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.

Vaccine 39 (2021) 1087–1095

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Immuno-informatics approach for B-cell and T-cell epitope based peptide vaccine design against novel COVID-19 virus

Jitender Singh [a], Deepti Malik [b], Ashvinder Raina [a,*]

[a] Post Graduate Institute of Medical Education and Research, Sector-12, Chandigarh 160 012, India
[b] Assistant Professor, AIIMS, Bilaspur, India

ARTICLE INFO

Article history:
Received 20 May 2020
Received in revised form 8 November 2020
Accepted 4 January 2021
Available online 9 January 2021

Keywords:
COVID-19
Vaccine
Epitopes
B-cell
T-cell
3D modeling
Docking

ABSTRACT

COVID-19 has brought the world to a standstill with a wave of destruction in country after country with tremendous loss of lives and livelihood in advanced to developing nations. Whole world is staring at the prospect of repeated lockdowns with another wave of COVID-19 predicted to hit the world in September of 2020. The second wave is assumed to be even more destructive with severe impact across much of the world. The only way to defeat this pandemic is to quickly develop a safe and effective vaccine against this raging menace and initiate a global vaccination drive. Our study is an attempt to deploy various computational methods to identify B-cell and T-cell epitopes from the spike surface glycoprotein of SARS-COV-2 which have the novel potential for vaccine development against COVID-19. For this we have taken 8 unique strains with one each from India, China, France, USA, Italy, Australia, Iran and Pakistan. The strain data was extracted from NCBI Database. By analyzing the immune parameters like surface accessibility, antigenicity, variability, conservancy, flexibility, hydrophilicity, allergenicity and toxicity of the conserved sequences of spike glycoprotein using various databases and bioinformatics tools, we identified two potential novel linear (SGTNGTKRFDN and ASVYAWNRK) and one structural B-cell epitope as well as two T-cell epitopes (RLFRKSNLK and IPTNFTISV) which can be used as epitope-based peptide vaccines. Docking simulation assay revealed that above T-cell epitopes have minimum free binding energy and showed strong hydrogen bond interaction which strengthened its potential as being a T-cell epitope for the epitope-based novel vaccine against SARS-CoV-2. This study allows us to claim that B-cell and T-cell epitopes mentioned above provide potential pathways for developing an exploratory vaccine against spike surface glycoprotein of SARS-CoV-2 with high confidence for the identified strains. We will need to confirm our findings with biological assays.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

COVID-19 pandemic, a severe acute respiratory syndrome caused by coronavirus 2 (SARS-CoV-2), has emerged as a serious public health threat [1]. Humanity needs to quickly understand the underlying mechanism of SARS-CoV-2 and how it invades and infects the host cells and try and develop effective and safe vaccines against this virus. Our study proposes development of an effective vaccine for SARS-CoV-2 by analyzing linear B-cell and T-cell epitopes from the spike surface glycoprotein of SARS-CoV-2 and its conservancy in all the available strains of this virus. Fourteen proteins found till now from SARS-CoV-2 virus with proteome id UP000464024 as in uniprot database (https://www.uniprot.org/proteomes) [2]. The most important protein for SARS-CoV-2 is spike protein domain S1 which enables SARS-CoV-2 to bind onto human cells using the ACE2 and CLEC4M/DC-SIGNR receptors which leads to virus invading host cell endosomes leading to conformational changes in the S glycoprotein whereas Spike protein domain S2 plays the role of fusion protein helping SARS-CoV-2 to fuse with host cell membranes [3]. Spike glycoprotein which consists of 1273 amino acid residues primary sequences has three conformational states: pre-fusion native state, pre-hairpin intermediate state, and post-fusion hairpin state. As SARS-CoV-2 is binding with the host cell membrane, the coiled coil regions (heptad repeats) create a hairpin structure, kind of a trimer which positions the fusion peptide very close to C-terminal region of the ectodomain [4]. This structure facilitates the fusion of SARS-CoV-2 with host cell membranes [5]. This shows that spike glycoprotein provides SARS-CoV-2 with the ability to seamlessly bind with host cell membranes, giving a clear path to develop vaccines which should potentially induce antibodies that have the ability to

* Corresponding author.
E-mail address: drashvinder@gmail.com (A. Raina).

https://doi.org/10.1016/j.vaccine.2021.01.011
0264-410X/© 2021 Elsevier Ltd. All rights reserved.

block viral fusion and binding abilities with host cell membranes leading to effective blocking of SARS-CoV-2. The S protein plays a key role in the induction of neutralizing-antibody and T-cell responses, as well as protective immunity, during infection with SARS-CoV-2 [6]. The present study aims to identify B-cell and T-cell epitopes from the spike surface glycoprotein of SARS-CoV-2 and analyse their conservancy in all the available strains of COVID-19. Epitopes with high specificity confidence were generated. Potential B cell epitopes were discovered having high conservancy implying wider protection against multiple strains. The two linear B-cell epitopes were highly antigenic with properties like surface accessibility, flexibility and hydrophilicity. We also observed one structural B cell epitope which has high probability to be used for vaccine due to its best threshold value (Propensity and DiscoTope Score). For the T cell epitopes, we performed a population coverage analysis and proposed a set of epitopes that is estimated to provide broad coverage globally. Our findings provide a screened set of epitopes that can help guide experimental efforts towards the development of vaccines against SARS-CoV-2. 3D structures and docking studies of finally selected conserved T-cell epitopes among the selected strains and selected MHC I alleles were performed enabling us to identify two potential novel T-cell epitopes which can be used as epitope-based peptide vaccines. To the best of our knowledge, our study has discovered novel B- and T-cell epitopes that could act as potential candidates for vaccine development against SARS-CoV-2. Overall, the results of this study can provide the basis for further research that could lead to a novel therapeutic approach for patients with SARS-CoV-2.

2. Materials and methods

2.1. Protein sequence retrieval

The present study comprises of eight different strains of spike glycoprotein from eight different countries including India, China, France, USA, Italy, Australia, Iran and Pakistan. Sequences of the spike surface glycoprotein of SARS-CoV-2 for all the strains of the virus were retrieved from the NCBI database [7]. The sequences were extracted from the database in the FASTA format. The length of each of the eight sequences for spike surface glycoprotein of SARS-CoV-2 was 1273 amino acids. The NCBI accession no of all the selected strains number as India (Accession no QJF11884.1), China (QHD43416.1), France (QIX12148.1), USA (QID21058.1), Italy (QIC50498.1), Australia (QHR84449.1), Iran (QIX12195.1) and Pakistan (QIS60276.1).

2.2. Variability analysis of spike surface glycoprotein

To analyse the level of conservation, the retrieved sequences were aligned by using EBI-Clustal Omega program [8] and multiple sequence alignment (MSA) program. MSA was visualized using Jalview [9]. The absolute site variability in MSA was calculated using Protein Variability Server (PVS) [10]. PVS makes use of several variability metrics to compute absolute variation in multiple protein-sequence alignments (MSAs) [10].

