. There are many factors to be considered in the prediction of epitopes by bioinformatics method, such as the surface probablity and flexibility of the epitopes. At the same time, it is necessary to exclude the structurally stable and non-deformable α-helix, β-sheet, glycosylation sites which may obscure the epitopes or alter the antigenicity, etc [8]. Even so, the predicted epitopes are still inaccurate [4]. Compared with the current study on SARS-CoV-2, this work adopted various prediction methods and 3D structure databases developed in recent years, which were based on artificial neural network, Hidden Markov Model(HMM), Support Vector Machine(SVM), etc, such as ABCpred, BepiPred2.0, SEPPA 3.0, IEDB, etc. Compared with prediction by a single method [9] or on the basis of epitopes of SARS [10], these methods and databases greatly improved the accuracy of prediction and had more bioinformatic meaning. We comprehensively analyzed the prediction results from the tools which were widely used, set up screening criteria on the basis of primary structure, secondary structure and tertiary structure, so that the prediction results would more accurate and reliable.

The S protein, the E protein and the M protein are surface proteins of SARS-CoV-2, which have the potential as antigenic molecules. However, the current study on the epitopes prediction of SARS-CoV-2 [11], due to the S protein has been reported to be the directly binding molecule of SARS-CoV-2 to ACE2[12], the prediction of epitopes is mainly focusing on the S protein, with few studies on the E protein and the M protein. In this work, we analyzed the S protein, the E protein and the M protein, predicted their epitopes. On the basis, 7 B cell epitopes were predicted, including 2 conformational

and 6 linear B cell epitopes, one of the conformational and one of the linear are coincide. All of the epitope A, B, C, D located on the surface of the tail of the S protein, which is relatively easy to bind. The epitope E is located on the head of the S protein, which is the key area where the S protein recognizes and binds to ACE2 [12,13], has the potential to block the infection process. The epitope F and the epitope G located on the end of the head of the E protein, the two epitopes coincide, this may due to they are all consecutive and the secondary structure avoided the α-helix and the β-sheet. The epitope H is derived from the M protein, the structure and conservation could not be determined due to the inability to predict reliable structure. However, it could be known from the surface probablity scores that the epitope H is more likely to be located on the surface of the M protein.

The higher the conservation score calculated by the Consurf Server is, the more likely the site is to be mutated in the evolutionary process. When the score<1, the site is likely to be a conservative site; when the score is between 1 and 2, the site is a site which is likely to be a relatively easy mutation; when the score>2, the site is likely to be an easy mutation site [14]. In the 7 epitopes obtained, all the epitopes of the S, E, M protein were absolute conservative among all SARS-CoV-2 sequences. For the human coronavirus dataset and the coronavirus dataset, only the average conservative score of the epitope D is higher than 1, which is prone to mutation. The epitope D should not be used as an epitope of the S protein. The conservation of the epitope H could not be calculated by the PDB file, the application value of the epitope H needed further experimental verification. Although the epitopes could be integrally considered to be conservative, the independent residues of these epitopes could still easy to mutate. Except the epitope E, all of 6 dominate epitopes contain 1-2 residues which has a conservative score higher than 1(Table 3C), indicating that these residues were likely to be easy mutation sites. These residues mostly located at the head or the tail of the epitopes, therefore, the mutation of these residues should be paid attention to, and the length of the epitopes should be adjusted according to the actual effect in application. The scores of epitopes in different datasets were different, which could due to the quantity of sequences in the datasets and the structures were analyzed in different

situations.

In this work, we predicted 6 reliable epitopes: A, B, C, E, F/G and H. The reliability of the epitopes of the S protein was relatively better than that of the epitopes of the E protein and the M protein, indicating that the S protein is still the optimal choice for the prediction of epitopes and the development of vaccine. All of the 6 epitopes were able to achieve absolute conservation in SARS-CoV-2, and to achieve relative conservation in the data set, including SARS, etc. Therefore, the epitopes not only have the potential to be directly applied on the treatment in this disease, but also have the potential to prevent the possible threats caused by other types of coronavirus. In addition, although various factors of prediction were integrated in this work, more experimental data are needed to further verify whether all the 6 epitopes can induce the body to produce corresponding antibodies and generate specific humoral immunity, due to the limited data set and other factors.

Materials and methods

Materials

All of the analysis was based on the NCBI Reference Sequence: NC_045512.2. We obtained the sequence of S, E and M protein and its proximal sequences by BLAST, which got 420, 334 and 329 sequences in total from NCBI database respectively. We obtained the whole genome sequence of SARS-CoV-2 from Genbank and GISAID(959 in total), which were used to be a dataset after genome annotation. The genome sequences, which performed mistakes of translation, were deleted.

Methods

Basic analysis of surface protein of SARS-CoV-2

The physical and chemical properties of target protein were analyzed by the Port-Param tool in ExPASy [15], including the primary structure of the target protein, molecular formula, theoretical isoelectric point, the protein instability index(the index<40 means the protein was stable), *etc*. Online software, ProtScale, was used to deeply analyze the hydrophilicity and hydrophobicity of target protein and the

distribution of hydrophilicity and hydrophobicity of polypeptide chains [15]. SARS-CoV-2 carried the S/E/M protein through the virus envelope, the transmembrane region of the protein was predicted online by TMHMM 2.0 [16].

