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.
Immunoinformatics based prediction of recombinant multi-epitope vaccine for the control and prevention of SARS-CoV-2

Muhammad Asif Rasheed a, Sohail Raza i, Ali Zohaib c, Muhammad Ilyas Riaz b, Amina Amin a, Muhammad Awais h, Sami Ullah Khan d, M. Ijaz Khan e, Yu-Ming Chu f,g,h,*

a Department of Biosciences, COMSATS University Islamabad, Sahiwal Campus, 57000 Sahiwal, Pakistan
b Department of Microbiology, University of Veterinary and Animal Sciences, Lahore 54000, Pakistan
c Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology H-12 Campus, Islamabad, Pakistan
d Department of Mathematics, COMSATS University Islamabad, Sahiwal 57000, Pakistan
e Department of Mathematics, Riphah International University, Faisalabad Campus, Faisalabad 38000, Pakistan
f Department of Mathematics, Huzhou University, Huzhou 313000, PR China
g Hunan Provincial Key Laboratory of Mathematical Modeling and Analysis in Engineering, Changsha University of Science & Technology, Changsha 410114, PR China
h University Institute of Biochemistry and Biotechnology, PMAS-Arid Agriculture University, Rawalpindi 43600, Pakistan
i Institute of Microbiology, University of Veterinary and Animal Sciences, Lahore 54000, Pakistan

Received 9 January 2021; revised 21 January 2021; accepted 26 January 2021
Available online 1 February 2021

KEYWORDS
Coronavirus; Epitopes; SARS-CoV-2; Peptide; Vaccine

Abstract The emergence of SARS-CoV-2 has been reported during December 2019, in the city of Wuhan, China. The transmission of this virus via human to human interaction has already been described. The novel virus has become pandemic and declared as a comprehensive emergency worldwide by World Health Organization due to its exponential spread within and outside China. There is a need of time to create a therapeutic agent and a vaccine to cure and control this lethal SARS-CoV-2. Conventionally, the vaccine development process is time taking, tiresome and requires more economical inputs with manpower. However, bioinformatics offers a key solution to compute the possibilities. The present study focuses on the utilization of bioinformatics platforms to forecast B and T cell epitopes that belong to SARS-CoV-2 spike glycoprotein. The protein is thought to have an involvement in triggering of momentous immune response. NCBI database was explored to collect the surface glycoprotein sequence and was analyzed to determine the immunogenic epitopes. This prediction analysis was carried out using IEDB web based server and the prediction of protein structure was done by homology modeling approach. This study resulted in prediction of 5T cell and 13B cell epitopes. Moreover, GPGPG linker was used to make
these predicted epitopes a single peptide prior to further analysis. Afterwards, a 3D model of the final vaccine peptide was constructed, and the structure quality of the final construct was checked by Ramachandran Plot analysis and ProSA-web. Moreover, docking analysis highlighted three interactions of epitope against HLA-B7 including Lys 178, Gln 303 and Thr 31 residues. In conclusion, the predicted multi epitope peptide can be suggested as therapeutic or prophylactic candidate vaccine against SARS-CoV-2 after further confirmation by immunological assays.

© 2021 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Alexandria University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

With the advent of twentieth century, new viruses start to emerge having global health importance. One of the recent examples is ‘Wuhan coronavirus’. This zoonotic virus has potential to infect humans. This virus was firstly identified in Wuhan, China in persons exposed to the wet market or seafood market [1]. The infection of this virus leads to respiratory damage with signs and symptoms of fever, nasal discharge, cough and sore throat. Fatal cases involved acute respiratory distress, pneumonia, and multi-organ failure. The SARS-CoV-2 virus infection appears to be milder than other human coronaviruses (SARS and MERS) in terms of case fatality, severity, and transmissibility [2]. To date, more than 0.3 million cases with more than 13,000 deaths of SARS-CoV-2 infection have been reported from the globe. Initially, there were limited evidences of virus human to human transfer because no health care personal in the hospitals were infected [3]. After few days, as the cases starts to increase officials admits the horizontal transfer of SARS-CoV-2, spread through nasal secretions of the infected person [4, 5]. In the seafood market of Wuhan, along with other sea animals chicken, bats, snake and other wild animals were also sold. In the beginning, it was predicted that spike protein of the SARS-CoV-2 contribute its transfer from snake to humans [6]. Till now, there are no evidences available of snake infection by this SARS-CoV-2 and serves as host for it [7].

The coronavirus belongs to the family Coronaviridae, based on their genetic properties devided into four genera including Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. The coronaviruses are enveloped, non-segmented, single stranded positive sense RNA genome. The genome size is ranging from 26 kbp to 32 kbp, largest genome in all RNA viruses. The coronaviruses are 80–120 nm in diameter and spherical or pleomorphic in shape [8]. The genome encodes for four structural proteins: the spike protein (S), Membrane protein (M), Nucleocapsid protein (N), and envelop protein (E) [9]. For most of the coronaviruses, all structural proteins are required to produce complete virus particle; however few coronaviruses do not require full structural protein to produce complete infectious virus particle [10]. Each protein not only plays its role in the structure of the virus but also play their part in other aspects of viral replication. The S protein involves in viral attachment and fusion with host cell to facilitate viral entry into the cell [11]. The N protein involved in the nucleocapsid formation and host cellular response during viral infection [12]. The M protein involved in the structure of the virus and has central role in the assembly of the coronavirus [13]. The E protein has an important role in the assembly of the virion [10]. Among these, S protein played an important role in the in the production of immunity against coronaviruses.