2.3. Prediction of antigenicity of spike surface glycoprotein

In order to develop a peptide vaccine, it is essential to identify proteins which display antigenic features. A reference spike surface glycoprotein having accession number QHD43416.1 was tested for its antigenicity. VaxiJen v2.0 server [11] and Kolaskar & Tongaonkar method [12] were used to predict the antigenic property of the given sequence. VaxiJen uses alignment-independent prediction to predict the antigenicity of a given protein [12].

2.4. B-cell epitope prediction

Three tools were utilized for the prediction of linear B-cell epitopes to ensure enhanced accuracy in results. Sequence was fed into BepiPred 2.0 IEBD [13], BepiPred-2.0: Sequential B-Cell Epitope Predictor [14] and ABCpred [15] web servers. BepiPred 2.0 IEBD [13] server analyzes epitopes using Hidden Markov model and propensity scale. The BepiPred-2.0 Sequential B-Cell Epitope Predictor server predicts B-cell epitopes from a protein sequence using Random Forest algorithm trained on epitopes and non-epitope amino acids with threshold value set above 0.5 for all of the servers [14]. The 0.75 threshold and a window length of 12 amino acids were set as parameters for the prediction of epitopes in ABCpred server [15]. The predicted epitopes from these three servers were scrutinized and epitopes that were commonly recognized by all three servers were selected for further analysis. The structural epitopes in the 3D protein structures were analyzed with DiscoTope: Structure-based Antibody Prediction tool [16] and the threshold in the settings was set at −7.0, which increase specificity and sensitivity. By this we analyzed the positive and negative prediction of residues and positive predicted residues were considered for epitope based vaccine.

2.5. Surface accessible regions prediction

In order to determine the surface accessibility of the epitope, Emini surface accessible prediction tool of the IEDB was used [17,18]. Epitopes that were found to have surface accessibility were selected and analyzed for conservancy.

2.6. B cell epitopes conservancy analysis

An ideal epitope should be conserved so that it provides wider protection against multiple strains [19]. Conservancy analysis of surface accessible epitopes with all the spike surface glycoprotein strains of SARS-CoV-2 sequences was analyzed by the IEDB epitope conservancy tool [20]. For the prediction of the conservancy, sequence identity cut-off was set to 90%.

2.7. Antigenicity prediction of selected B-cell epitopes

Selected B-cell epitopes were checked for their antigenicity by VaxiJen v2.0 server with a threshold value of 0.4 [11].

2.8. Prediction of flexibility and hydrophilicity of B cell epitopes

Studies have reported that hydrophilicity and flexibility of a peptide are related to its antigenicity [21]. For this, conserved epitopes were submitted to Karplus and Schulz (KS) flexibility online tool [22], and Parker hydrophilicity prediction tool for flexibility and hydrophilicity predictions respectively [23].

2.9. Prediction of T cell epitopes and conservancy analysis

T cell epitopes were identified with the help of NetCTL online server [24,25]. The threshold was set to 1.50 and the sensitivity and specificity were set to 0.54 and 0.993 respectively. MHC-I alleles interacting with each of the selected epitopes were determined by MHC-I prediction server [26] of IEDB. Stabilized matrix method (SMM) [26] was used for the prediction of half maximal inhibitory concentration (IC50) of peptide binding to MHC-I alleles. The cut-off value of IC50 was set to 50 nM. For the analysis of binding of the epitope to the allele all the available MHC class I alleles were selected and the peptide lengths were set to 9 amino- acids. For MHC class II binding prediction, whole protein sequences of spike surface glycoprotein was submitted in the NetMHCII 2.3 Server

[27]. IC50 values and % rank of the epitopes binding to MHC II molecules were calculated using the Stabilized Matrix Base Method (SMM). Epitopes with length of 15 amino acid and those that interacted with highest number of alleles was selected with IC50 cut-off value set to 50 nM with threshold % rank which predicts binding affinity set to 0.5. Predicted T-cell epitopes were submitted to IEDB conservancy analysis tool with a sequence identity threshold of 90 percent [20] and world population coverage analysis [28].

2.10. Prediction for allergenicity and toxicity of the selected T-cell epitopes

For determining allergenicity of the selected T cell epitopes, AllerTOP v. 2.0 was used which reports results with 94% sensitivity [29]. ToxinPred web server was used to determine toxicity of the selected T-cell epitopes [30].

2.11. World population coverage analysis of the predicted T-cell epitopic peptides

To analyse world population coverage, selected T-cell epitopic sequences with corresponding Class I and II HLA alleles were submitted to the population coverage analysis tool of IEDB using default analysis parameters [31].The alleles used for the input predictions were HLA-A*03:01, HLA-C*12:03, HLA-A*30:01, HLA-A*11:01, HLA-DRB1*11, HLA-0101, HLA-A0201,HLA-A0202, HLA-A0203, HLA-A0205, HLA-A0206, HLA-A0207, DRB1_0101, DRB1_0103, DRB1_0301, DRB1_0401, DRB1_0402, DRB1_0403, DRB1_0404, DRB1_0405, DRB1_0701, DRB1_0801, DRB1.

*2.12. 3D structures of conserved T-cell epitopes and selected HLA-C 12*03*

3D structures of the selected peptides were constructed using PEP-FOLD Peptide Structure Prediction server [32,33]. The best models provided by the server were chosen for the docking assay. The MHC-I allele, HLA-C 12*03, was found to interact with majority of the predicted T cell epitopes. Therefore, 3-D model of HLA-C 12*03 was constructed using SWISS MODEL server [34]. The model was chosen based on GMQE and QMEAN scores of the model and the sequence identity and coverage of the template. The HLA-C 12*03 3D model was evaluated by PROCHECK software [35] and ProSA web tool [36]. Ramachandran plot was generated using the PROCHECK software analyses [35].

2.13. Docking simulation assay

In silico docking was carried out between conserved T-cell epitopes with the HLA-C 12*03 allele. PyRx was utilised for the docking purpose [37]. PyRx is combination of several softwares such as AutoDock Vina, AutoDock 4.2, Mayavi, Open Babel, etc. In this present study, AutoDock Vina was used in PyRx for the docking purpose [38]. The free energy was calculated for binding of the epitopes to the binding groove of the HLA-C*12:03. The binding of the epitope to the allele was then visualized using PyMol [39].

