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.

International Immunopharmacology 91 (2021) 107276

Contents lists available at ScienceDirect

International Immunopharmacology

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

Genome-wide analysis of Indian SARS-CoV-2 genomes to identify T-cell and B-cell epitopes from conserved regions based on immunogenicity and antigenicity

Nimisha Ghosh [a,1], Nikhil Sharma [b,1], Indrajit Saha [c,*,1], Sudipto Saha [d]

[a] Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Orissa, India
[b] Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India
[c] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India
[d] Division of Bioinformatics Bose Institute, Kolkata, West Bengal, India

ARTICLE INFO

Keywords:
B-cell epitopes
Conserved regions
SARS-CoV-2
Peptide based vaccine
Physico-chemical properties
T-cell epitopes

ABSTRACT

SARS-CoV-2 has a high transmission rate and shows frequent mutations, thus making vaccine development an arduous task. However, researchers around the globe are working hard to find a solution e.g. synthetic vaccine. Here, we have performed genome-wide analysis of 566 Indian SARS-CoV-2 genomes to extract the potential conserved regions for identifying peptide based synthetic vaccines, viz. epitopes with high immunogenicity and antigenicity. In this regard, different multiple sequence alignment techniques are used to align the SARS-CoV-2 genomes separately. Subsequently, consensus conserved regions are identified after finding the conserved regions from each aligned result of alignment techniques. Further, the consensus conserved regions are refined considering that their lengths are greater than or equal to 60nt and their corresponding proteins are devoid of any stop codons. Subsequently, their specificity as query coverage are verified using Nucleotide BLAST. Finally, with these consensus conserved regions, T-cell and B-cell epitopes are identified based on their immunogenic and antigenic scores which are then used to rank the conserved regions. As a result, we have ranked 23 consensus conserved regions that are associated with different proteins. This ranking also resulted in 34 MHC-I and 37 MHC-II restricted T-cell epitopes with 16 and 19 unique HLA alleles and 29 B-cell epitopes. After ranking, the consensus conserved region from NSP3 gene is obtained that is highly immunogenic and antigenic. In order to judge the relevance of the identified epitopes, the physico-chemical properties and binding conformation of the MHC-I and MHC-II restricted T-cell epitopes are shown with respect to HLA alleles.

1. Introduction

In December 2019, China reported a sudden outbreak of pneumonia due to an unknown source in Hubei province, Wuhan city [1] which later got attributed to a virus named SARS-CoV-2. SARS-CoV-2 belongs to the family of Coronaviridae which also houses SARS-CoV-1 [2,3] and MERS-CoV [4] virus. Genomic sequence analysis of the newly reported virus was found to be highly similar to that of SARS-CoV (95%–100%), thus showing the evolutionary similarity between SARS-CoV and SARS-CoV-2 [5]. By October 2020, India has registered over 7.65 million cases [6], making it one of the most affected countries in the world. Symptoms of the COVID-19 vary from fever, cough, myalgia, dyspnoea and diarrhoea to severe respiratory distress which may require life support systems. In severe cases, it may even lead to death [7]. Considering these consequences, World Health Organisation (WHO) suggested to interrupt human–human contact in the form of total lock downs along with pre-cautionary measures such as face masks and hand sanitizers to control the spread of COVID-19. Hence, it is the need of the hour to find a cure for COVID-19 in the form of vaccine.

Classical methods of vaccine design like attenuation of the virus through external sources such as micro-organisms to mitigate its harm or virulence usually depends on the response of the virus itself. Sometimes mutations in the virus genome can result in autoimmune response eventually making the virus even more virulent. Hence, such classic

* Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).
[1] Equally contributed.

https://doi.org/10.1016/j.intimp.2020.107276
Received 21 October 2020; Received in revised form 21 November 2020; Accepted 2 December 2020
Available online 16 December 2020

vaccine design approaches are time consuming, expensive and may not provide an effective response. With the evolution in bioinformatics and genome analysis, it is now possible to study the DNA, RNA and molecular evolution of a virus which can aid in development of vaccine through approaches such as reverse vaccinology. Reverse vaccinology involves pinpointing the protein sites that results into synthetic peptide based vaccines [8,9]. The preparation of epitope based vaccine is carried out in sequential form, starting from scanning the genome of the pathogen to locating the surface proteins, followed by extracting the best epitopes situated on the surface and also testing these synthetic designs against any autoimmune response [9]. The antigens provided by the epitopes are the sites to which antibodies bind, hence selection of the best epitopes is one of the crucial and foremost steps in vaccine design. In regard to this, Skwarczynski et al. [8] have suggested several factors which influence the selection of epitopes, such as immune response to the pathogen, hypersensitivity responses and coverage of different peptide against different pathogen subtypes. Further, these epitopes can be classified into two classes i.e. MHC-I, MHC-II associated T-cell epitopes [10] and B-cell epitopes [11] based on their responses against recognized foreign pathogens. The antigens provided by MHC-I interact directly with the CD8 cells evoking the cellular response [8]. MHC-II antigens bind to the surface of the pathogens to initiate the T-helper cells (CD4) which are responsible for activating the Th1 and Th2 type helper cells in the form of cytotoxic T-lymphocyte (CTL) and humoral response through antigens loaded in MHC-I and B-cell epitopes. Hence, the selection of T-cell and B-cell epitopes is a crucial process in order to provide a reliable vaccine.