Prediction of the 3D structure of target protein

With the amino acid sequences of the surface protein of SARS-CoV-2 of NC_045512.2 as templates, based on homology modeling method, we predicted the 3D structure through the online server SWISS-MODEL[17], selected and optimized the optimal structure based on the template identity and GMQE value[17], the rationality of the structure was evaluated by Ramachandran plot [18] with PDBsum server. The structures were displayed and analyzed by SWISS-pdb Viewer v4.10 [19].

Prediction of conformational B cell epitopes of target protein of SARS-CoV-2

Based on the structures, the conformational B cell epitopes were predicted by SEPPA 3.0 [20] and Ellipro [21] respectively, and the common predicted conformational B cell epitopes from two methods were selected for the further analysis.

Prediction of linear B cell epitopes of target protein of SARS-CoV-2

The Protean module of DNAStar was used to predict the flexibility[21], surface probablity [22] and antigenic index [23] of the target protein of SARS-CoV-2. The linear B cell epitope was predicted by ABCpred [24] and BepiPred 2.0 [25] respectively and the common predicted linear B cell epitopes from two methods were selected for the further analysis. Coupled with the secondary structure, the tertiary structure and the glycosylation sites [26] *etc*, the linear B cell epitopes were finally determined.

Analysis of epitope conservation

Based on the PDB model and the multiple alignment result, we used the Consurf Server to analyze the conservation of amino acid sites of the epitopes online[27]. The conservation of epitopes on the surface protein of SARS-CoV-2 was analyzed by multiple alignment with MAFFT and Logo was drawn with Weblogo [28,29].

Authors' contribution

JL conceived the study and participated in its design and coordination. HD

participated in the design of the study and helped draft the manuscript. YB participated in analysis of conservation, sequence alignment and manuscript drafting. BZ participated in antigenic prediction. FC participated in drafting the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

This work was supported by the National Key R&D Program of China (2018YFC0910201), the Key R&D Program of Guangdong Province (2019B020226001), and the Science and the Technology Planning Project of Guangzhou (201704020176). The Student Entrepreneurship and Innovation Center of the school of biology and biological engineering, South China University of Technology, also provided a lot of help during the preparation of the project.

References

[1] Han Q, Lin Q, Jin S, You L. Recent insights into 2019-nCoV: a brief but comprehensive review. J Infect 2020. https://doi.org/10.1016/j.jinf.2020.02.010.

[2] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med 2020. https://doi.org/10.1093/jtm/taaa021.

[3] El-Manzalawy Y, Honavar V. Recent advances in B-cell epitope prediction methods. Immunome Res 2010;6 Suppl 2:S2. https://doi.org/10.1186/1745-7580-6-S2-S2.

[4] Sun P, Ju H, Liu Z, Ning Q, Zhang J, Zhao X et al. Bioinformatics resources and tools for conformational B-cell epitope prediction. Comput Math Methods Med

2013;2013:943636. https://doi.org/10.1155/2013/943636.

[5] Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins 1992;12(4):345–64. https://doi.org/10.1002/prot.340120407.

[6] Groot AS de, Sbai H, Aubin CS, McMurry J, Martin W. Immuno-informatics: Mining genomes for vaccine components. Immunol Cell Biol 2002;80(3):255–69. https://doi.org/10.1046/j.1440-1711.2002.01092.x.

[7] Yang X, Yu X. An introduction to epitope prediction methods and software. Rev Med Virol 2009;19(2):77–96. https://doi.org/10.1002/rmv.602.

[8] Chen W, Zhong Y, Qin Y, Sun S, Li Z. The evolutionary pattern of glycosylation sites in influenza virus (H5N1) hemagglutinin and neuraminidase. PLoS ONE 2012;7(11):e49224. https://doi.org/10.1371/journal.pone.0049224.

[9] Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe 2020. https://doi.org/10.1016/j.chom.2020.03.002.

[10] Yuan M, Wu NC, Zhu X, Lee C-CD, So RTY, Lv H et al. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV; 2020.

[11] Lv H, Wu NC, Tsang OT-Y, Yuan M, Perera RAPM, Leung WS et al. Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections; 2020.

[12] Tian X, Li C, Huang A, Xia S, Lu S, Shi Z et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody; 2020.

[13] Yan R, Zhang Y, Guo Y, Xia L, Zhou Q. Structural basis for the recognition of the 2019-nCoV by human ACE2; 2020.

[14] WANG S, Li G, Di WU, Cao Z. Mutation Feature Analysis on Epitope and Receptor Binding Sites of Influenza A H1N1 Hemagglutinin. ACTA BIOPHYSICA SINICA 2012;28(6):486.

https://doi.org/10.3724/SP.J.1260.2012.20015.

[15] Walker JM. The Proteomics Protocols Handbook. Dordrecht: Springer; 2005.

[16] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999;294(5):1351–62. https://doi.org/10.1006/jmbi.1999.3310.

[17] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018;46(W1):W296-W303. https://doi.org/10.1093/nar/gky427.

[18] Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26(2):283–91. https://doi.org/10.1107/S0021889892009944.

[19] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18(15):2714–23. https://doi.org/10.1002/elps.1150181505.

[20] Zhou C, Chen Z, Zhang L, Yan D, Mao T, Tang K et al. SEPPA 3.0-enhanced spatial epitope prediction enabling glycoprotein antigens. Nucleic Acids Res 2019;47(W1):W388-W394. https://doi.org/10.1093/nar/gkz413.

[21] Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften 1985;72(4):212–3. https://doi.org/10.1007/BF01195768.