The SARS-CoV-2 is the seventh member of the family that infects humans. During genomic analysis this virus formed different clade in sartecovirus subgenus, subfamily orthocoronavirinae. The SARS-CoV-2 is different from other known human coronaviruses. The SARS-CoV-2 was isolated using human airway epithelial cells from the bronchialveolar lavage of the patients admitted in the hospital [14]. The closest linkage of SARS-CoV-2 is with two SARS like coronaviruses that are transfered through bats. Coronaviruses of the other species are genetically different from SARS-CoV-2, reflecting that new coronavirus did not originate from other animal hosts [3].

Presently, limited vaccine or therapeutic agents are available to treat and protect from human coronavirus infections. Development of effective treatment and vaccine is a research priority to control emerging and re-emerging diseases. To this end, the combined use of genetics, immunology, and bioinformatics is called as immuninformatics, a novel approach to find out the effective ways to control the diseases [15]. The conventional approaches for vaccine development not only required more time but also ineffective to control viral diseases spread through RNA viruses. Hence, the immuninformatics based approach has become much popular with the advancement in the genomic and protein sequence databases [16]. Therefore, the main aim of this study was to predict T-cell and B-cell epitopes present in the surface glycoprotein of novel SARS-CoV-2 that could be used to develop promising vaccine against SARS-CoV-2.

2. Material and method

2.1. Retrieval of sequence

To retrieve the protein sequence from NCBI protein database, QHD43416.1 accession number was used. The sequence was related to surface glycoprotein of the isolated coronaviruses.

2.2. The prediction of B cell epitopes and their joining

In context of immunity, B-cell epitopes induce defensive immunity [17]. The antibodies are going to attach with B cell epitopes. In order to predict of the location of linear B cell epitopes, Hidden Markov models (HMMs) were used to predict the locations of B cell epitopes. Moreover, to predict B cell epitopes in the retrieved protein, Immune Epitope Database (IEDB) was used. The epitopes were predicted by B cell linear epitope prediction 2.0. The predicted epitopes were filtered and joined according to the length of the predicted B cell epitopes.
Moreover, the redundancy was removed. All B cell epitopes between lengths of 10 to 30 amino acids were selected. Moreover, to join these B cell epitopes, GPGPG spacer was used. GPGPG had been selected based on certain properties [18].

2.3. The T cell epitopes prediction and joining

To predict the T cell epitopes, similar approach of IEDB resource analysis was used which was used for B cell epitope prediction. IEDB recommended 2.22 method was used for prediction. Each field (except sequences and alleles) was satisfied with default settings. Alleles were selected according to HLA targeting efficiency. A*24–02, A*68–01, A*68–02, B*39–01 and A*32–02 were used for the analysis as they were used previously by the researchers [19]. T cell epitopes were filtered according to the two main approaches. The first approach was to filter the epitopes on the basis of IC50 value. Hence, the epitopes having less than 500 nm value of IC50 were selected. The second approach was to select peptides with percentile ranks below the 1% [20]. At the end, the epitopes were arranged according to their positions and redundancies. Then A linker of GPGPG was introduced to join the obtained epitopes. The addition of this spacer region was aimed to reduce the formation of junctional complexes.

2.4. Prediction for the protein structure

To predict the protein structure, a suitable template was searched in protein databank (PDB) database. Hence, homology modeling approach was used. This method can provide good models if the sequence identity is greater than 75% against the target protein and the template protein [21]. The BLAST result against PDB database highlighted that the protein has 75.12% identity with 95% query coverage against PDB protein (6ACC). Homology modeling was performed using swiss model webserver using 6ACC template. The .pdb file was downloaded for further analysis. Moreover, 3Dpro server was used to predict the tertiary structure of predicted and combined B and T cell epitopes. In 3Dpro, statistical terms in energy functions are used which are based on predicted structural features and Protein Databank knowledge. During modeling of the protein, a set based on fragment replacement and random perturbations is used to model the target protein. Moreover, the modeling strategy is based on simulated annealing with linear cooling and decisions are based on these strategies. Hence, at the end of the modeling, many models are predicted using random seeds and the energy of the models is calculated. Hence, the model with lowest energy is selected (Scratch Protein Predictor, 2015). To refine the predicted models, Galaxyweb webserver was used. A distinct feature of the server is to detect unreliable region and refine by Ab initio method (Ko et al., 2012). Furthermore, predicted structure validation was done by using Rampage analysis (Ho and Brasseur, 2005).

2.5. Highlighting epitopes on the structure

The positions of predicted epitopes on 3D structure of predicted protein was observed using Pymol [22]. The residues on the structure were selected according to the filtered epitopes and the epitopes were highlighted in red spheres.