3. Results and discussion

It has been reported by UniProt and NCBI database that spike surface glycoprotein of SARS-CoV-2 contains 1273 amino acids. Multiple studies across the globe have clearly pointed out that the S Protein (one of various proteins of SARS-CoV-2) has a vital role in inducing the neutralizing-antibody and T-cell responses which lead to development of protective immunity against the

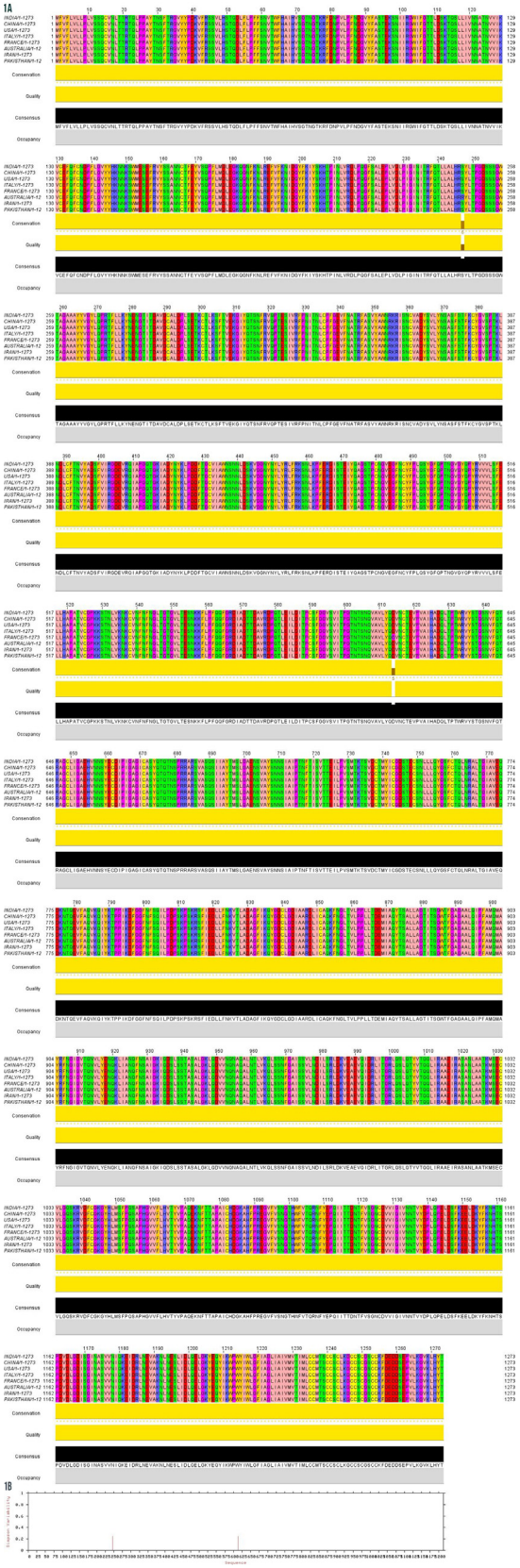

Fig. 1. (A) Jalview representation of multiple sequence alignment of all the 8 different strains sequences of spike surface glycoprotein of COVID-19. (B) The protein variability index of spike surface glycoprotein was determined by using PVS server. The threshold of conservancy was kept at 1.0 for our analysis. X axis indicates the position of amino acid sequence and Y axis indicates the Shannon entropy.