[22] Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol 1985;55(3):836–9.

[23] Jameson BA, Wolf H. The antigenic index: a novel algorithm for predicting antigenic determinants. Comput Appl Biosci 1988;4(1):181–6. https://doi.org/10.1093/bioinformatics/4.1.181.

[24] Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 2006;65(1):40–8. https://doi.org/10.1002/prot.21078.

[25] Larsen JEP, Lund O, Nielsen M. Improved method for predicting linear B-cell

epitopes. Immunome Res 2006;2:2. https://doi.org/10.1186/1745-7580-2-2.

[26] Gupta R, Jung E, Brunak S. Prediction of N-glycosylation sites in human proteins 2004;46:203–6.

[27] Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 2016;44(W1):W344-50. https://doi.org/10.1093/nar/gkw408.

[28] Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14(6):1188–90. https://doi.org/10.1101/gr.849004.

[29] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990;18(20):6097–100. https://doi.org/10.1093/nar/18.20.6097.

Figure legends

Figure 1 The 3D structure prediction and Ramachandran plot analysis of the S and E protein

A. The 3D structure of the S protein predicted by homology modeling. It is a trimer, the head contains RBD (receptor Binding Domain) [24], the tail contains the basic elements required for the membrane fusion, the end of the tail is a transmembrane region and is embedded in the envelope of SARS-CoV-2. **B.** The 3D structure of the E protein predicted by homology modeling. It is a pentamer with ion channel activity [25]. Its head is short, the middle of the tail is a transmembrane region which help the E protein embed in the envelope of SARS-CoV-2. **C.** The Ramachandran plot analysis of the 3D structure of the S protein (without Gly and Pro). Most residues located in the red (core) regions, and few in the white regions. **D.** The Ramachandran plot analysis of the 3D structure of the E protein (without Gly and Pro). All of the residues located on the red(core) region.

Figure 2 The secondary structures and properties analysis of the S, E and M protein with the Protean tool of DNAStar

A. Analysis of the S protein. It contains most α-helix and β-sheet, some Turn and Coli region, several discontinuous high flexibility fragments, fluctuant surface probablity with a few of positive peak and several antigenicity regions with positive peak. **B.** Analysis of the E protein. It contains most α-helix and β-sheet, some Turn and Coli region, three high flexibility fragments, few surface probablity regions and two antigenicity regions with positive peak in the begin and the end of polypeptide chain, respectively. **C.** Analysis of the M protein. It contains most α-helix and β-sheet, some Turn and Coli region, several high flexibility fragments, few surface probablity regions, two antigenicity regiona with positive single peak in the begin and middle of peptide chain, respectively, and consecutive positive peaks in the end.

Figure 3 The predicted epitopes of the S and E protein

A. The predicted linear B-cell epitopes of the S protein. The epitope A, B, C located in the forepart of the tail, the epitope D located in the back part of the tail and is close to the transmembrane region. **B.** The predicted conformational B-cell epitope of the S protein. It located in the RBD of the head which is the vital sites binding with ACE2. **C.** The predicted B-cell epitope of the E protein. The epitope G is the linear epitope and the F is the conformational epitope, which are coincide.

Tables

Table 1 The plot statistics of the Ramachandran plot
Table 2 The composition and the antigenic index of the epitopes of SARS-CoV-2
Table 3A The conservation of the epitopes in SARS-CoV-2
Table 3B The conservation of the epitopes in human coronavirus
Table 3C The conservation of the epitopes in coronavirus

Supplementary information

Figure S1 Deep analysis of hydrophilicity and hydrophobicity of surface protein of SARS-CoV-2

The online software, ProtScale, was used to predict the hydrophilicity and hydrophobicity of the surface protein deeply. **A.** The S protein has a maximum score of hydrophobicity, 3.222 at the 7^{th} site, which revealed a strong hydrophobicity; a minimum score of hydrophobicity, -2.589 at the 679^{th} site, which revealed a strong

hydrophilicity. The score of hydrophilicity and hydrophobicity on the polypeptide chain of S protein constantly fluctuates, with most of the scores being negative, which revealed the possibility that the protein had bisexual properties on the basis of hydrophilicity. **B.** The E protein has a maximum score of hydrophobicity, 3.489 at the 21st and the 25th site, which revealed a strong hydrophobicity; a minimum score of hydrophobicity, -1.550 at the 65th site, which revealed a strong hydrophilicity. Most of the scores of the residues being positive, which revealed the possibility that the protein has obvious hydrophobicity. **C.** The M protein has a maximum score of hydrophobicity, 2.978 at the 84th site, which revealed a strong hydrophobicity; a minimum score of hydrophobicity, -1.956 at the 211th and the 212th site, which revealed a strong hydrophilicity. The scores of hydrophilicity and hydrophobicity on the polypeptide chain of M protein showed large fluctuations, and the number of positive scores and negative scores were similar, the positive scores accounted for the majority, which revealed the possibility that the protein had bisexual properties on the basis of hydrophobicity.