2.6. Enriched analysis of the protein

In order to perform the enriched analysis of the protein, ProtParam webserver was used. The webserver depends on physical and chemical properties of the protein. It calculates the properties of the protein which are stored in TrEMBL or Swiss Prot. Many characteristics of a protein are included. The examples of such characteristics are atomic composition, aliphatic index, half-life estimation, extinction coefficient, grand average of hydropathicity (GRAVY), instability index, theoretical pI and molecular weight.

2.7. Antigenicity of single peptide epitope checked by ANTIGENpro

In order to predict antigenicity of a protein, ANTIGENpro was used. This tool is alignment free, sequence based and pathogen independent predictor of protein antigenicity. The antigenicity of the protein is predicted by machine learning algorithms. In order to summarize the results, support vector machine (SVM) classifier is used. Based on the probability, the tool predicts either the protein is likely to be antigenic or not. (Scratch Protein Predictor, 2015). The first server to predict the protective antigen is Vaxijen. Moreover, it performs alignment independent prediction and antigen classification which is based on the physicochemical properties of the query protein.

2.8. Allergenicity prediction using AllergenFP

The tool AllergenFP was used to predict either the protein is allergen or not. The algorithm of the tool transforms the strings into vectors by auto cross covariance (ACC) transformation. Moreover, in algorithm design of the program, principal component analysis of data matrix consist of 237 physicochemical properties of all amino acids. Among the components, first principal component or E1 imitates the hydrophobicity of amino acids. Moreover, E2, E3, E4 and E5 reflect the size, helix forming tendency, abundance and beta strand forming tendency respectively. ACC transformation make uniform length of the protein. At the end, the query protein is classified as allergen or non allergen based on Tanimoto coefficient (AllerTOP 2.0 2018). The webserver can be accessed by the URL http://ddg-pharmfac.net/AllergenFP/ and AllerTOP version 2.0 https://www.ddg-pharmfac.net/AllerTOP/.

2.9. Solubility checked by SOLpro

This webserver was used to predict either the protein is likely to be soluble upon overexpression in E. coli. The algorithm of the sever is based on SVM. During the analysis, a distinct set of features describing the sequence is used. At the end, SVM summarizes the results of predictions and we detect either the protein is soluble or not (Scratch Protein Predictor 2015). The webserver can be accessed by the URL http://download.igb.uci.edu/.

2.10. Prediction of TM helices

In order to predict the transmembrane helices in proteins, TMHMM was used to predict TM helices. The algorithm of
the webserver takes the query protein input in FASTA format. To get the output of the program, there are long and short output format options. It is used to find out potential transmembrane helices in vaccine (Server 2015). The webserver can be accessed by the url TMHMM version 2.0 [http://www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=5E6888480000202382AF3A72&wait=20].

2.11. Protein docking analysis against HLA-B7

In order to perform the docking analysis, HLA-B7 protein was used as the main protein. The PDB id of the protein is 3YCL and the resolution is 1.7 Å [23]. To perform the docking analysis, Molecular Operating Environment (MOE) tool was used. The predicted epitopes structure was docked against the main protein and the binding potential was analyzed. The program was used as mentioned by the researchers [24,25]. Briefly, protonation, removal of already bound peptides and energy minimization was performed. Moreover, water particles were removed. The docking analysis was performed by triangular matcher algorithm. The output consist of top ten positioned poses of the complex of protein and ligand. At the end, the docking analysis was examined by Ligplot tool. Moreover, Pymol was used to produce figures of docked complexes.

3. Results

3.1. Structural analysis of SARS-CoV-2 surface glycoprotein

To predict the structure of the protein, homology modeling approach was used. The predicted structure is shown in Fig. 1. The physicochemical properties of novel SARS-CoV-2 surface glycoprotein are predicted using Protparam which exhibited that it has 141.17847 kDa molecular weight, with 1273 amino acids. This reflects good antigenic nature of this protein. The isoelectric point (PI) of this protein was 6.24 that predicts its almost neutral in nature. The predicted structure was docked against the main protein and the binding potential was analyzed. The predicted epitopes structure was docked against the main protein and the binding potential was analyzed. The program was used as mentioned by the researchers [24,25]. Briefly, protonation, removal of already bound peptides and energy minimization was performed. Moreover, water particles were removed. The docking analysis was performed by triangular matcher algorithm. The output consist of top ten positioned poses of the complex of protein and ligand. At the end, the docking analysis was examined by Ligplot tool. Moreover, Pymol was used to produce figures of docked complexes.

3.2. The B cell epitopes prediction and joining

The Fig. 2 (A) represents a graph of predicted B cell epitopes that was developed by average, minimum and maximum score of the amino acid residues. The average score for the graph was −0.066, minimum score was −0.001 while maximum score was 2.291 as shown in Fig. 2 (A). The identified B cell epitopes are shown in Table 1. After filtering, total 13 epitopes were predicted based on the length of the epitope. The longest epitope has length of 27 amino acids with starting and ending positions of 461 and 487 respectively. Moreover, the smallest epitope has length of 10 amino acids with starting and ending positions of 44 and 53 respectively. The antigenicity index was predicted using Kolaskar & Tongaonkar Antigenicity prediction tool and the threshold value was 1.041 Fig. 2(B). The plot of surface accessibility epitope was predicted using default parameters with threshold value of 1.00 Fig. 2 (C). The hydrophilicity of the predicted epitopes was calculated using default parameters with threshold value 1.238 Fig. 2(D).