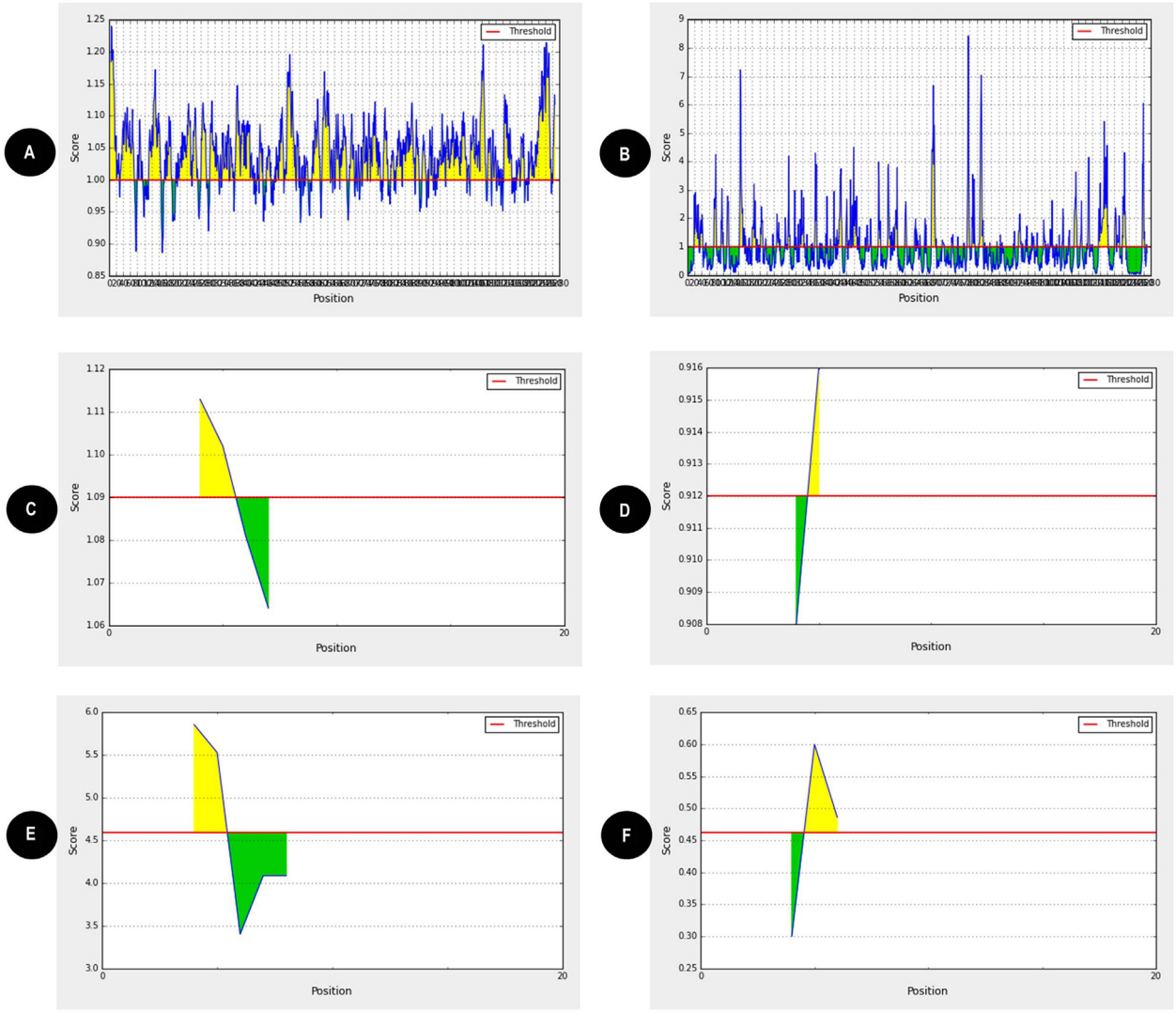

Fig. 2. (A) The spike surface glycoprotein is highly antigenic. VaxiJen v2.0 server and Kolaskar and Tongaonkar method were used to predict the antigenic property of the given sequence. The threshold was kept at 1.00; highlighted region above the threshold are antigenic. X axis indicates the position of amino acid sequence and Y axis indicates the antigenic score of amino acid residues. (B) The spike surface glycoprotein is surface accessible. The red horizontal line indicates surface accessibility cutoff and highlighted regions above threshold line are surface accessible epitopes. The threshold was kept at 1.00 for our analysis. X axis indicates the position of amino acid sequence and Y axis indicates surface accessible epitopes indicating minimum score as 0.031 and maximum score 8.415. Emini surface accessible prediction tool of the IEDB was used to determine the surface accessibility of the epitope. (C) Representative image indicating flexibility of two B-cell epitopes SGTNGTKRFDN and ASVYAWNRK (D). With a window size of 7 amino acids and center position as 4, both epitopes were found to have flexibility above the threshold of 0.9. Flexibility predictions were determined using Karplus and Schulz (KS) flexibility online tool. Image indicating hydrophilicity of two B-cell epitopes SGTNGTKRFDN (E) and ASVYAWNRK (F). With a window size of 7 amino acids and centre position as 4, highlighted region specifies that both the epitopes were found to have hydrophilicity above the threshold of 0.40. Hydrophilicity predictions were determined using Parker hydrophilicity prediction tool. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Table indicating there was an overlap of Eleven B-cell epitopes by all three prediction servers'. Their lengths and sequence positions were also given by the servers.

Sr. No	Sequence	Length	Start	End
1	KTPPIKDFGGF	11	790	801
2	QLPPAYTNSFTR	12	23	34
3	LGVYYHKNNK	10	141	150
4	ADAGFIKQYG	10	829	838
5	GDEVRQIAPGQT	12	404	415
6	NSASFSTFKCY	11	370	380
7	SGTNGTKRFDN	11	71	81
8	FSTFKCYGVSPT	12	374	385
9	SNLKPFERDIST	12	459	470
10	CYGVSPTKLNDL	12	379	390
11	ASVYAWNRK	9	348	357

SARS-CoV-2 infection [6]. Worldwide massive efforts are being made to either develop a vaccine or therapeutic based on S protein. The present study aims to identify B-cell and T-cell epitopes from the spike surface glycoprotein of SARS-CoV-2 and analyze their conservancy in all the available strains of SARS-CoV-2. For this, we have taken 8 unique strains of Spike glycoprotein with one each from India (Accession no QJF11884.1), China (QHD43416.1), France (QIX12148.1), USA (QID21058.1), Italy (QIC50498.1), Australia (QHR84449.1), Iran (QIX12195.1) and Pakistan (QIS60276.1), as these countries suffered initially with this dreaded disease. The strain data was extracted from NCBI Database and for the prediction of conserved sequences, multiple sequence alignments (MSA) using Clustal Omega [8] and protein variability analysis [10] was performed. From the multiple sequence alignments, spike surface glycoprotein was found to be highly conserved in 6 strains

Table 2
Table indicating selected 11B-cell epitopes were compared with these 29 surface accessible peptides and we found that 5 out of the selected 11B-cell epitopes were found to have consensus sequences with the 29 predicted surface accessible peptides. These five B-cell epitopes were KTPPIKDFGGF, QLPPAYTNSFTR, SGTNGTKRFDN, SNLKPFERDIST and ASVYAWNRK.

Sr. No	Sequence	Length	Start	End
1	KTPPIKDFGGF	11	790	801
2	QLPPAYTNSFTR	12	23	34
3	SGTNGTKRFDN	11	71	81
4	SNLKPFERDIST	12	459	470
5	ASVYAWNRK	9	348	357

Table 3
Table indicating that epitopes SGTNGTKRFDN and ASVYAWNRK which were predicted to be highly antigenic with values 0.5906 and 0.5788 respectively. The threshold for the antigenicity was set at 0.5 as default parameter. The other three epitopes KTPPIKDFGGF, QLPPAYTNSFTR and SNLKPFERDIST were anticipated to be non-antigens. The antigenicity prediction was analyzed using VaxiJen server.

Sr. No	Sequence	Overall Prediction for the Protective Antigen	Nature
1	KTPPIKDFGGF	−0.4958	NON-ANTIGEN
2	QLPPAYTNSFTR	−0.0689	NON-ANTIGEN
3	SGTNGTKRFDN	0.5906	ANTIGEN
4	SNLKPFERDIST	0.2978	NON-ANTIGEN
5	ASVYAWNRK	0.5788	ANTIGEN

Table 4
(A) Table indicating allergenicity of all the twelve selected T-cell epitopes using AllerTOP v.2.0. We found that out of 12 selected T-cell epitopes, 7 epitopes were found to be non-allergens and 5 were found to be allergens. (B) Table indicating toxicity of all seven non-allergen T-cell epitopes using ToxinPred server. We found that all the seven epitopes were non toxic in nature.

A

Sr. No.	Epitopes	Nature
1	WTAGAAAYY	NON-ALLERGEN
2	YLQPRTFLL	ALLERGEN
3	RLFRKSNLK	NON-ALLERGEN
4	NYNYLYRLF	NON-ALLERGEN
5	IPTNFTISV	NON-ALLERGEN
6	GRLQSLQTY	ALLERGEN
7	YRLFRKSNL	ALLERGEN
8	TRFQTLLAL	ALLERGEN
9	AEIRASANL	NON-ALLERGEN
10	HADQLTPTW	NON-ALLERGEN
11	TLLALHRSY	ALLERGEN
12	YSSANNCTF	NON-ALLERGEN

B

Epitope Sequence	Prediction
WTAGAAAYY	Non-Toxin
RLFRKSNLK	Non-Toxin
NYNYLYRLF	Non-Toxin
IPTNFTISV	Non-Toxin
AEIRASANL	Non-Toxin
HADQLTPTW	Non-Toxin
YSSANNCTF	Non-Toxin

from different countries except for certain changes observed in the strain of Australia and France. In the Australian strain, amino acid S-247 in the sequence of spike glycoprotein was replaced by R-247 and in the France strain, amino acid D-614 was replaced by G-613 [40] (Fig. 1A). To further validate such changes, Protein Variability Server [10] was used to determine the absolute variability. We discovered that there were 1271 highly conserved amino acids which comprise more than 98% of the length of spike surface glycoprotein. Three regions with amino acid positions 1–246, 248–613 and 615–1273 has sequence similarity in all the strains except for the regions with amino acid positions 247 and 614 (Fig. 1B). This proves the point that spike surface glycoprotein is conserved in most pathogenic coronavirus-2(SARS-CoV-2) strains and that the strain found in Australia and France is different from the strains found in other countries.