Figure S2 The transmembrane region of the surface protein of SARS-CoV-2

The S, E and M protein are embedded in the envelope of SARS-CoV-2, the transmembrane helix was predicted by TMHMM 2.0 server. All of three amino acid indexes were higher than 18, indicating the reliability of the prediction. **A.** For the S protein, an outside-in transmembrane helix was predicted in the 23 residues of amino acids from position 1214th to position 1236th at the N-terminal. The amino acid index was 23.97303. **B.** For the E protein, an inside-out transmembrane helix was predicted in the 23 residues of amino acids from position 12th to position 34th at the N-terminal. The amino acid index was 25.72521. **C.** For the M protein, 2 outside-in transmembrane helices were predicted, which were a helix in the 20 residues of amino acids from position 20th to position 39th and a helix in the 23 residues of amino acids from position 78th to position 100th at the N-terminal. An inside-out helix was predicted in the 23 residues of amino acids from position 51st to position 73rd at the N-terminal. The amino acid index was 64.90522.

Figure S3 The antigenic conservation of the surface protein in SARS-CoV-2

A. The epitope A was absolutely conservative in 756 SARS-CoV-2 genomes. **B.** The epitope B was absolutely conservative in 756 SARS-CoV-2 genomes. **C.** The epitope C was absolutely conservative in 756 SARS-CoV-2 genomes. **D.** The epitope D was absolutely conservative in 756 SARS-CoV-2 genomes. **E.** The epitope E was absolutely conservative in 756 SARS-CoV-2 genomes. **F.** The epitope F/G was absolutely conservative in 939 SARS-CoV-2 genomes. **G.** The epitope H was absolutely conservative in 913 SARS-CoV-2 genomes.

Figure S4 The antigenic conservation of the surface protein in human coronavirus

A. The conservation of the epitope A in 331 human coronavirus genomes. **B.** The conservation of the epitope B in 331 human coronavirus genomes. **C.** The conservation of the epitope C in 331 human coronavirus genomes. **D.** The conservation of the epitope D in 331 human coronavirus genomes. **E.** The conservation of the epitope E in 331 human coronavirus genomes. **F.** The conservation of the epitope F/G in 268 human coronavirus genomes. **G.** The conservation of the epitope H in 268 human coronavirus genomes.

Figure S5 The antigenic conservation of the surface protein in coronavirus

A. The conservation of the epitope A in 403 human coronavirus genomes. **B.** The conservation of the epitope B in 403 human coronavirus genomes. **C.** The conservation of the epitope C in 403 human coronavirus genomes. **D.** The conservation of the epitope D in 403 human coronavirus genomes. **E.** The conservation of the epitope E in 403 human coronavirus genomes. **F.** The conservation of the epitope F/G in 334 human coronavirus genomes. **G.** The conservation of the epitope in 327 human coronavirus genomes.

Table S1 Bepipred2.0 linear epitope prediction results

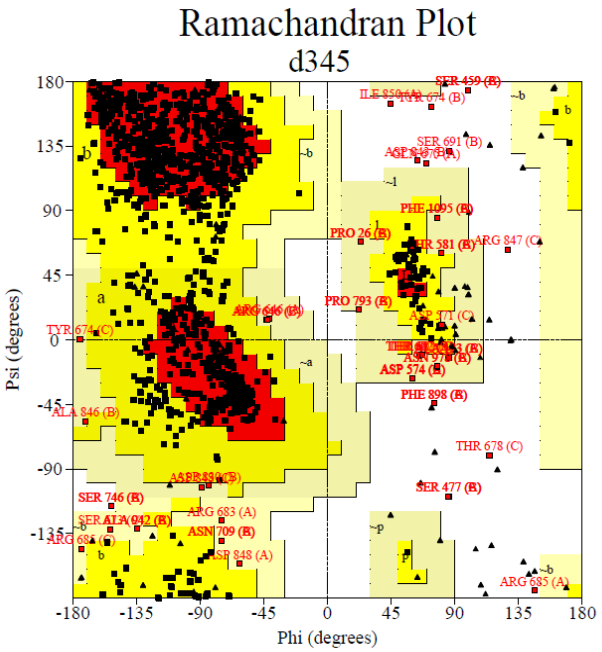
Table S2 ABCpred linear epitope prediction results

Table S3 Prediction results of conformational B cell epitopes of surface protein by Ellipo

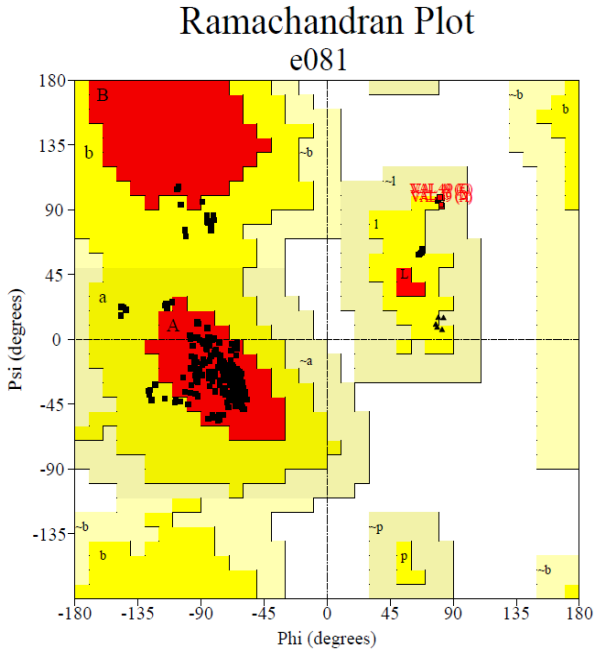
Table S4 Prediction results of conformational B cell epitopes of surface protein by SEPPA3.0