3.3. The prediction of T cell epitopes and joining

The prediction of T cell epitopes was performed using IEDB web server. Altogether we predicted five T cell epitopes after filtration of the data as mentioned in Table 2. As mentioned in material and method, the filtration of the predicted data was performed on the basis of IC50 and percentile rank of the predicted peptides. Hence, the selected peptides were having IC50 value less than 500 nm. Moreover, the peptides with percentile ranks below the 1% were selected. After filtration of the data, the epitopes mentioned in Table 2 were used for multi epitope peptide vaccine.

3.4. Highlighting the epitopes on the structure

The structure of the protein was predicted by homology modeling. The predicted structure is shown in Fig. 3. The predicted epitopes using IEDB web server were highlighted on the structure of the protein using Pymol. Altogether, 13 predicted epitopes were highlighted on the structure of the protein as shown on Fig. 3. Moreover, the structure of predicted B and T cell epitopes is shown in the following figure. The predicted structure’s quality was 91.9% predicted by Ramachandran plot.
3.5. Combine B cell and T cell epitopes to make a single peptide gene using GPGPG linker

To make a single peptide, the predicted B and T cell epitopes were combined using GPGPG spacer as shown below:

Table 1 Predicted B cell epitopes.

| Sr No. | Start | End | Peptide | Length |
|--------|-------|-----|---------|--------|
| 1      | 6     | 30  | VLLPLVSSQCVNLTTRTQLPPAYTN | 25     |
| 2      | 44    | 53  | RSSVLHSTQD | 10     |
| 3      | 62    | 81  | VTFHAIHVSGTNGTKRFDN | 20     |
| 4      | 90    | 101 | YFASTEKSNI | 12     |
| 5      | 107   | 131 | GTTLDSKTQSSLIVNNATNVVIKVC | 25     |
| 6      | 327   | 340 | VRFPNITNLCPFGE | 14     |
| 7      | 431   | 447 | GCVIAWNSNILDSKVVGG | 17     |
| 8      | 461   | 487 | LKPFOERTIIYQAAGSTPCNGVEFNN | 27     |
| 9      | 496   | 508 | GFQPURATIONGQFF | 13     |
| 10     | 807   | 817 | PDPSKPSKRFSF | 12     |
| 11     | 1066  | 1087| TYVPAQEKNFTAPAICHDGKA | 22     |
| 12     | 1096  | 1107| VSNGLTHWFTVTRQ | 12     |
| 13     | 1114  | 1139| IITDNTFYSGNCMDVIVGNTFYD | 26     |

Fig. 2 The graphs were generated using IEDB resource analysis by keeping the parameter at Default. The Y-axes depict correspondent score for all residues while the X axis represents their equivalent positions. (A) Plot of predicted bepipred B cell linear epitopes using threshold value 0.5 (B) Plot of Kolaskar & Tongaonkar Antigenicity prediction using online tools with threshold value 1.041 (C) The plot of surface accessibility epitope was predicted using default parameters to Emini surface accessibility prediction having threshold value of 1.00 (D) The hydrophilicity of the predicted epitopes were calculated using Parker hydrophilicity prediction tool having default parameters with threshold value 1.238.
3.6. Enriched analysis of the combined epitope protein

The structure of the multi-epitope peptide was predicted by 3Dpro Fig. 4 A. According to the analysis, pH of the peptide was 8.68 with Molecular weight of 39.67KDa. The Theoretical pI of the protein was 8.83. Moreover, total numbers of negatively charged residues were 18 while total numbers of positively charged residues were 24. The atomic composition of the peptide highlighted that it contains 1780 Carbon, 2693 Hydrogen, 487 Nitrogen, 529 Oxygen and 9 Sulfur atoms. Total numbers of atoms are 5498. The formula of the peptide is C1780H2693N487O529S9 while estimated half-life of the peptide is more than 20 h In vitro in yeast, 30 h In vitro in mammalian reticulocytes and more than 10 h In vivo in Escherichia coli. Moreover, the computed instability index (II) of the protein is 21.58. Hence, the protein was classified as stable protein.

3.7. Quality assessment of multi-epitope protein structure

The quality of the protein structure was determined by Ramachandran plot analysis. According to Ramachandran analysis, 91.9% residues of amino acids were present in the most favorable area of the plot. Moreover, 4.2% and 3.9% amino acid residues were fall in allowed regions and outer regions respectively as shown in Fig. 4 B. Moreover, in ProsA-Web the z-score of the multi-epitope protein structure was −3.23 as mentioned in Fig. 4 C.

3.8. Transmembrane probability analysis and Phobius prediction for secretory protein

The analysis was performed by TMHMM to check either the protein is transmembrane or not. According to TMHMM posterior probability, the protein was found non transmembrane as shown in the Fig. 5 A. Phobius analysis was performed to check either the protein is secretory or not. According to the analysis, it was found that the protein is non-secretory as shown in the Fig. 5 B.