Next, after determining the conserved region, we wanted to explore if the spike surface glycoprotein of SARS-CoV-2 protein could be used as promising vaccine candidate. For this we accessed the antigenicity of the spike surface glycoprotein. On evaluation of surface spike glycoprotein sequence with accession number QHD43416.1 by VaxiJen server [11] we identified it as a probable antigen with a value of 0.4747. The threshold value for the antigenicity of virus was 0.4. To further validate the antigenicity of this spike surface glycoprotein, another antigenicity prediction tool by Kolaskar & Tongaonkar [12] was used. In this, a window size of 10 amino acids was set to determine the antigenicity of the central amino acid for each of residue of Spike surface glycoprotein. It discovered that most of the amino acid residues out of 1273 in the protein were above the threshold value of 1.00 with the minimum and maximum scores between 0.886 and 1.240 with an average

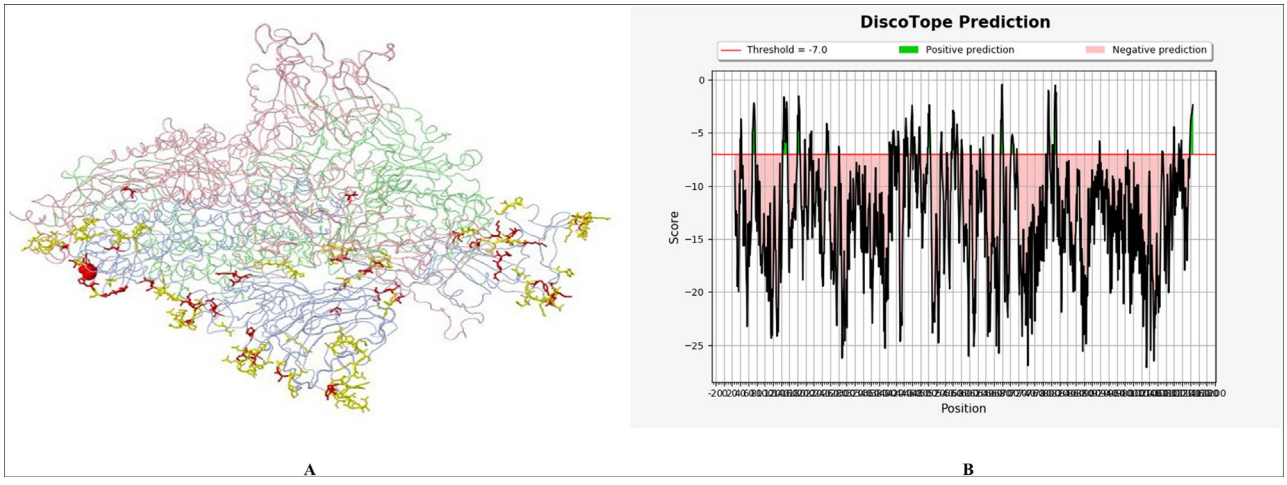

Fig. 3. Prediction of structural B cell epitopes from 3D protein structure, (A) 3D structure showing structural peptides which are shown in red mark (B) A graphical structure showing the score predicted by DiscoTope online tool, the epitopes are positive and above the threshold value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Table indicating world population coverage and geographical analysis of all the selected seven T-cell epitopes using IEDB Population Coverage analysis tool. Population coverage by most probable epitopes varied from (46.41 −67.16%) with both MHC class I and II alleles. Geographical area coverage ranged from 47.28% to 67.02%.

Epitope Sequence	World population coverage
WTAGAAAYY	17.34%
RLFRKSNLK	67.16%
NYNYLYRLF	32.15%
IPTNFTISV	51.84%
AEIRASANL	19.34%
HADQLTPTW	46.41%
YSSANNCTF	9.92%
Population/area	Coverage
Australia	47.28%
Europe	52.33%
Iran	50.03%
Northeast Asia	67.02%
South Asia	55.6%
United States	48.3%

score of 1.041 (Fig. 2A). These results unambiguously state that spike surface glycoprotein has strong antigenic capabilities to provoke enough immune response and hence can be used as promising vaccine candidate.

Next, our aim was to predict potential B-cell and T-cell epitopes which can be potentially used as vaccine candidates. For the prediction of potential linear B-cell epitopes, three different software packages namely BepiPred 2.0 IEBD [13], BepiPred-2.0: Sequential B-Cell Epitope Predictor [14] and ABCpred [15] were utilized. The number of peptides detected by ABCpred, BepiPred-2.0 and Bepipred Linear were 33, 24 and 24 respectively (Table S1). These identified epitopes were further comprehensively studied, and we found that there were 11 common B-cell antigenic epitopes of proteins predicted by all the three prediction tools (Table 1). The locations of these epitopes are 790–801, 23–34, 141–150, 829–838, 404–415, 370–380, 71–81, 374–385, 459–470, 379–390 and 348–357 (Table 1). Multiple studies have proved that an ideal epitope should be accessible to an antibody or a cell surface receptor [19]. After the determination of 11 potential B-cell antigenic epitopes, next we wanted to analyze the surface accessibility properties of these B-cell epitopes. At threshold cutoff 1.0, the surface accessibility of the spike glycoprotein was determined by Emini surface accessibility prediction tool [18]. Total 29 surface accessible peptides consisting of amino acids of varying length were found to have scores above the threshold value (Fig. 2B and Table S2). The selected 11B-cell epitopes were compared with these 29 surface accessible peptides and from the results we found that 5 out of the selected 11B-cell epitopes were found to have consensus sequences with the 29 predicted surface accessible peptides (Table 2). It is a well acknowledged fact that the use of conserved epitopes provides broader protection across multiple strains, or even species, than epitopes derived from highly variable genomic regions [41]. So, in an epitope-based vaccine development, an ideal epitope should be highly similar or conserved. Keeping this in mind, our next aim was to analyze conservancy of all the predicted 5B-cell surface accessible peptides in all the 8 strains from different countries. The conservancies of all the predicted B cell epitopes were evaluated by the IEDB conservancy analysis tool [20]. The results showed that these 5B-cell peptides showed 100% identity among all the strains sequences (Table S3). Next, these 5 highly conserved B-cell surface accessible peptides were scrutinized individually for the prediction of antigenicity. The antigenicity prediction was analyzed using VaxiJen server [11]. The threshold for the antigenicity was set at 0.5 as default parameter. We found that only 2 highly conserved B-cell surface accessible peptides were highly antigenic. The other three epitopes did not qualify as antigens and hence they were eliminated from the study (Table 3). Evidence suggests that apart from antigenicity, flexibility and accessibility are two other fundamental properties of an epitope which are essential to induce an immune response [21]. Subsequently, we analyzed the flexibility and hydrophilicity of these two highly antigenic epitopes. Flexibility was predicted using Karplus and Schulz flexibility prediction [22]. With a window size of 7 amino acids and center position as 4, both epitopes were found to have flexibility above the threshold of 0.9 (Fig. 2C, D and Table S4). Hydrophilicity of these two highly antigenic epitopes was assessed by IEDB Parker hydrophilicity analysis [23]. With a window size of 7 amino acids and centre position as 4, both epitopes were found to have hydrophilicity above the threshold of 0.40 (Fig. 2E, F and Table S5). These studies hence conclude that these 2 linear B-cell epitopes SGTNGTKRFDN and ASVYAWNRK are highly flexible and hydrophilic in nature. For structural epitopes prediction in the 3D protein structures DiscoTope: Structure-based Antibody Prediction tool was used [16]. A threshold value was set at −7.0 and the version1.1 was selected for analysis. From the result we identified one, 50 residue structural epitope which was above threshold value and showed good Disco-Tope score (Fig. 3A-B and Table S6). The positive predicted residues were considered for epitope based vaccine.