Table S5 Conservation analysis of epitopes in different datasets by Consurf

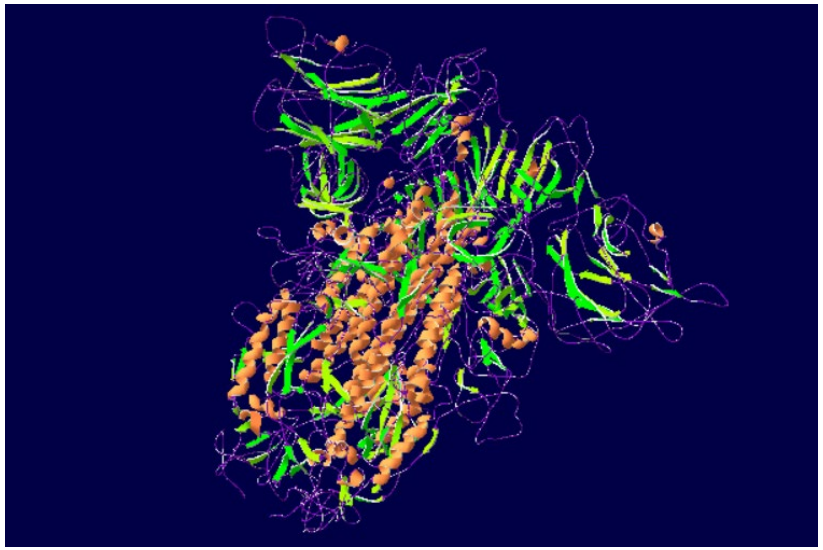
A


Ramachandran Plot
d345
Psi (degrees)
Phi (degrees)
180
135
90
45
0
-45
-90
-135
-180

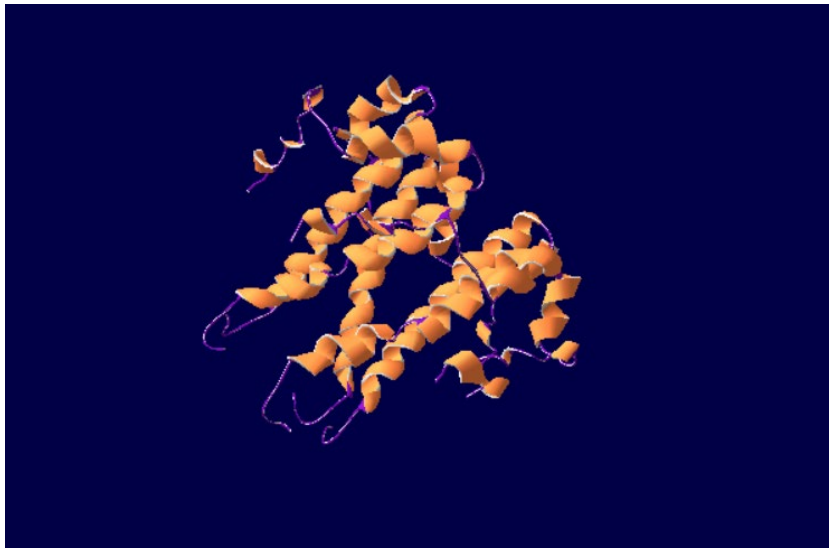
C


Ramachandran Plot
e081
Psi (degrees)
Phi (degrees)
180
135
90
45
0
-45
-90
-135
-180