By considering the several advantages presented in form of peptide-based vaccine, many studies have been carried out to design a vaccine in order to provide a stable solution against the threat as presented by SARS-CoV-2 virus. Earlier, it was found that spike (S) glycoprotein of SARS-CoV-2 can act as an intermediary to bind to the host cells with a very strong affinity, thus eventually attracting various experiments towards targeting this protein site as the potential target for vaccine design and diagnostics [12]. Following this, many types of vaccine designs have been proposed based on RNA, vectored, recombinant protein sequence and cell-cultures while focusing on the spike protein or whole virion [13]. Additionally, in Lin et al. [14] heptad repeats 1 and 2 (HR1 and HR2) in the spike protein have been predicted followed by the peptides with the help of molecular dynamics simulation between the fusion of the viral membrane and the host cell membrane, eventually limiting the spread of the virus within the host cells. Another study carried out by Vashi et al. [15] predicted 24 potential epitope fragments of which 20 were on the surface of spike protein. This information can be helpful for designing potential immunogenic peptide-based vaccines. Similar study has been conducted by Rakib et al. [16] in which spike protein region has been analysed through multiple sequence analysis in different SARS-CoV-2 genomes to predict the most immunogenic peptide fragments. In this study, a multi-epitope based vaccine has been proposed through analysing the S1 and S2 domains of spike proteins of the SARS-CoV-2 genomes in order to provide the best epitopes [17] for designing a vaccine. However, it is important to note that other protein sites can also be targeted for vaccine design as well [18]. This depends on how the T-cell interacts inside the different protein region of SARS-CoV-2. Grifoni et al. [18] have identified that 70–100% of epitope pools detect CD8 and CD4 T-cells for SARS-CoV-2. CD4+ cells interact with the other proteins like membrane (M), nucleocapsid (N) and ORF1ab proteins like NSP3, NSP4 and NSP12, but the dominance of CD4+ cells is very high within the spike region. On the other hand, no such dominant reactivity was identified in case of CD8+ cells in spike protein region. Hence, MHC-I restricted epitopes derived from M, NSP6, ORF3a or N proteins can also be considered for vaccine design. Noorimotlagh et al. [19] have conducted a review on several papers and have inferred a set of T-cell and B-cell epitopes from the Spike and Nucleocapsid proteins with high antigenicity. Genomic analysis conducted by Yadav et al. [20] on the first two cases reported in India resulted in the introduction of two non-identical strains of SARS-CoV-2. With time, more mutation points have been discovered [21] as well. This alteration in the protein region of the genome can lead to vaccine failures as was noticed in the case of Influenza virus in 2013–14 [22]. Hence, stable vaccine design is the need of the hour. Moreover, for such RNA viruses which undergo rapid mutations, Nandy et al. [9] have suggested the extraction of genomic regions which are either not influenced or very less influenced by the process of mutation. This can be carried out by analysing large set of virus genomes with the help of sequence alignment techniques. Such similar regions inside different viral genomes can be then considered for synthetic peptide vaccine designs. In [23], Gupta et al. have developed a web resource “CoronaVR” and have identified a set of T-cell and B-cell epitopes that can be incorporated in vaccine design. On the other hand, Crooke et al. [24] have used available algorithms and webtools to identify 41 T-cell epitopes (5 HLA class I, 36 HLA class II) and 6 B-cell epitopes as probable targets for epitope-based vaccine design. Ong et al. [25] have used Vaxign and the recently developed Vaxign-ML reverse vaccinology tools to predict potential vaccine candidates for COVID-19. Apart from Spike, they have identified epitopes derived from NSP3, 3CL-pro, NSP8, NSP9 and NSP10 proteins to be highly likely candidates for vaccine design. There are other works like [26–33] as well pertaining to epitope identification in SARS-CoV-2 for vaccine design.

In the above discussed literature, prediction of epitopes has been performed by analysing the virus proteins whereas genetical mutations are the primary reason for change in structure of the virus proteins. This fact motivated us to analyse the 566 available Indian SARS-CoV-2 genomes to identify the conserved regions to predict the immunogenic and antigenic epitopes. For this purpose, we have used four different multiple sequence alignment techniques viz. ClustalW [34], MUSCLE [35], ClustalO [36,37] and MAFFT [38] to align the sequences. Consensus conserved regions (CCnR) are then identified after finding the conserved regions from each aligned results of the alignment techniques. Further, these conserved regions are filtered on the basis of (a) length should be greater than or equal to 60nt and (b) corresponding protein sequence should not have any stop codons. This is followed by the validation of specificity of the conserved regions as query coverage with the help of Nucleotide BLAST [39]. These filtered conserved regions are then used to identify the T-cell and B-cell epitopes based on their immunogenic and antigenic scores. Thereafter, these scores are used to rank the conserved regions. As a result, we have obtained 23 conserved regions encompassing NSP1, NSP2, NSP3, NSP4, 3CL-Proteinase, NSP10, RNA-directed RNA polymerase, Helicase, Spike glycoprotein and Nucleocapsid protein. Subsequently, the consensus conserved region in NSP3 gene has been found to be highly immunogenic and antigenic. It provides MHC-I and MHC-II restricted T-cell epitopes and B-cell epitopes, FLKKDAPYI, ITFLKKDAPYIVGDV, TLVSDIDITFLKKDAP as immunogenic and TAVVIPTKK, IDITFLKKDAPYIVG, LHPDSATLVSDIDITF as antigenic respectively. Also, different immunogenic and antigenic epitopes associated to other conserved regions are provided as well. Finally, to validate the identified epitopes, the conformational 2D non-covalent structure of the chosen epitopes is studied. Moreover, the physico-chemical properties of the epitopes along with Ramachandran plot and Z-scores are also reported in the paper.

2. Materials and methods

In this section, at first the data preparation is elaborated followed by the discussion on the pipeline of the proposed work. For the benefit of the readers, brief discussions on epitope based vaccine, T-cell and B-cell epitopes and their prediction tools, physico-chemical properties of epitopes and docking of T-cell epitopes are given in the supplementary file. Moreover, prediction tools for T-cell and B-cell epitopes are reported in Supplementary Tables S1 and S2.