Subsequently, we identified potential T-cell epitopes using NetCTL server [24]. From the analysis, we found several T-cell epitopes but we shortlisted only those T-cell epitopes which had threshold value more than 3.0 from the different serotypes and based on this 30 T-cell epitopes were certified for further analysis (Table S7). Next, we examined the conserved sequences in all these thirty T-cell epitopes among all the 8 strains of spike protein from 8 different countries. High combinatorial scores predicted by multiple sequence alignments (MSA) suggested that all these 30 T-cell

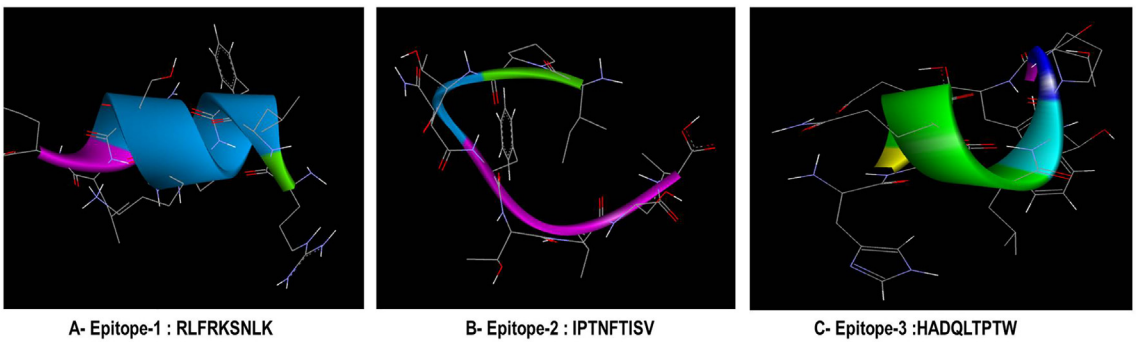

Fig. 4. Predicted 3-D structure of T-Cell epitopes RLFRKSNLK (A), IPTNFTISV (B) and HADQLTPTW (C). 3D structures of the selected peptides were constructed using PEP-FOLD Peptide Structure Prediction server.

epitopes were highly conserved in all the 8 different strains (Table S8). Afterwards, these selected 30 T-cell epitopes were made to bind with MHC class I alleles which was determined using MHC-I binding prediction server [26,42] based on IC50 cutoff values of 50 nM, peptide length 9 amino acids and results are shown in Table S9. For highly accurate results, the entire genomic sequence of spike surface glycoprotein was submitted in NetMHCII 2.3 Server for MHC class II alleles binding using MHC class II binding prediction tool [27]. As MHC class II can fit much longer peptides, epitopes with length 15 amino acids, IC50 cut off value of 50 nM and which fit into the binding grooves of MHC-II were selected (Table S10). From the results we found several T-cell epitopes that strongly bind with MHC class II alleles but in our study we shortlisted those T-cell epitopes which were commonly present in the binding grooves of both MHC I and MHC II. There were 12 overlapping T-cell epitopes which strongly bind with both MHC I and MHC II alleles (Table S11) and these were finally selected for allergenicity and toxicity assessment [28]. For the allergenicity prediction AllerTOP v. 2.0 [29] was used to predict the allergenicity of the 12 selected T-cell epitopes. From the results we found that out of 12 selected T-cell epitopes, 7 epitopes were found to be non-allergens and 5 were found to be allergens and hence these 5 T-cell epitopes were eliminated for further analysis (Table 4A). We also assessed the toxicity of these selected 7 T-cell epitopes using ToxinPred server [30]. The results showed that all the selected 7 T-cell epitopes were found to be non-toxic to cell proving their potential as candidate vaccines (Table 4B). Furthermore, studies have proved that for a peptide to turn out as good vaccine candidate, the predicted peptide(s) should effectively cover human population in wide areas [43]. Keeping this in mind these 7 selected T-cell epitopes were examined for their world population coverage and geographical distribution using IEDB Population Coverage analysis tool [31]. Population coverage by most probable epitopes varied from (46.41-67.16%) with both MHC class I and II alleles. Top three T-cell epitopes, with highest world population coverage were chosen for further analysis (Table 5). Next, our aim was to perform molecular docking between these three T cell epitopes peptides and MHC-I allele. Since HLA-C*12:03 (MHC-I allele) was found to interact with majority of the predicted T cell epitopes, therefore we decided to model 3D structure of these three T cell epitope peptides and HLA-C*12:03 allele. The population frequency of the HLA-C*12:03 alleles in various countries was analyzed using HLA Database (http://www.allelefrequencies.net/). The frequency found to be India is 0.0800, China 0.0730, France 0.0730, USA 0.0911, Italy 0.1500, Australia o.0510, Iran 0.0938 and Pakistan 0.1070.

3D structure of selected three T cell epitope peptides were built using PEPFOLD Peptide Structure Prediction server (Fig. 4A–C). 3-D structure of the HLA-C*12:03 allele was generated in the SWISS MODEL server by homology modeling [34] (Fig. 5A). The template