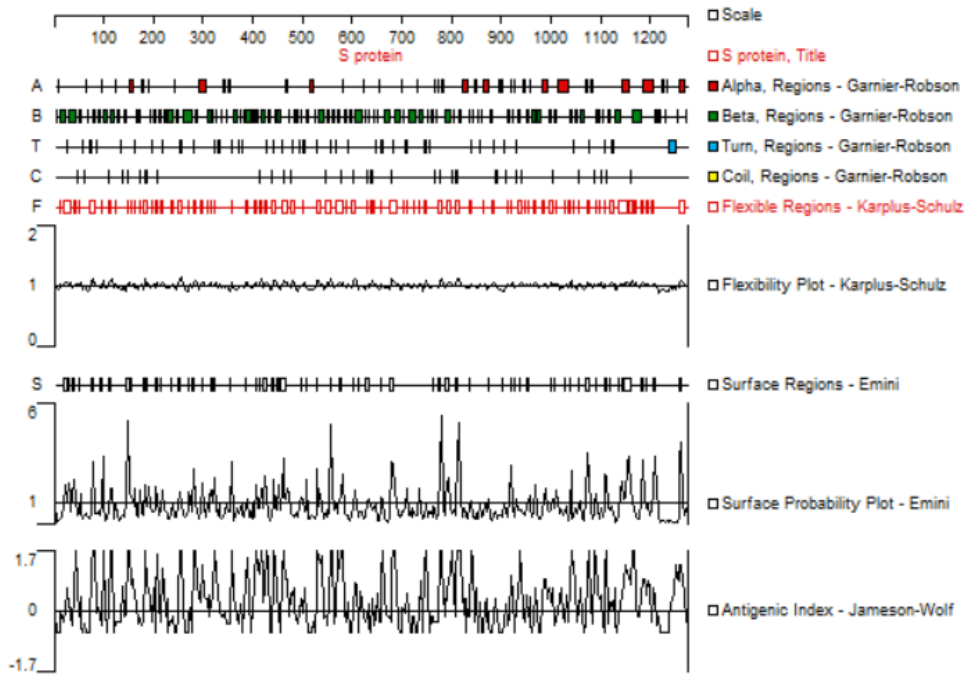
B



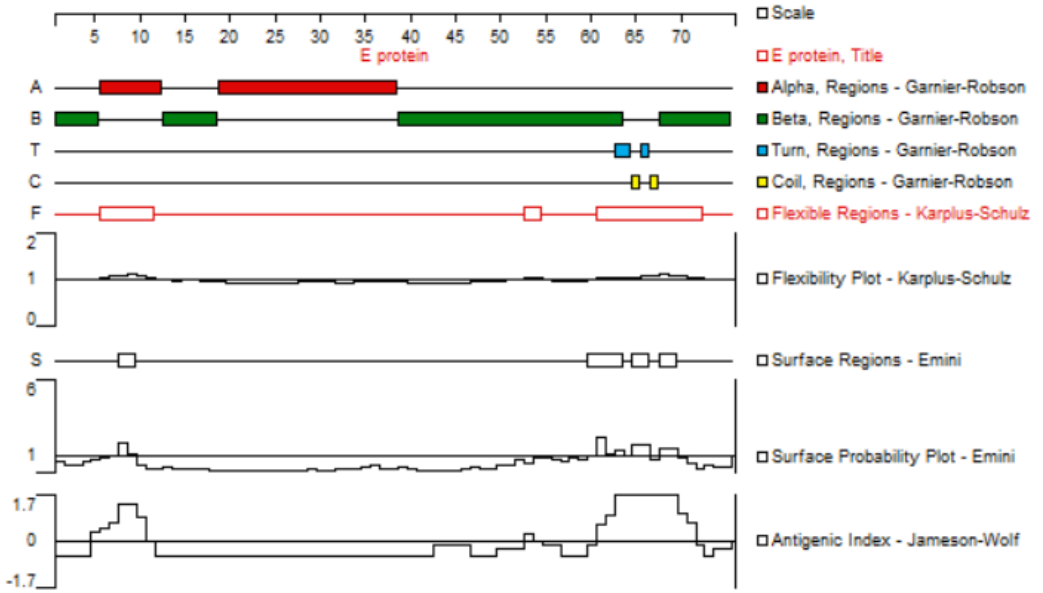
D



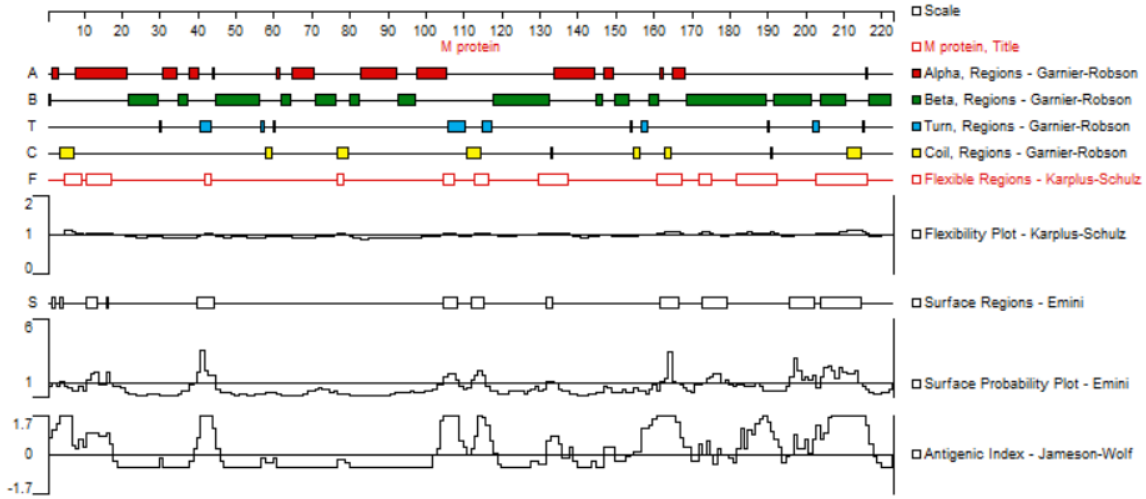
A


100 200 300 400 500 600 700 800 900 1000 1100 1200
S protein
Scale
S protein, Title
A
B
T
C
F
Alpha, Regions - Garnier-Robson
Beta, Regions - Garnier-Robson
Turn, Regions - Garnier-Robson
Coil, Regions - Garnier-Robson
Flexible Regions - Karplus-Schulz
2
1
0
Flexibility Plot - Karplus-Schulz
S
Surface Regions - Emini
6
1
Surface Probability Plot - Emini
1.7
0
-1.7
Antigenic Index - Jameson-Wolf