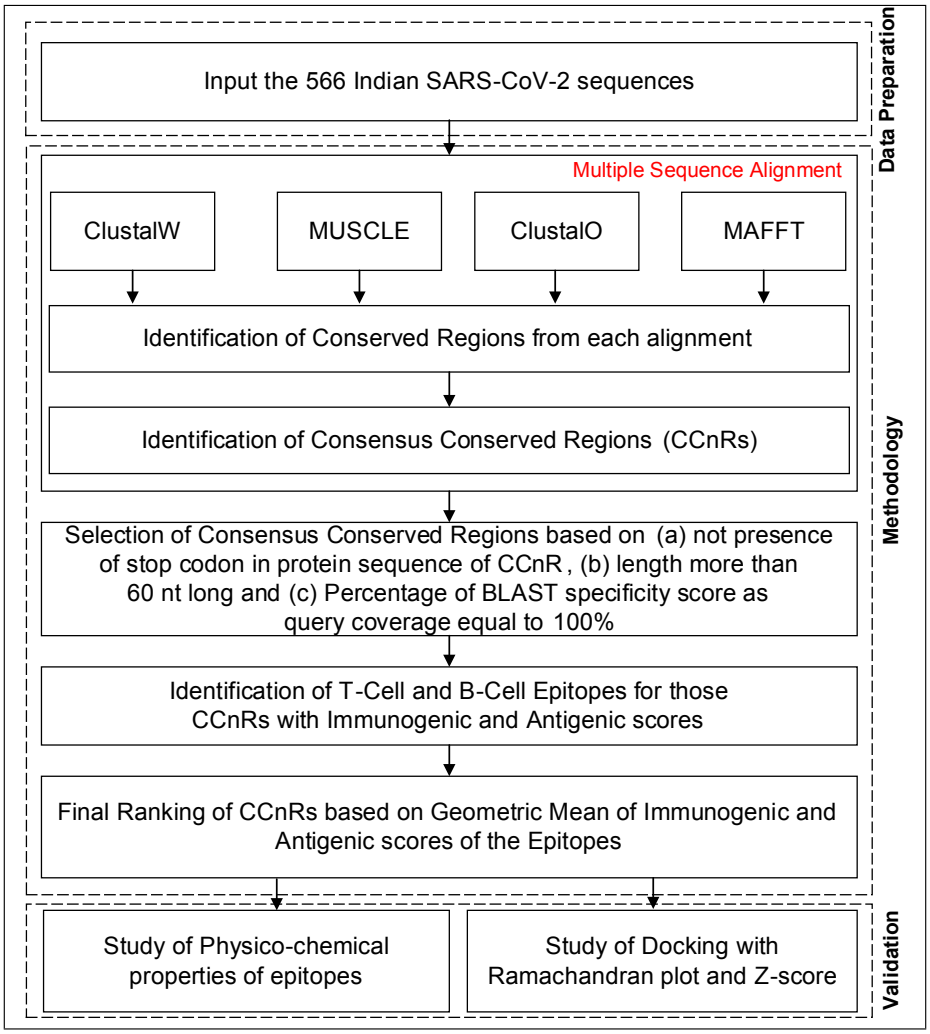

Fig. 1. Pipeline of the Workflow.

2.1. Data preparation

In order to map the SARS-CoV-2 proteins, we have used the reference SARS-CoV-2 genome (NC_045512.2)[2] and 44583 available protein sequences from the National Center for Biotechnology (NCBI). To generate the protein sequence, we have taken the reference sequence of SARS-CoV-2 genome and considered the reading frame concepts. A reading frame divides the sequence of nucleotides of the reference sequence into a set of successive, non-overlapping triplets. There are three possible reading frames: Frame 1 which starts from the first nucleotide of a reference sequence and creates the triplets, Frame 2 which starts from the second nucleotide and creates the triplets and Frame 3 which starts from the third nucleotide and creates the triplets. For each frame, these triplets are then translated into the corresponding proteins based on the codon table[3]. Finally, we have obtained 25 such unique proteins which were best matched to Frame 2. Also, the recent genomic sequences of Indian SARS-CoV-2 virus have been collected from Global Initiative on Sharing All Influenza Data (GISAID)[4] in fasta format. It contains 566 complete and near complete genomes with sequence ID. The average length of the 566 genomes is 29,831 bp. These 566 SARS-CoV-2 sequences are aligned using multiple sequencing alignment (MSA) techniques to extract the conserved regions. Also, the coded protein associated to each conserved region are extracted. For the alignment of sequences, High Performance Computing (HPC) facility of NITTTR, Kolkata is used. The HPC cluster has a master node with dual Intel Xeon Gold 6130 Processor having 32 Cores, 2.10 GHz, 22 MB L3 Cache and 128 GB DDR4 RAM and 2 GPU and 4 CPU computing nodes with dual Intel Xeon Gold 6152 Processor having 44 Cores, 2.1 GHz, 30 MB L3 Cache and 192 GB DDR4 RAM each, while GPU nodes have NVIDIA Tesla V100 GPU with 16 GB memory each. MSA was performed using the 2 GPU and 4 CPU computing nodes.

2.2. Pipeline of the workflow

The pipeline of the workflow is shown in Fig. 1. To start with, we have focused on finding the conserved regions in the 566 Indian SARS-CoV-2 genome sequence which are not affected by genetic mutations. For the same, initially we have constructed a Consensus Multiple Sequence Alignment (CMSA) approach in which we have used four different alignment techniques: ClustalW, MUSCLE, ClustalO and MAFFT in order to align the 566 SARS-CoV-2 sequences. Subsequently, consensus conserved regions (CCnR) are identified after finding the conserved regions from each aligned result of alignment techniques. ClustalW initially performs pairwise alignment of all sequences by using the k-tuple method. Thereafter, MSA is created by progressively aligning the most closely related sequences based on Neighbor-Joining guide tree method. In MUSCLE technique, two distance measures are used: k-mer

[2] https://www.ncbi.nlm.nih.gov/nuccore/1798174254.
[3] https://en.wikipedia.org/wiki/DNA_codon_table.
[4] https://www.gisaid.org/.

for unaligned pairs and Kimura method for aligned pairs of sequences. Initially, a draft MSA is produced in MUSCLE using the k-mer method. Then, a progressive alignment is constructed based on the guide tree as produced by the UPGMA method. This initial tree is then re-estimated using the Kimura distance method after which UPGMA method is once again used to produce a new guide tree, thereby creating a second MSA. New MSAs are finally created by realigning the two sequences created previously. ClustalO uses the k-tuple method to produce pairwise alignment. Then mBed is used to cluster the sequences followed by k-means clustering algorithm. Next, the guide tree is built using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method. Finally, MSA is constructed using the HHalign package. MAFFT uses two different heuristic methods, progressive (FFT-NS-2) and iterative refinement (FFT-NS-i). The main aim of MAFFT is to merge local and global algorithms for MSA. Initially, FFT-NS-2 is used to calculate all-pairwise distances to create a provisional MSA from which refined distances are calculated. Then, FFT-NS-i is performed to get the final MSA. Thereafter, to identify the conserved regions, these aligned sequences are used to compute the entropy(E).

$$E = \ln 5 + \sum S_x^y \ln[S_x^y] \quad (1)$$

where S_x^y indicates the frequency of each residue x occurring at position y and 5 represents the four possible residues as nucleotide plus gap. To identify the conserved regions (CnRs) for each alignment technique, a minimum segment length of 15 is considered with maximum average entropy as 0.2. Further, maximum entropy per position is taken as 0.2 with no gaps after finding the consensus sequence for the 566 genomic sequences. All these values are taken after following the literature. Thereafter, the CCnRs are identified considering the CnRs of all the alignment techniques. Next, a refinement process is carried out for the CCnRs based on the criteria that their length is greater then or equal to 60nt and no stop codon is present in the associated protein sequence. Moreover, Nucleotide BLAST is used to verify the specificity of the CCnRs as query coverage as well. Subsequently, T-cell and B-cell epitopes are identified from these CCnRs. To predict the T-cell and B-cell epitopes and to find their corresponding immunogenic scores, each CCnR is subjected to IEDB[5] and ABCPred[6] respectively. As recommended by IEDB, for the prediction of MHC-I and MHC-II T-cell epitopes, NetMHCpan[7] and Consensus Approach[8] [40] are selected respectively whereas for B-cell epitopes, prediction is carried out by ABCPred which uses Recurrent Neural Network. Then, by using the predicted epitopes, antigenic scores are calculated with the help of VaxiJen2.0[9]. For each CCnR, multiple T-cell and B-cell epitopes are identified along with their corresponding immunogenic and antigenic scores. Subsequently, for each CCnR the highest immunogenic and antigenic scores are considered to select the corresponding epitopes. Furthermore, these scores are used to rank the CCnRs based on geometric mean as given in Eq. (2). The use of geometric mean is to avoid the skewness of immunogenic and antigenic scores obtained for T-cell and B-cell epitopes so that proper ranking of the consensus conserved regions can be performed. Moreover, to validate the identified epitopes, the conformational 2D non-covalent structures of the identified epitopes are studied using LigPlot+ [41]. Furthermore, BepiPred2.0 server[10] [42] is used for the verification of the predicted B-cell epitopes.Also, the physico-chemical properties of the epitopes along with Ramachandran plot are reported through PyMOL [43] and its extensive libraries Autodock Vina (for docking) [44] and PyMOD 3 [45] while for the Z-score calculation ProSA[11] [46] online server is used.