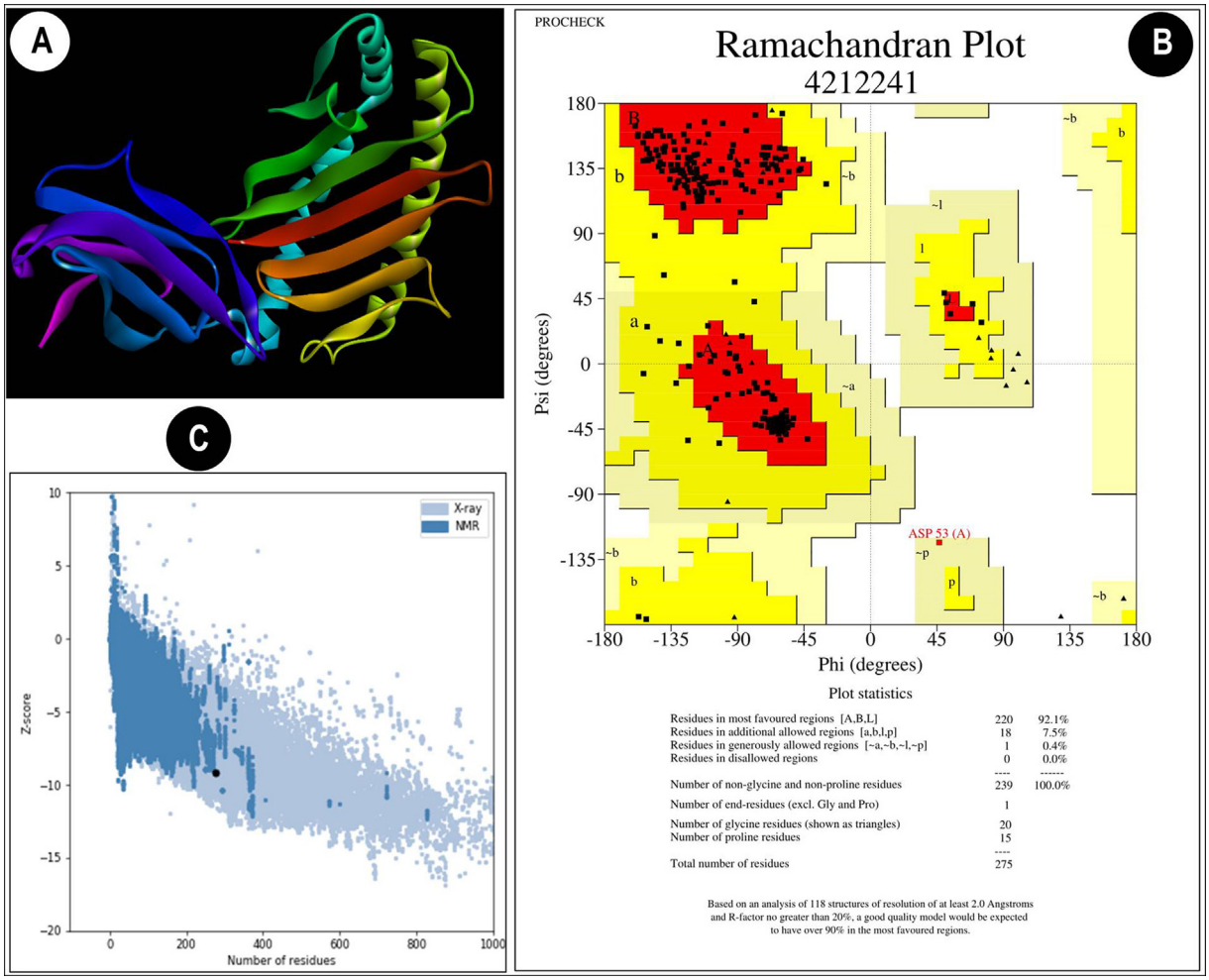

Fig. 5. (A) 3-D structure of the allele HLA-C*12:03 was generated in the SWISS MODEL server by 3D modeling. (B) Ramachandran plot of HLA-C*12:03 along with statistics displaying residues in the most favorable and disallowed regions (C) Z-score for quality of the 3D structure HLA-C*12:03, which provides an overall model quality, was −9.16 confirming good quality of the generated model.

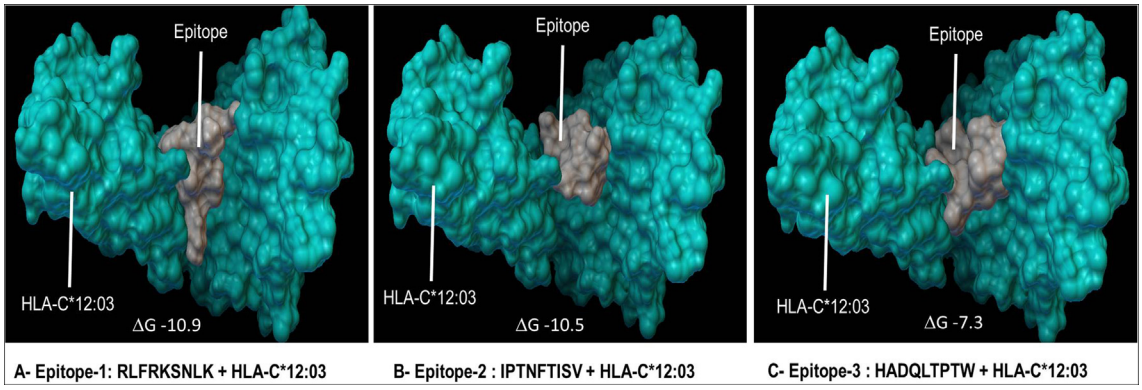

Fig. 6. The free binding energy of all the three conserved T-cell epitopes RLFRKSNLK, IPTNFTISV and HADQLTPTW bound with the HLA-C*12:03 were determined using AutoDOCK Vina tool in PyRx according to the equation: Free energy of binding = Intermol energy + Internal energy + Torsional energy-Unbound energy. The energy values calculated for binding of the epitopes RLFRKSNLK (A) IPTNFTISV (B) and HADQLTPTW (C) to the binding groove of the HLA-C*12:03 were ΔG = -10.9, −10.5 and −7.3 kcal/mol respectively.

model 5vgd.1.A was selected on the basis of GQME and QMEAN4 scores of 0.76 and 0.11 respectively and we found that there was 96.03% target-template sequence identity. The model was further validated using Ramachandran plot and Z-score. Ramachandran plot generated by PROCHECK software showed that 92.1% residues were in the favorable region (Fig. 5B). In addition the G- factor was normal (−0.04), Z- score determined using ProSAz-score [36,44] was −9.16. These findings evidently confirmed that the model we generated is a good quality model (Fig. 5 C). The free binding energy of all the three conserved T-cell epitopes RLFRKSNLK, IPTNFTISV and HADQLTPTW bound with the HLA-C*12:03 were determined using AutoDOCK Vina tool in PyRx [37] according to the equation: Free energy of binding = Intermol energy + Internal energy + Torsional energy-Unbound energy. In docking analysis, intermolecular forces include vander waals forces, hydrogen bonds, solvation and electrostatic energy. The energy values calculated for binding of the epitopes RLFRKSNLK, IPTNFTISV and HADQLTPTW to the binding groove of the HLA-C*12:03 were ΔG = −10.9, −10.5 and −7.3 kcal/mol respectively. The binding energy of the epitope RLFRKSNLK, IPTNFTISV and HADQLTPTW to the HLA-C*12:03 allele was then visualized using PyMol (Fig. 6A–C). From the docking scores, two epitopes RLFRKSNLK & IPTNFTISV showed lower binding free energy −10.9 kcal/mol and −10.5 kcal/-mol in contrast to the third epitope HADQLTPTW which had binding free energy −7.3 kcal/mol. Overall, these results confirmed that two T-cell epitopes RLFRKSNLK & IPTNFTISV have the best binding affinity with HLA-C*12:03 and hence shows promising approach as vaccine candidates. There are many similar studies which identified multiple specific regions in SARS-CoV-2 that have highly homology to SARS-CoV virus and also used for the prediction of potential B and T cell epitope for SARS-CoV-2 using bioinformatic approach [45]. Another study by Stephen et al predicted B cell epitopes using structure of viral glycoprotein which revealed that there are four epitopes which are located in the receptor binding domain of S protein [46]. Battacharya et al [47] and Feng et al [48] in their studies have also shown that spike glycoprotein can be used to obtain immunogenic epitope for the development of vaccine. Therefore, from the results we infer that our study can provide a direction towards an effective vaccine development against SARS-CoV-2