3.9. Protein-peptide interaction visualization using Ligplot

The visualization of docking analysis i.e. protein-peptide interaction visualization was performed using Ligplot program. This program produces 2D schematic representations of docking complex. The input file was related to Protein Data Bank file (.pdb) while the output file consist of docking interactions. The docking analysis highlighted three interactions of epitope against HLA-B7 as shown in Fig. 6. The interacting residues include Lys 178, Gln 303 and Thr 31. Among them, the highest binding affinity was found against Thr 31 with 2.74 Å distance. Moreover, least binding affinity was found against Lys 178 with 3.18 Å distance.

4. Discussion

Emerging and pandemic zoonotic diseases have become an emerging public health concern for the world. Zoonotic diseases have also posed a threat to global health security. Almost 60 percent of infectious diseases and 75 percent of emerging infectious diseases are zoonotic in origin. Infectious diseases are responsible for 15.8% deaths globally and 43.7% deaths in Low-income countries. Each year, zoonotic diseases are account for 2.7 million human deaths worldwide [26,27]. Along with human deaths, zoonotic diseases also lead to heavy economic losses to the world specially; low income countries. For instance, last epidemic of Ebola in 2014 was responsible

| Allele          | start | end | length | Peptide          | method       | Percentile Rank | ann_ic50 | ann_rank |
|----------------|-------|-----|--------|------------------|--------------|----------------|----------|----------|
| HLA-A*24:02    | 1     | 13  | 13     | PTWRVYSTGSNVF   | ann          | 0.57           | 352.54   | 0.57     |
| HLA-A*68:01    | 14    | 23  | 10     | HVTYVPAQEK      | Consensus (ann/smm) | 0.83       | 20.9     | 0.16     |
| HLA-A*24:02    | 24    | 37  | 14     | KVGNYNLYRLFR    | ann          | 0.6            | 386.63   | 0.6      |
| HLA-A*68:01    | 40    | 53  | 14     | LCFTNVYADSFVIR  | ann          | 0.49           | 55.79    | 0.49     |
| HLA-A*24:02    | 54    | 67  | 14     | AYYVGYLQPRFLLL  | ann          | 0.17           | 111.15   | 0.17     |

**Table 2** Predicted T cell epitopes on the basis of percentile rank and ic50.

![Figure 3](image-url)

**Fig. 3** Highlighting B cell epitopes on the structure of the protein. Altogether, 13 predicted epitopes are highlighted on the structure of the protein as shown in the figure. The positions of predicted epitopes on 3D structure of predicted protein was observed using Pymol software and highlighted red in colour.
Emerging viruses pose a great challenge to medicine and science because little is known about them before they emerge. Human coronaviruses are one of the most important human viruses that pose great threat to global health. During past two decades, many coronaviruses caused serious problems in both animals and humans [29]. In 2003, severe acute respiratory syndrome coronavirus (SARS-CoV) epidemic was lead to 299 deaths among 1755 infected people with death ratio of 6:1. The horseshoe bats were later traced as primary reservoir of SARS-CoV [30]. Later on in 2012, Middle East respiratory syndrome coronavirus (MERS-CoV) was discovered from the patients died due to mysterious fatal disease [31]. The dromedary camels were found to be the source of MERS-CoV spread in humans [32]. Recently, in December 2019, a novel human corona virus (SARS-CoV-2) is discovered in wuhan, China. More than 0.3 million people are infected with this virus with more than 13,000 deaths reported. Scientists have started working on therapeutic, vaccine and diagnostic kits for SARS-CoV-2. Conventional approaches to make vaccine production and diagnostics require handsome amount of time. However, modern use of information technology in the discipline of virology has become very helpful to develop effective vaccine and diagnostics in a short time.

Bioinformatics is a good option to be used in development of vaccines and diagnostics for newly emerged viruses [15]. The use of this approach can able to reduce the time and cost. In order to construct a potent vaccine and effective diagnosis, understanding of the epitope and antibody interaction is required. The surface glycoprotein of human corona viruses is the primary target for neutralizing antibodies thus a good target for vaccine development and diagnostics. The surface glycoprotein of SARS-CoV-2 consists of 1273 amino acids. The surface glycoprotein of SARS-CoV-2 is relatively smaller than MERS-CoV, has 1353 amino acids and comparatively closer to SARS-CoV, has 1255 amino acids [33]. This predict that SARS-CoV-2 is closer to SARS-CoV moreover; recent studies also confirm that SARS-CoV-2 is 75% to 80% genetically similar to SARS-CoV [5,14]. The surface glycoprotein mainly involve in the entry of the virus into the cell during virus replication cycle [34]. The surface glycoproteins of coronaviruses are able to produce neutralizing antibodies, and block virus entry or neutralize the viral infection [35,36].