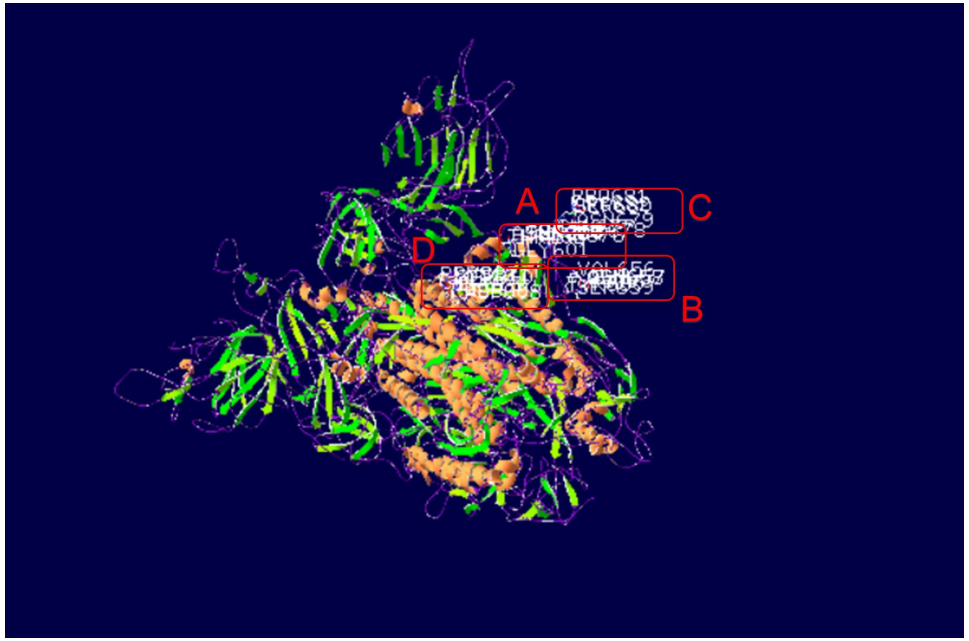
B


5 10 15 20 25 30 35 40 45 50 55 60 65 70
E protein
Scale
E protein, Title
A
B
T
C
F
Alpha, Regions - Garnier-Robson
Beta, Regions - Garnier-Robson
Turn, Regions - Garnier-Robson
Coil, Regions - Garnier-Robson
Flexible Regions - Karplus-Schulz
2
1
0
Flexibility Plot - Karplus-Schulz
S
Surface Regions - Emini
6
1
Surface Probability Plot - Emini
1.7
0
-1.7
Antigenic Index - Jameson-Wolf

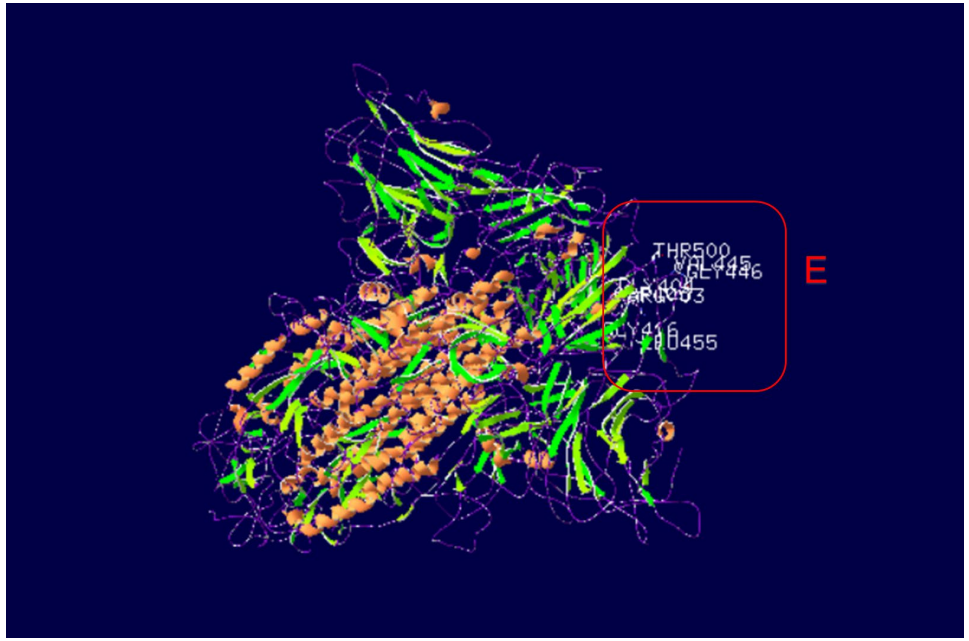
C


10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220
M protein
Scale
M protein, Title
A
B
T
C
F
Alpha, Regions - Garnier-Robson
Beta, Regions - Garnier-Robson
Turn, Regions - Garnier-Robson
Coil, Regions - Garnier-Robson
Flexible Regions - Karplus-Schulz
2
1
0
Flexibility Plot - Karplus-Schulz
S
Surface Regions - Emini
6
1
Surface Probability Plot - Emini
1.7
0
-1.7
Antigenic Index - Jameson-Wolf