4. Conclusion

The present study discovered two highly conserved B-cell linear epitopes as well as one structural epitopes from the target protein, and also two T-cell epitopes which may be used as potential vaccine candidates for the prevention of SARS-CoV-2 infection. SGTNGTKRFDN, and ASVYAWNRK were predicted to be B-cell linear epitopes in addition to structural peptides and RLFRKSNLK and IPTNFTISV were predicted as T-cell epitope. These B-cell and T-cell epitopes which were unique to the spike glycoprotein were highly conserved (more than 98 percent), provide wider protection against multiple strains and are highly antigenic amongst all the 8 unique strains of spike glycoprotein taken from 8 different countries. Our results provide a potential pathway suggesting that these B-cell and T-cell epitopes can be used to develop epitope-based peptide vaccines against all pathogenic strains of SARS-CoV-2. Our study provides a clear blueprint for a path towards an efficacious vaccine against the scourge of SARS-CoV-2. However, these immunoinformatics analyses require several *in-vitro* and *in-vivo* validations before formulating the vaccine to resist SARS-CoV-2.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.vaccine.2021.01.011.

References

[1] Sohrabi C, Alsaf Z. World Health Organization declares global emergency: A review of the novel coronavirus (COVID-19). Int J Surg 2019;76(2020):71–6.
[2] Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9.
[3] Du L, He Y, Zhou Y, Liu S, Zheng B, Jiang S. The spike protein of SARS-CoV-a target for vaccine and therapeutic development. Nat Rev Microbiol 2009;7:226–36.
[4] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181:271–280.e8.
[5] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367:1260–3.
[6] Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV — a target for vaccine and therapeutic development. Nat Rev Microbiol 2009;7:226–36.
[7] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2009;37(Database):D26–31.
[8] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7. 539–539.

[9] Waterhouse A, Procter J, Martin D, Clamp M, Barton G. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009;25:1189–91.
[10] Garcia-Boronat M, Diez-Rivero C, Reinherz E, Reche P. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. Nucleic Acids Research 2008;36(Web Server):W35–41.
[11] Doytchinova I, Flower D. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinf 2007;8:4.
[12] Kolaskar A, Tongaonkar P. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett 1990;276:172–4.
[13] Larsen J, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res 2006;2(1):2.
[14] Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 2017;45(Web Server):W24–9.
[15] Saha S, Raghava G. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins Struct Funct Bioinf 2006;65(1):40–8.
[16] Andersen PH, Nielsen M, Lund O. Prediction of residues in discontinuous B cell epitopes using protein 3D structures. Protein Sci 2006;15:2558–67.
[17] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. The Immune Epitope Database 2.0. Nucleic Acids Research 2010;38 (Database):D854–62.
[18] 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.
[19] Caoili S. B-cell epitope prediction for peptide-based vaccine design. Proceedings of the first ACM international conference on bioinformatics and computational biology – BCB '10, 2010.
[20] Bui H, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. BMC Bioinf 2007;8(1):361.
[21] Novotny J, Handschumacher M, Haber E, Bruccoleri RE, Carlson WB, Fanning DW, et al. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc Natl Acad Sci 1986;83 (2):226–30.
[22] Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. Naturwissenschaften 1985;72(4):212–3.
[23] Parker JM, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: Correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. Biochemistry 1986;25(19):5425–32.
[24] Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, et al. NetCTL-1.0: An integrative approach to CTL epitope prediction. A combined algorithm integrating MHC-I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol 2005;35(8):2295–303.
[25] Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. NetCTL-1.2: Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinf 2007;8:424.
[26] Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinf 2005;6(1):132.
[27] Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. M Immunol 2018;154(3):394–406.
[28] Yasmin T, Akter S, Debnath M, Ebihara A, Nakagawa T, Nabi AH. In silico proposition to predict cluster of B- and T-cell epitopes for the usefulness of vaccine design from invasive, virulent and membrane associated proteins of C. jejuni. In Silico Pharmacol 2016;4(1):5.
[29] Dimitrov I, Flower DR, Doytchinova I. AllerTOP – a server for in silico prediction of allergens. BMC Bioinf 2013;14(Suppl 6).
[30] Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GP. In silico approach for predicting toxicity of peptides and proteins. PLoS ONE 2013;8(9).
[31] Bui H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinf 2006;7(1):153.
[32] Thevenet P, Shen Y, Maupetit J, Guyon F, Derreumaux P, Tuffery P. PEP-FOLD: An updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. Nucleic Acids Res 2012;40(W1):W288–93.
[33] Maupetit J, Derreumaux P, Tuffery P. PEP-FOLD: An online resource for de novo peptide structure prediction. Nucleic Acids Res 2009;37(Web Server).
[34] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 2014;42(W1).
[35] 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.
[36] Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 2007;35(Web Server).
[37] Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. Methods Mol Biol Chem Biol 2015;1263:243–50.
[38] Quiroga R, Villarreal MA. Vinardo: A scoring function based on Autodock Vina improves scoring, docking, and virtual screening. PLoS ONE 2016;11(5): e0155183.
[39] Seeliger D, Groot BL. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. J Comput Aided Mol Des 2010;24(5):417–22.
[40] Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al., Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2, bioRxiv2020.04.29.069054. Doi: https://doi.org/10.1101/2020.04.29.069054.
[41] Eickhoff CS, Terry FE, Peng LL, Meza KA, Sakala IG, Van Aartsen D, et al. Highly conserved influenza T cell epitopes induce broadly protective immunity. Vaccine 2019;36:5371–81.
[42] Peters B, Tong W, Sidney J, Sette A, Weng Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. Bioinformatics 2003;19(14):1765–72.
[43] Schubert B, Lund O, Nielsen M. Evaluation of peptide selection approaches for epitope-based vaccine design. Tissue Antigens 2013;82(4):243–51.
[44] Parvege MM, Rahman M, Nibir YM, Hossain MS. Two highly similar LAEDDTNAQKT and LTDKIGTEI epitopes in G glycoprotein may be useful for effective epitope based vaccine design against pathogenic Henipavirus. Comput Biol Chem 2016;61:270–80.
[45] 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;27:671–680.e2.
[46] Crooke SN, Ovsyannikova IG, Kennedy RB, Poland GA. Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome. https://doi.org/10.1101/2020.05.14.093757.
[47] Bhattacharya M, Sharma AR, Patra P, Ghosh P, Sharma G, Patra BC, et al., Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): Immunoinformatics approach. https://doi.org/10.1002/jmv.25736.
[48] Feng Y, Qiu M, Zou S, Li Y, Luo K, Chen R, et al., Multi-epitope vaccine design using an immunoinformatics approach for 2019 novel coronavirus in China (SARS-CoV-2). https://doi.org/10.1101/2020.03.03.962332.