The SARS-CoV-2 spike (S) protein is a 1273 amino acid protein consists of N-terminus signal peptide, the S1 subunit responsible for receptor binding and S2 subunit responsible for membrane fusion [37]. The S1 domain comprises of N-terminus domain and most important receptor binding domain; while S2 domain consists of fusion peptide, heptapeptide repeat sequence 1, heptapeptide repeat sequence 2, transmembrane (TM) domain, and heptapeptide repeat sequence 3.
domain and cytoplasmic domain [38]. In native state, the S protein exists as an inactive form. The proteases of the target cell activate the S protein during the infection by cleaving into two subunits (S1 and S2) [39]. The SARS-CoV-2 S protein is predicted to be the conserved in all human coronaviruses. Overall, S protein is involved in viral attachment and entry into the host cell. Because of its important functions, S protein has become one of the main targets for vaccines and antivirals. Recently, it has found that polyclonal antibodies against SARS-CoV-2 S protein neutralize the virus mediated entry into the host cells that encourages the use of S protein target for immunotherapies [40].

B cell epitopes are the region of amino acids recognized by the B cell receptor or specific antibodies [41]. The B cell epitope was predicted using Immune Epitope Database. After filtering, a total of 13 epitopes were predicted in surface glycoprotein of SARS-CoV-2 based on the length of the epitopes. The joining of B cell epitopes was done by the GPGPG linkers. The linkers are having a prominent role in the epitope based vaccine. The addition of GPGPG linkers help to decreases the binding affinity specifically around the core binding regions. Moreover, rich GPGPG linker is associated with beta turns with no effect on main structure of the protein. Moreover, the linkers not only prevent the immune processing of antigen but also prevent the generation of junctional epitopes. The linkers also help to present the selected epitopes to HLA-II [42,43].

CD4 or and CD8+ receptors of T cell are important part of the cell where short linear sequence of T cell epitopes are
Identification of epitopes in coronavirus

This analysis will be helpful to develop effective low cost epitope based vaccines and diagnostics to diagnose and control imminent novel SARS-CoV-2 challenge.

5. Author’s contribution
Muhammad Asif Rasheed and Soail Raza are equal contributing authors.

Declaration of Competing Interest
The authors declared that they have no conflict of interest and the paper presents their own work which does not been infringe any third-party rights, especially authorship of any part of the article is an original contribution, not published before and not being under consideration for publication elsewhere.

References
[1] Stanley Perlman, Another decade, another coronavirus, N. Engl. J. Med. (2020).
[2] D.S. Hui, I.A. E, T.A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T.D. McHugh, Z.A. Memish, C. Drosten, A. Zumla, E. Petersen, The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China, Int. J. Infect. Dis. 91 (2020) 264–266.
[3] X.Y. Ning Dong, Lianwei Ye, Kaichao Chen, Edward Wai-Chi Chan, Mengsu Yang, Sheng Chen Genomic and protein structure modelling analysis depicts t 1 he origin and infectivity of 2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China, bioRxiv (2020).
[4] J. Parry, China coronavirus: cases surge as official admits human to human transmission, Brit. Med. J. (2020).
[5] S.Y. Jasper Fuk-Woo Chan*, Kin-Hang Kok*, Kelvin Kai-Wang To*, Hin Chu*, Jin Yang, Fanfan Xing, Jieling Liu, Cyril Chik-Yan Yip, Rosana Wang-Shan Poon, Hoi-Wah Tsoi, Simon Kam-Fai Lo, Kwok-Hung Chan, Vincent Kwok-Man Poon, Wan-Mui Chan, Jonathan Daniel Ip, Jian-Piao Cai, Vincent Chi-Chung Cheng, Honglin Chen, Christopher Kim-Ming Hui, Kwok-Yung Yuen, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a family of a cluster family, The Lancet (2020) 30154–30159.
[6] W.W. Wei Ji, Xiaofang Zhao, Junjie Zai, Xingguang Li, Homologous recombination within the spike glycoprotein of the newly identified coronavirus 2019-nCoV may boost cross-species transmission from snake to human, J. Med. Virol. (2020).
[7] E.C.a.D. Cyranoski, Why snakes probably aren’t spreading the new China virus, Nature (2020).
[8] T.S. Fung, D.X. Liu, Human coronavirus: host-pathogen interaction, Annu. Rev. Microbiol. 73 (2019) 529–557.
[9] D. Schoeman, B.C. Fielding, Coronavirus envelope protein: current knowledge, Virol. J. 16 (2019) 69.
[10] T.R. Ruch, C.E. Machamer, The coronavirus E protein: assembly and beyond, Viruses 4 (2012) 363–382.
[11] R.N. Kirchdoerfer, C.A. Cottrill, N. Wang, J. Pallesen, H.M. Yassine, H.L. Turner, K.S. Corbett, B.S. Graham, J.S. McLellan, A.B. Ward, Pre-fusion structure of a human coronavirus spike protein, Nature 531 (2016) 118–121.
[12] v.Z.M. R. McBride, B.C. Fielding, The coronavirus nucleocapsid is a multifunctional protein, Viruses 6 (2014) 2997–3018.

Fig. 6 Docking analysis results highlight the interactions. The docking analysis highlighted three interactions of epitope against HLA-B7. The interacting residues include Lys 178, Gol 303 and Thr 31. Among them, the highest binding affinity was found against Thr 31 with 2.74 Å distance. Moreover, least binding affinity was found against Lys 178 with 3.18 Å distance.