A



B



C

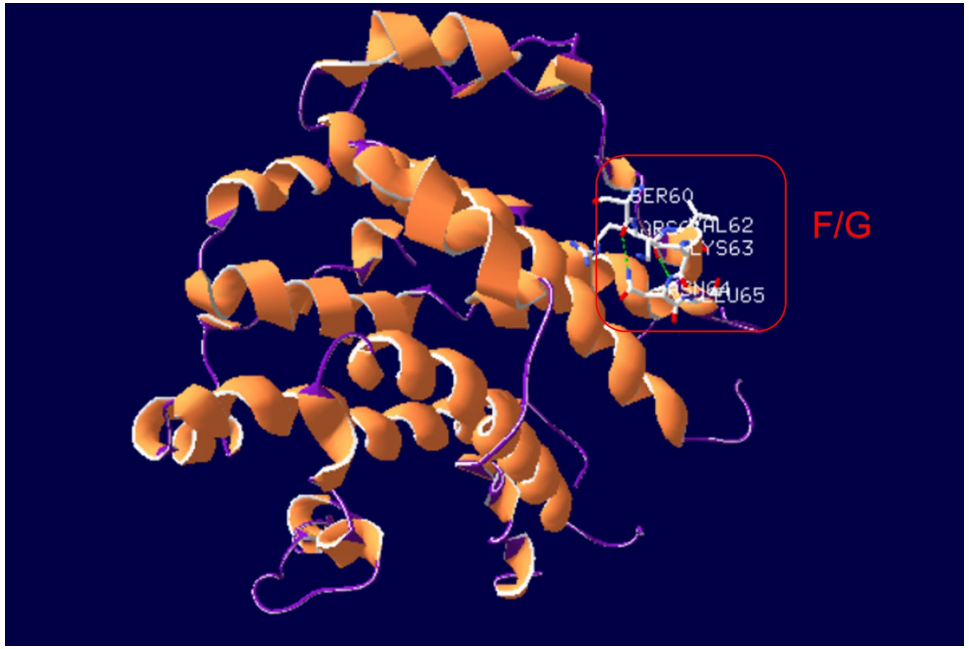


Table 1 The plot statistics of the Ramachandran plot

Plot statistics-S			Plot statistics-E		
Residues in most favoured regions [A, B, L]	2352	78.80%	Residues in most favoured regions [A, B, L]	228	84.40%
Residues in additional allowed regions [a, b, l, p]	577	19.30%	Residues in additional allowed regions [a, b, l, p]	38	14.10%
Residues in generously allowed regions [~a, ~b, ~l, ~p]	36	1.20%	Residues in generously allowed regions [~a, ~b, ~l, ~p]	4	1.50%
Residues in disallowed regions	20	0.70%	Residues in disallowed regions	0	0.00%

Table 2 The composition and the antigenic index of the epitopes of SARS-CoV-2

Name	Position	Amino acid	Antigenic index
A	601-605	GTNTS	0.525
B	656-660	VNNSY	0.575
C	676-682	TQTNSPR	0.675
D	808-813	DPSKPS	0.580
E	**403-405，416，445-446，455，500**	**RGD,G,VG,L,T**	**0.871**
F	59-64	SRVKNL	0.588
G	**59-64**	**SRVKNL**	**0.767**
H	211-215	SSSSD	0.656

Note: The scores of the epitope E and the epitope G were calculated by Ellipro, the others were calculated by Bepipred 2.0.

Table 3A The conservation of the epitopes in SARS-CoV-2

Name	Position	Conservation Score	Average	Name	Position	Conservation Score	Average
A	601	-0.155	-0.273	E	403	-0.237	-0.211
	602	-0.3			404	-0.155	
	603	-0.289			405	-0.241	
	604	-0.3			416	-0.155	
	605	-0.323			445	-0.281	
					446	-0.155	
B	656	-0.281	-0.266		455	-0.16	
	657	-0.289			500	-0.3	
	658	-0.289					
	659	-0.323		F/G	60	-0.841	-0.448
	660	-0.146			61	-0.439	
					62	-0.548	
C	676	-0.3	0.319		63	0.466	
	677	-0.241			64	-0.048	
	678	-0.3			65	-1.277	
	679	-0.289					
	680	-0.323		H	null	null	
	681	-0.173					
	682	3.861					
D	808	-0.241	0.456				
	809	-0.173					
	810	-0.323					
	811	-0.21					
	812	4.005					
	813	-0.323					

Note: The calculation was independent and based on the SARS-CoV-2 data set.

Table 3B The conservation of the epitopes in human coronavirus

Name	Position	Conservation Score	Average	Name	Position	Conservation Score	Average
A	601	0.703	0.457	E	403	0.143	0.427
	602	-0.821			404	0.509	
	603	-0.03			405	0.406	
	604	1.807			416	0.858	
	605	0.639			445	0.807	
					446	0.005	
B	656	0.199	0.724		455	0.706	
	657	1.345			500	-0.015	
	658	1.191					
	659	-0.7		F/G	60	0.096	0.263
	660	1.585			61	0.346	
					62	0.634	
C	676	0.79	0.356		63	-0.438	
	677	0.639			64	0.387	
	678	-0.108			65	0.555	
	679	0.359					
	680	-0.585		H	null	null	
	681	0.956					
	682	0.44					
D	808	1.357	1.323				
	809	1.44					
	810	0.943					
	811	1.688					
	812	1.73					
	813	0.777					

Note: The calculation was independent and based on the human coronavirus data set.

Table 3C The conservation of the epitopes in coronavirus

Name	Position	Conservation Score	Average	Name	Position	Conservation Score	Average
A	601	-0.072	0.256	E	403	-1.464	-0.467
	602	-1.167			404	1.594	
	603	1.492			405	-1.49	
	604	-0.55			416	-0.588	
	605	1.576			445	0.067	
					446	-0.141	
B	656	1.186	-0.618		455	-0.005	
	657	-1.318			500	-1.708	
	658	-0.843					
	659	-1.643		F/G	60	-0.204	0.235
	660	-0.472			61	0.042	
					62	0.246	
C	676	-1.469	-0.854		63	-0.559	
	677	0.027			64	0.775	
	678	-1.493			65	1.108	
	679	-1.408					
	680	-1.714		H	null	null	
	681	-1.007					
	682	1.089					
D	808	1.522	1.247				
	809	1.367					
	810	1.5					
	811	1.554					
	812	1.374					
	813	0.164					

Note: The calculation was independent and based on the coronavirus data set.