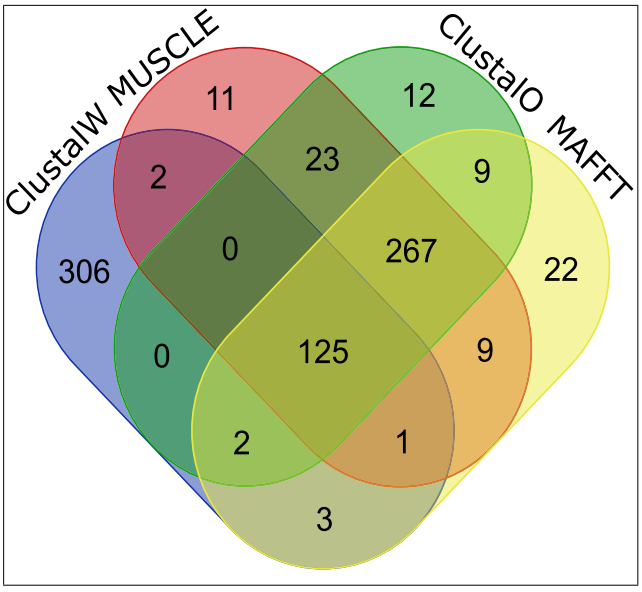

Fig. 2. 125 Consensus Conserved Regions (CCnRs) from the four alignment techniques.

$$R_{\mathrm{CCnR}} = Rank\left(\left[IS_{\mathrm{MHC-I}} \times IS_{\mathrm{MHC-II}} \times IS_{\mathrm{B-cell}}\right] \times \left[AS_{\mathrm{MHC-I}} \times AS_{\mathrm{MHC-II}} \times AS_{\mathrm{B-cell}}\right]\right)^{\frac{1}{6}}) \quad (2)$$

where, R_{CCnR} represents rank of consensus conserved region (CCnR) based on geometric mean of immunogenic and antigenic scores of T-cell and B-cell epitopes, IS_i and AS_i are the scaled immunogenic and antigenic scores for MHC-I, MHC-II and B-cell epitopes respectively.

3. Results and discussions

3.1. Ranking of the CCnRs

Experiments in this study are carried out according to the flowchart as mentioned in Fig. 1. Initially, 566 Indian SARS-CoV-2 genomes are aligned by using Consensus Multiple Sequence Alignment (CMSA) techniques, ClustalW, MUSCLE, ClustalO and MAFFT. Subsequently, we have obtained 125 CCnRs by considering all the alignment techniques. This is shown in Fig. 2 where 438, 439, 438 and 438 conserved regions (CnRs) from ClustalW, MUSCLE, ClustalO and MAFFT respectively are provided resulting in 125 CCnRs. This is followed by mapping of the CCnRs to 11 coding regions i.e. ORF1ab, Spike, ORF3a, Envelope, Membrane, ORF6, ORF7a, ORF7b, ORF8, Nucleocapsid and ORF10. The corresponding protein sequence for each CCnR has been taken from Frame 2. Now, the 125 CCnRs are filtered based on the criteria that (a) their length should be greater than or equal to 60nt and (b) no stop codons should be present in the corresponding proteins. A BLAST specificity score as query coverage equal to 100% is also considered during the filtering process. As a result, 23 CCnRs have been identified. Subsequently, these CCnRs are ranked on the basis of geometric mean of highly immunogenic and antigenic scores of the corresponding MHC-I, MHC-II T-cell and B-cell epitopes. It is worth mentioning that the immunogenic and antigenic scores are scaled within the range of 0–1 to bring the scores of all the epitopes for different CCnRs to a uniform scale and mentioned throughout the paper while the actual scores are given as Supplementary in excel file. After ranking, top 5 CCnRs along with their corresponding protein sequences, lengths, blast specificity scores,

[5] https://www.iedb.org/.
[6] https://webs.iiitd.edu.in/raghava/abcpred/ABC_submission.html.
[7] http://tools.iedb.org/mhci/.
[8] http://tools.iedb.org/mhcii/.
[9] http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html.
[10] http://tools.iedb.org/bcell/.
[11] https://prosa.services.came.sbg.ac.at/prosa.php.

Table 1
Top 5 Consensus Conserved Regions (CCnRs) as derived from SARS-CoV-2 with associated details.

Consensus Conserved Region (CCnR)	Protein Sequence of CCnR	Length of CCnR	BLAST Specificity Score of CCnR	% of BLAST Specificity Score as Query Coverage	Coding Region (CR)	Starting Coordinate of CR	Ending Coordinate of CR	Length of CR	Coded Protein from CR
4012-CACAGAAAACTTGTTACTTTATATTGACATTAATGGCAATCTTCATCCAGATTCTGCCACTCTTGTTAGTGACATTGACATCACTTTCTTAAAGAAAAGATGCTCCATATATAGTGGGTGATGTTGTTCAAGAGGGTGTTTTAACTGCTGTGGTTATACCTACTAAAAAGGCTGGTGGCACTACTGAAATGCTAGCGAAAGCTTT-4215	TENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKAGGTTEMLAKA	204	377	100	ORF1ab	266	21555	21290	NSP3
10463-TTAAGGGTTCATTCCTTAATGGTTCATGTGGTAGTGTTGGTTTTAACATAGATTATGACTGTGTCTCTTTTTGTTAC-10539	KGSFLNGSCGSVGFNIDYDCVSFCY	77	143	100	ORF1ab	266	21555	21290	3CL-Proteinase
13291-TTTTGTGACTTAAAAGGTAAGTATGTACAAATACCTACAACTTGTGCTAATGACCCTGTGGGTTTTACACTTAAAAACACAGTCTGTACCGTCTGCGGTAT-13391	FCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCG	101	187	100	ORF1ab	266	21555	21290	NSP10
5307-TAACACTCCAACAAATAGAGTTGAAGTTTAATCCACCTGCTCTACAAGATGCTTATTACAG-5367	TLQQIELKFNPPALQDAYY	61	113	100	ORF1ab	266	21555	21290	NSP3
9564-ATTCTTACCTGGTGTTTATTCTGTTATTTACTTGTACTTGACATTTTATCTTACTAATGATGTTTCTTTTTTAGCACATATTCAGTGGATGGTT-9657	FLPGVYSVIYLYLTFYLTNDVSFLAHIQWMV	94	174	100	ORF1ab	266	21555	21290	NSP4