![Image](52x545 to 278x738)
[13] P.S. Masters, The molecular biology of coronaviruses, Adv. Virus Res. 66 (2006) 193–292.
[14] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G.F. Gao, W. Tan, A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. (2020).
[15] S. Raza, K. Siddique, M. Rabbani, T. Yaqub, A.A. Anjum, M. Ibrahim, M. Azhhar, F. Jamil, M.A. Rasheed, In silico analysis of four structural proteins of aphyovirus serotypes revealed significant B and T cell epitopes, Microb. Pathog. 128 (2019) 254–262.
[16] A.R. Oany, A.A. Emran, T.P. Jyoti, Design of an epitope-based peptide vaccine against spike protein of human coronavirus: an in silico approach, Drug Des. Devel. Ther. 8 (2014) 1139–1149.
[17] J.E. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes, Immunome Res. 2 (2006) 2.
[18] X. Ruan, D.E. Loyola, C.L. Marolda, J.M. Perez-Donoso, M. A. Valvano, The WaaL O-antigen lipopolysaccharide ligase has features in common with metal ion-independent inverting glycosyltransferases, Glycobio 22 (2012) 288–299.
[19] A. Di Gian Promendoni, M.J. Matewish, A. Bisaillon, J.R. Stehle, J. S. Lam, P. Castric, Glycosylation of Pseudomonas aeruginosa 1244 pilin: glycan substrate specificity, Mol. Microbiol. 46 (2002) 519–530.
[20] C. Gebhart, M.V. Ielmini, B. Reiz, N.L. Price, F.E. Aas, M. Koomey, M.F. Feldman, Characterization of exogenous bacterial oligosaccharidyltransferases in Escherichia coli reveals the potential for O-linked protein glycosylation in Vibrio cholerae and Burkholderia thailandensis, Glycobio 22 (2012) 962–974.
[21] F. Sievers, D.G. Higgins, Clustal Omega, accurate alignment of very large numbers of sequences, Methods Mol. Biol. 1079 (2014) 105–116.
[22] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729.
[23] R.M. Brennan, J. Petersen, M.A. Neller, J.J. Miles, J.M. Burrows, C. Smith, J. McClusky, R. Khanna, J. Rossjohn, S. R. Burrows, The impact of a large and frequent deletion in the human TCR beta locus on antiviral immunity, J. Immunol. 188 (2012) 2742–2748.
[24] M. Gurusamy, J.F. Abdul, Lead optimization studies towards finding NS2B/NS3 protease targetspecific inhibitors as potential anti-dengue drug-like compounds, Curr. Drug Discov. Technol. 16 (2019) 307–314.
[25] H. Liu, R. Wu, Y. Sun, Y. Ye, J. Chen, X. Luo, X. Shen, H. Liu, Identification of novel thiadiazoloacylamide analogues as inhibitors of dengue-2 virus NS2B/NS3 protease, Bioorg. Med. Chem. 22 (2014) 6344–6352.
[26] K.E. Jones, N.G. Patel, M.A. Levy, A. Storeygard, D. Balk, J.L. Gittleman, P. Daszak, Global trends in emerging infectious diseases, Nature 451 (2008) 990–993.
[27] S.J. Salyer, R. Silver, K. Simone, C. Barton Behravesh, Prioritizing zoonesos for global health capacity building-themes from one health zoonotic disease workshops in 7 countries, 2014–2016, Emerg. Infect. Dis. 23 (2017).
[28] L.H. Nel, The role of non-governmental organisations in controlling rabies: the Global Alliance for Rabies Control, Partners for Rabies Prevention and the Blueprint for Rabies Prevention and Control, Rev. Sci. Tech. 37 (2018) 751–759.
[29] S.K. Lau, J.F. Chan, Coronaviruses: emerging and re-emerging pathogens in humans and animals, Viral. J. 12 (2015) 299.
[30] X.Y. Ge, J.L. Li, X.L. Yang, A.A. Chmura, G. Zhu, J.H. Epstein, J.K. Mazet, B. Hu, W. Zhang, C. Peng, Y.J. Zhang, C. M. Luo, B. Tan, N. Wang, Y. Zhu, G. Cramer, S.Y. Zhang, L. F. Wang, P. Daszak, Z.L. Shi, Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor, Nature 503 (2013) 535–538.
[31] A.M. Zaki, S. van Boeheim, T.M. Bestebroer, A.D. Osterhaus, R.A. Fouchier, Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia, N. Engl. J. Med. 367 (2012) 1814–1820.
[32] B.L. Haagmans, S.H. Al Dhahiry, C.B. Reusken, V.S. Raj, M. Galiana, R. Myers, G.J. Godeke, M. Jonges, E. Farag, A. Diah, H. Ghoabasby, F. Alhajri, M. Al-Thani, S.A. Al-Marri, H.E. Al Romainh, A. Al Khal, A. Bermingham, A.D. Osterhaus, M.M. AlHajri, M.P. Koopmans, Middle East respiratory syndrome coronavirus in domedary camels: an outbreak investigation, Lancet Infect. Dis. 14 (2014) 140–145.
[33] C.M. Petit, J.M. Melancon, V.N. Chouljenko, R. Colgrove, M. Farzan, D.M. Knipe, K.G. Kousoulas, Genetic analysis of the SARS-coronavirus spike glycoprotein functional domains involved in cell-surface expression and cell-to-cell fusion, Virology 341 (2005) 215–230.
[34] F. Li, Structure, function, and evolution of coronavirus spike proteins, Annu. Rev. Virol. 3 (2016) 237–261.
[35] M.H. Kim, H.J. Kim, J. Chang, Superior immune responses induced by intranasal immunization with recombinant adenovirus-based vaccine expressing full-length Spike protein of Middle East respiratory syndrome coronavirus, PLoS ONE 14 (2019) e0220196.
[36] Z.Y. Yang, W.P. Kong, Y. Huang, A. Roberts, B.R. Murphy, K. Subbarao, G.J. Nabel, A DNA vaccine induces SARS coronavirus neutralization and protective immunity in mice, Nature 428 (2004) 561–564.
[37] Y. Huang, C. Yang, X.F. Xu, W. Xu, S.W. Liu, Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19, Acta Pharmacol. Sin 41 (2020) 1141–1149.
[38] S. Xia, Y. Zhu, M. Liu, Q. Lan, W. Xu, Y. Wu, T. Ying, S. Liu, Z. Shi, S. Jiang, L. Lu, Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein, Cell Mol. Immunol. 17 (2020) 765–767.
[39] S. Bertram, R. Dijkman, M. Habjan, A. Heurich, S. Gierer, I. Glowacka, K. Welsch, M. Winkler, H. Schneider, H. Hofmann-Winkler, V. Thiél, S. Pohlmann, TMRPSS2 activates the human coronavirus 229E for cathepsin-independent host cell entry and is expressed in viral target cells in the respiratory epithelium, J. Virol. 87 (2013) 6160–6160.
[40] A.C. Walls, Y.J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell 183 (2020) 1735.
[41] H. Yu, L.F. Jiang, D.Y. Fang, H.J. Yan, J.J. Zhou, J.M. Zhou, Y. Liang, Y. Gao, W. Zhao, B.G. Long, Selection of SARS-coronavirus-specific B cell epitopes by phage peptide library screening and evaluation of the immunological effect of epitope-based peptides on mice, Virology 359 (2013) 6150–6160.
[42] F.A. Beatriz Meza, Arturo Pedro Sierra-Beltrán, Javier Torres, Carlos Angulo, A novel design of a multi-antigenic, multistage and multi-epitope vaccine against helicobacter pylori: an in silico approach, Infect. Genet. Evol. 49 (2017) 309–317.
[43] Z.K. Navid Nezafat, Mahboobeh Eslami, Milad Mohkam, Sanam Zandian, Younes Ghasemi, Designing an efficient multi-epitope peptide vaccine against Vibrio cholerae via combined immunoinformatics and protein interaction based approaches, Comput. Biol. Chem. 62 (2016) 8–95.
[44] A.E.E. Tarek, A. Ahmad, Laila H. El-Sayed, T-cell epitope mapping for the design of powerful vaccines, Vaccine Rep. 6 (2016) 13–22.
[45] A.I. Ashik, M. Hasan, A.T. Tasnim, M.B. Chowdhury, T. Hossain, S. Ahmed, An immunoinformatics study on the spike protein of SARS-CoV-2 revealing potential epitopes as vaccine candidates, Heliyon 6 (2020) e04865.
[46] J. He, F. Huang, J. Zhang, Q. Chen, Z. Zheng, Q. Zhou, D. Chen, J. Li, J. Chen, Vaccine design based on 16 epitopes of SARS-CoV-2 spike protein, J. Med. Virol. (2020).