percentage of BLAST specificity scores as query coverage, coding regions with their starting and ending coordinates, lengths and coded proteins are also mentioned in Table 1. Moreover, the ranking with the scores of these top 5 CCnRs is reported in Table 2. It is found from Table 1, that the top 5 CCnRs belong to the coding region which codes NSP3, 3CL-Proteinase, NSP10 and NSP4 proteins respectively. Please note that all the 23 CCnRs are reported in Supplementary Table S3 while their ranking details are given in Supplementary Table S4.

It is important to note that although structural proteins are the popular candidates for vaccine, vaccine protection can be correlated to non-structural proteins. In this regard, [47] showed that NS1 which is a non-structural protein can bring about protective immunity against

Table 2
Ranking procedure done on the basis of Geometric Mean of Binding and Antigenic Scores of T-cell and B-cell epitopes from each CCnR.

Consensus Conserved Region (CCnR)	Protein Sequence	Coded Protein	MHC-I restricted T-cell		MHC-II restricted T-cell		B-cell Epitopes		Final Score
			Immunogenic Score	Antigenic score	Immunogenic score	Antigenic Score	Immunogenic Score	Antigenic Score	
10463-CACAGAAAACTTGTTACTTTATATTGACATTAATGGCAATCTTCATCCAGATTCTGCCACTCTTGTTAGTGACATTGACATCACTTTCTTAAAGAAAAGATGCTCCATATATAGTGGGTGATGTTGTTCAAGAGGGTGTTTTAACTGCTGTGGTTATACCTACTAAAAAGGCTGGTGGCACTACTGAAATGCTAGCGAAAGCTTT-10539	TENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKAGGTTEMLAKA	NSP3	0.8640	0.7361	0.9804	0.6382	0.8810	1	0.84
9104-TTAAGGGTTCATTCCTTAATGGTTCATGTGGTAGTGTTGGTTTTAACATAGATTATGACTGTGTCTCTTTTTGTTAC-9211	KGSFLNGSCGSVGFNIDYDCVSFCY	3CL-Proteinase	0.6552	0.9049	0.9114	0.7499	0.7143	0.7401	0.77
21661-TTTTGTGACTTAAAAGGTAAGTATGTACAAATACCTACAACTTGTGCTAATGACCCTGTGGGTTTTACACTTAAAAACACAGTCTGTACCGTCTGCGGTAT-21728	FCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCG	NSP10	0.9136	0.7542	0.9818	0.3852	0.9048	0.6813	0.74
5220-TAACACTCCAACAAATAGAGTTGAAGTTTAATCCACCTGCTCTACAAGATGCTTATTACAG-5288	TLQQIELKFNPPALQDAYY	NSP3	0.8106	1	0.9485	0.6714	0.3333	0.8433	0.72
6706-ATTCTTACCTGGTGTTTATTCTGTTATTTACTTGTACTTGACATTTTATCTTACTAATGATGTTTCTTTTTTAGCACATATTCAGTGGATGGTT-6839	FLPGVYSVIYLYLTFYLTNDVSFLAHIQWMV	NSP4	0.9980	0.7866	0.9933	0.3326	0.9762	0.4726	0.70

Table 3
List of Immunogenic and Antigenic Epitopes for MHC-I, MHC-II restricted T-cell and B-cell Epitopes.

Protein Sequence	Coded Proteins	Type	MHC-I restricted T-cell				MHC-II restricted T-cell				B-cell		
			Epitope	Alleles	Scaled Score of		Epitope	Alleles	Scaled Score of		Epitope	Scaled Score of	
					Immunogenicity	Antigenicity			Immunogenicity	Antigenicity		Immunogenicity	Antigenicity
TENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKAGGTTEMLAKA	NSP3	Immunogenic	FLKKDAPYI	HLA-A*31:01	0.8640	0.3890	ITFLKKDAPYIVGDV	HLA-DRB3*01:01	0.9804	0.3036	TLVSDIDITFLKKDAP	0.8810	0.7314
KGSFLNGSCGSVGFNIDYDCVSFCY	3CL-Proteinase	Immunogenic	FLNGSCGSV	HLA-A*02:03	0.6552	0.3342	CGSVGFNIDYDCVSF	HLA-DQA1*01:01/DQB1*05:01	0.9114	0.7499	CGSVGFNIDYDCVSFC	0.7143	0.7401
		Antigenic											
FCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCG	NSP10	Immunogenic	DLKGKYVQI	HLA-B*08:01	0.9136	0.7542	KGKYVQIPTTCANDP	HLA-DRB1*04:01	0.9818	0.1892	TTCANDPVGFTLKNTV	0.9048	0.6813
TLQQIELKFNPPALQDAYY	NSP3	Immunogenic	NPPALQDAY	HLA-B*35:01	0.8106	0.4557	QIELKFNPPALQDAY	HLA-DRB3*02:02	0.9485	0.6409	LQQIELKFNPPALQDA	0.3333	0.8433
FLPGVYSVIYLYLTFYLTNDVSFLAHIQWMV	NSP4	Immunogenic	VSFLAHIQW	HLA-B*57:01	0.9980	0.7866	GVYSVIYLYLTFYLT	HLA-DPA1*01:03/DPB1*02:01	0.9933	0.3326	YSVIYLYLTFYLTNDV	0.9762	0.4726
		Antigenic											

flaviviruses. Though, no neutralizing effect was shown by antibodies against NS1, some exuded complement-fixing activity and even passive transfer of anti-NS1 antibody or immunization with NS1 can lead to protection against viruses [48]. Furthermore, anti-NS1 antibody could be responsible to block NS1-induced pathogenic effects, reduce viral replication by complement-dependent cytotoxicity of infected cells and even attenuate NS1-induced disease development. This has led to NS1 being a prospective vaccine candidate against Dengue virus [49,50]. Another core advantage of NS1 is that being a non-structural protein, the anti-NS1 antibody will not instigate antibody-dependent enhancement (ADE), which is a virulence factor causing serious repercussions. Additionally, non-structural virus proteins can generate cytotoxic T

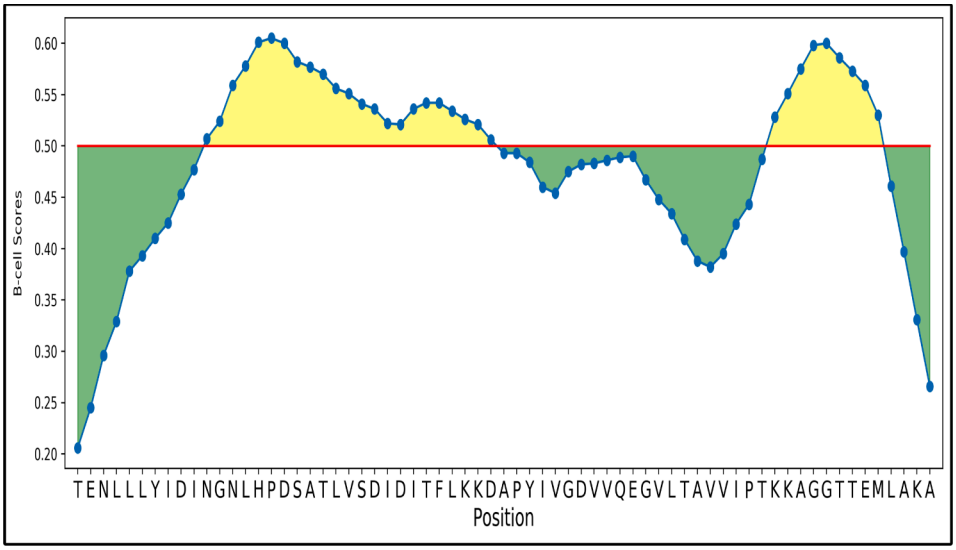

Fig. 3. Graphical representation of B-cell epitopes for TLVSDIDITFLKKDAP and LHPDSATLVSDIDITF with the threshold marked by red line.

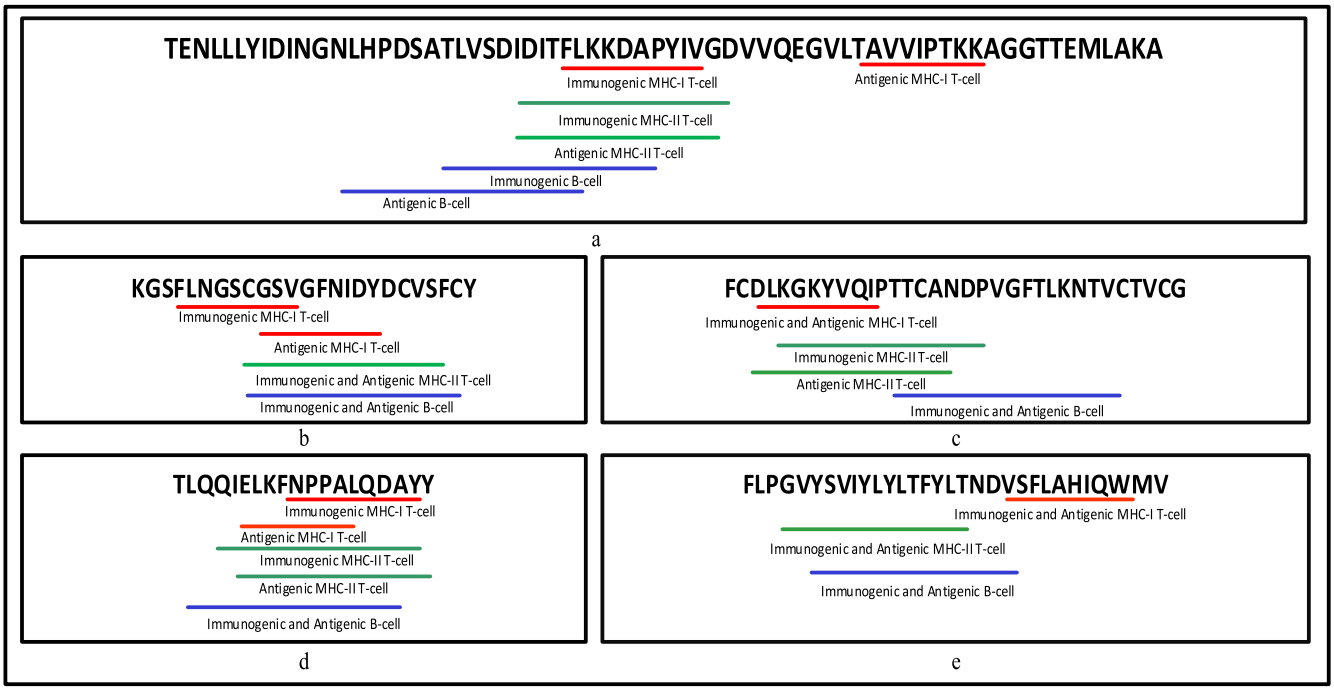

Fig. 4. MHC-I, MHC-II restricted T-cell and B-cell epitopes underlined in the protein sequences of top 5 CCnRs for (a) NSP3 (b) 3CL-Proteinase (c) NSP10 (d) NSP3 and (e) NSP4.

lymphocytes which are important to control infection. In [51], the authors have shown that the non-structural proteins of the hepatitis-C virus could generate HCV-specific broad-spectrum T-cell responses. Non-structural proteins have been used by [52] for vaccine design against Usutu Virus. Also, as targets for prophylactic or therapeutic vaccines, the non-structural proteins of HIV-1 were shown to be quite important [53]. Moreover, Ong et al. [25] have predicted NSP3 in SARS-CoV-2 to produce high protective antigenicity. Thus, we can hypothesize that apart from structural proteins non-structural proteins of SARS-CoV-2 can be possible targets as well for vaccine design which may induce cell-mediated or humoral immunity that is necessary to prevent viral invasion and/or replication.

3.2. Identification of MHC-I restricted T-cell epitopes

For epitope prediction from the 23 CCnRs, the associated protein sequences are used as inputs to the prediction tools. In this regard, MHC-I binding predictions are performed using IEDB [54] recommended NetMHCpan EL 4.1 (published recently in September 2020) targeting 27 unique HLA alleles. As a result, for each CCnR good binders in the form of immunogenic score, 4 best HLA epitopes are selected, in total 92 epitopes of length 9–11 mer each are obtained. Their antigenic scores are evaluated using VaxiJen2.0 [55]. In order to rank the CCnRs, only the best immunogenic and antigenic MHC-I restricted T-cell epitopes are considered. As a consequence, 34 such epitopes are identified and reported in Supplementary Table S5 for all the CCnRs while for the top 5 CCnRs, 8 epitopes are provided in Table 3. It is found that FLKKDAPYI and TAVVIPTKK are the highly immunogenic and antigenic MHC-I restricted T-cell epitopes from the NSP3 coded protein binded to HLA-A*31:01 and HLA-A*68:01 HLA alleles respectively. All the 92 MHC-I restricted T-cell epitopes along with their HLA alleles are provided in the supplementary as an excel file.