[47] A. Samad, F. Ahammad, Z. Nain, R. Alam, R.R. Imon, M. Hasan, M.S. Rahman, Designing a multi-epitope vaccine against SARS-CoV-2: an immunoinformatics approach, J. Biomol. Struct. Dyn. (2020) 1–17.

[48] S.G. Marsh, E.D. Albert, W.F. Bodmer, R.E. Bontrop, B. Dupont, H.A. Erlich, M. Fernandez-Vina, D.E. Geraghty, R. Holdsworth, C.K. Hurley, M. Lau, K.W. Lee, B. Mach, M. Maiers, W.R. Mayr, C.R. Muller, P. Parham, E.W. Petersdorf, T. Sasazuki, J.L. Strominger, A. Svejgaard, P.I. Terasaki, J.M. Tiercy, J. Trowsdale, Nomenclature for factors of the HLA system, Tissue Antigens 75 (2010) (2010) 291–455.

[49] A.S. De Groot, D.S. Rivera, J.A. McMurry, S. Buus, W. Martin, Identification of immunogenic HLA-B7 “Achilles’ heel” epitopes within highly conserved regions of HIV, Vaccine 26 (2008) 3059–3071.

[50] C.H. Lee, H. Koo hy, In silico identification of vaccine targets for 2019-nCoV, F1000Res. 9 (2020) 145.