3.3. Identification of MHC-II restricted T-cell epitopes

Similar procedures are carried out for MHC-II restricted T-cell epitopes as well using MHC-II binding prediction tool provided by IEDB with consensus prediction targeting a different set of 27 unique HLA alleles. Subsequently, we obtained 92 epitopes of length 15–17 mer each which are bounded to their alleles along with their corresponding immunogenic and antigenic scores. In order to rank the CCnRs, the best immunogenic and antigenic MHC-II restricted T-cell epitopes are considered, resulting in 37 epitopes which are reported in Supplementary Table S5 for all the CCnRs. The 8 epitopes for the top 5 CCnRs are reported in Table 3. From this table, it is seen that ITFLKKDAPYIVGDV and IDITFLKKDAPYIVG are the most immunogenic and antigenic MHC-II restricted T-cell epitopes corresponding to HLA-DRB3*01:01 allele. All the 92 MHC-II restricted T-cell epitopes along with their HLA alleles are provided in the supplementary as an excel file.

3.4. Identification of B-cell epitopes

After obtaining MHC-I and MHC-II T-cell epitopes, B-cell epitopes which are responsible for antigen productions are predicted using ABCPred [56] with the length of 15–18 mer and their antigenic scores are evaluated from the VaxiJen server. As a result, 61 epitopes are found. In order to rank the CCnRs, the best immunogenic and antigenic B-cell epitopes are considered which resulted in 29 epitopes. These epitopes are reported in Supplementary Table S5 for all the CCnRs while for the top 5 CCnRs, 6 B-cell epitopes are reported in Table 3. In this table, it is found that TLVSDIDITFLKKDAP and LHPDSATLVSDIDITF are the most immunogenic and antigenic B-cell epitopes. Here, it should be noted that for antigenicity evaluation, a threshold of 0.4 is maintained throughout the experiment by following the literature [20]. The graphical representation of TLVSDIDITFLKKDAP and LHPDSATLVSDIDITF is shown in Fig. 3 using BepiPred 2.0 where the total green and yellow regions represent the protein sequence TENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKAGGTTEMLAKA while the two yellow regions denote the B-cell epitopes TLVSDIDITFLKKDAP and LHPDSATLVSDIDITF respectively. The red line in the figure represents the threshold which is set to 0.5. For all the 23 CCnRs the results are shown in Supplementary Fig. S1 while the 61 B-cell epitopes are provided in the supplementary as an excel file.

3.5. Final panel of epitopes

Table 4 summarises the final panel of the 34 MHC-I, 37 MHC-II restricted T-cell epitopes and 29 B-cell epitopes for 23 CCnRs based on their highest immunogenic and antigenic scores. There are 16 unique HLA alleles for MHC-I and 19 unique HLA alleles for MHC-II restricted T-cell epitopes. The associated coded proteins for the 23 CCnRs are NSP1, NSP2, NSP3, NSP4, 3CL-Proteinase, NSP10, RNA-directed RNA polymerase, Helicase, Spike glycoprotein and Nucleocapsid protein. For better readability, the epitopes associated with the top 5 CCnRs are underlined in Fig. 4 whereas the epitopes for 23 CCnRs are underlined in Supplementary Fig. S2. The red lines, green lines and the blue lines respectively denote the MHC-I, MHC-II T-cells and B-cells respectively. Moreover, for the ease of the readers, all the details related to the 125 CCnRs, 92 MHC-I and MHC-II restricted T-cell epitopes and 61 B-cell epitopes are provided in the supplementary as an excel file, the link of which is given in Table S6. Additionally, a list of MHC-I and MHC-II restricted T-cell and B-cell epitopes for SARS-CoV-2 as collected from different sources in the literature like [15–17, 20, 23–33] are reported in Table 5. For space constraint, 3 of each MHC-I and MHC-II restricted T-cell and B-cell epitopes from each paper are mentioned in this table while the list of all the MHC-I and MHC-II restricted T-cell and B-cell epitopes are given in the supplementary as an excel file as given in Table S6. Thus, Tables 4 and 5 can provide the readers a better insight

Table 4
Overview of MHC-I, MHC-II restricted T-cell and B-cell epitopes for the 23 CCnRs.

Coded Proteins	Type	MHC-I restricted T-cell		MHC-II restricted T-cell		B-cell Epitopes
		Epitopes	HLA Alleles	Epitopes	HLA Alleles	
NSP3	Immunogenic	FLKKDAPYI	HLA-A*31:01	ITFLKKDAPYIVGDV	HLA-DRB3*01:01	TLVSDIDITFLKKDAP
	Antigenic	TAVVIPTKK	HLA-A*68:01	IDITFLKKDAPYIVG	HLA-DRB3*01:01	LHPDSATLVSDIDITF
3CL-Proteinase	Immunogenic	FLNGSCGSV	HLA-A*02:03	CGSVGFNIDYDCVSF	HLA-DQA1*01:01/DQB1*05:01	CGSVGFNIDYDCVSFC
	Antigenic	GSVGFNIDY	HLA-A*30:02			
NSP10	Immunogenic	DLKGKYVQI	HLA-B*08:01	KGKYVQIPTTCANDP	HLA-DRB1*04:01	TTCANDPVGFTLKNTV
	Antigenic			DLKGKYVQIPTTCAN	HLA-DRB1*04:01	
NSP3	Immunogenic	NPPALQDAY	HLA-B*35:01	QIELKFNPPALQDAY	HLA-DRB3*02:02	LQQIELKFNPPALQDA
	Antigenic	IELKFNPPAL	HLA-B*40:01	IELKFNPPALQDAYY	HLA-DRB3*02:02	
NSP4	Immunogenic	VSFLAHIQW	HLA-B*57:01	GVYSVIYLYLTFYLT	HLA-DPA1*01:03/DPB1*02:01	YSVIYLYLTFYLTNDV
	Antigenic					
NSP3	Immunogenic	QVNGLTSIKW	HLA-B*57:01	PQVNGLTSIKWADNN	HLA-DQA1*01:02/DQB1*06:02	KYPQVNGLTSIKWADN
	Antigenic			KYPQVNGLTSIKWAD	HLA-DQA1*01:02/DQB1*06:02	
Helicase	Immunogenic	RAQNMTMSY	HLA-A*30:02	YQLKLLIHHRAQNMT	HLA-DRB4*01:01	FWDYQLKLLIHHRAQN
	Antigenic			DYQLKLLIHHRAQNM	HLA-DRB4*01:02	IHHRAQNMTMSYSLKP
Spike glycoprotein	Immunogenic	HADQLTPTW	HLA-B*58:01	DIPIGAGICASYQTQ	HLA-DQA1*05:01/DQB1*03:01	GCLIGAEHVNNSYECD
	Antigenic					
NSP4	Immunogenic	ICISTKHFYW	HLA-B*57:01	KHFYWFFSNYLKRRV	HLA-DPA1*01:03/DPB1*04:01	ISTKHFYWFFSNYLKR
	Antigenic			TKHFYWFFSNYLKRR	HLA-DPA1*01:03/DPB1*04:01	
Nucleocapsid protein	Immunogenic	AQFAPSASAF	HLA-B*15:01	ATKAYNVTQAFGRR	HLA-DRB5*01:01	KSAAEASKKPRQKRTA
	Antigenic			KAYNVTQAFGRRGP	HLA-DRB5*01:01	GRRGPEQTQGNFGDQE
Spike glycoprotein	Immunogenic	FERDISTEI	HLA-B*40:01	VEGFNCYFPLQSYGF	HLA-DQA1*01:01/DQB1*05:01	GSTPCNGVEGFNCYFP
	Antigenic	YFPLQSYGF	HLA-A*24:02	NGVEGFNCYFPLQSY	HLA-DRB3*01:01	EGFNCYFPLQSYGFQP
NSP4	Immunogenic	NVLEGSVAY	HLA-B*35:01	PVPYCYDTNVLEGSV	HLA-DRB1*04:01	SGKPVPYCYDTNVLEG
	Antigenic	SGKPVPYCY	HLA-A*30:02	GKPVPYCYDTNVLEG	HLA-DRB1*04:01	
Helicase	Immunogenic	VLAYVDHSY	HLA-B*15:01	VDHSYVVNAVTTMSY	HLA-DRB3*02:02	LAYVDHSYVVNAVTTM
	Antigenic					
NSP3	Immunogenic	NYMPYFFTL	HLA-A*24:02	CTNYMPYFFTLLLQL	HLA-DPA1*03:01/DPB1*04:02	VCTNYMPYFFTLLLQL
	Antigenic					
NSP10	Immunogenic	FAVDAAKAY	HLA-B*35:01	LSFCAFAVDAAKAYK	HLA-DRB3*01:01	GTGQAITVTPEANMDQ
	Antigenic	VPANSTVLSF	HLA-B*35:01			KMLCTHTGTGQAITVT
3CL-Proteinase	Immunogenic	GTTTLNGLW	HLA-B*57:01	TTTLNGLWLDDVVYC	HLA-DQA1*01:01/DQB1*05:01	QVTCGTTTLNGLWLDD
	Antigenic			TLNGLWLDDVVYCPR	HLA-DQA1*01:01/DQB1*05:01	
NSP1	Immunogenic	HVGEIPVAY	HLA-B*15:01	VAYRKVLLRKNGNKG	HLA-DRB1*11:01	PHVGEIPVAYRKVLLR
	Antigenic	HVGEIPVAYR	HLA-A*68:01	IPVAYRKVLLRKNGN	HLA-DRB1*11:01	
NSP4	Immunogenic	RPDTRYVLM	HLA-B*07:02	LMDGSIIQFPNTYLE	HLA-DRB1*15:01	GSIIQFPNTYLEGSVR
	Antigenic					LRPDTRYVLMDGSIIQ
NSP4	Immunogenic	VCVSTSGRW	HLA-B*57:01	TSGRWVLNNDYYRSL	HLA-DRB3*02:02	YCRHGTCERSEAGVCV
	Antigenic			STSGRWVLNNDYYRS	HLA-DRB3*02:02	
RNA-directed RNA polymerase	Immunogenic	DTLSLTTNMK	HLA-A*68:01	TTNMKKQFIIYLRIV	HLA-DPA1*02:01/DPB1*05:01	LRDTLSLTTNMKKQFI
	Antigenic	LSLTTNMKK	HLA-A*11:01			
NSP2	Immunogenic	VTHSKGLYR	HLA-A*31:01	ETFVTHSKGLYRKCV	HLA-DRB5*01:01	LNLGETFVTHSKGLYR
	Antigenic	VTHSKGLYRK	HLA-A*03:01	LGETFVTHSKGLYRK	HLA-DRB5*01:01	
Spike glycoprotein	Immunogenic	VYYPDKVFR	HLA-A*31:01	TRGVYYPDKVFRSSV	HLA-DRB1*03:01	RGVYYPDKVFRSSVLH
	Antigenic	GVYYPDKVFR	HLA-A*31:01			
NSP2	Immunogenic	LEQPTSEAV	HLA-B*40:01	GDLQPLEQPTSEAVE	HLA-DQA1*03:01/DQB1*03:02	TGDLQPLEQPTSEAVE
	Antigenic	EVVLKTGDL	HLA-A*26:01	EVVLKTGDLQPLEQP	HLA-DRB1*08:02	

into the epitopes identified so far.

3.6. Study of physico-chemical properties of epitopes

To judge the relevance of the epitopes as found in this work, we have evaluated the physico-chemical properties for each selected epitope. The values of each physico-chemical property lie between 0 and 1. Tables 6–8 show the physico-chemical properties for MHC-I, MHC-II restricted T-cell and B-cell epitopes respectively for the top 5 CCnRs whereas for all the 23 CCnRs, the results are reported in Supplementary Tables S7-S9 respectively. For example, in Table 6 MHC-I restricted T-cell epitope FLKKDAPYI has a positively charged value of 0.222, a negatively charged value of 0.111, polarity of 0.111, non-polarity of 0.556, alphaticity of 0.444, aromaticity of 0.222, acidicity of 0.111, Basicity of 0.222, hydrophobicity of 0.556, hydrophilicity of 0.333, a neutral value of 0.111, hydroxylic value of 0 and sulphur content is 0 as well. Similarly, for other epitopes their physico-chemical properties can be found in the tables.

3.7. Study of docking with Ramachandran plot and Z-score

To further validate the identified epitopes, the conformational 2D non-covalent structures of the identified MHC-I and MHC-II restricted T-cell epitopes are studied using LigPlot+. For the highly immunogenic and antigenic epitopes of each CCnR, molecular docking is computed using Autodock Vina in order to extract the stable binding conformation of each predicted epitope allele pair. For MHC-I restricted T-cell epitopes, 12 binding scores are generated from Autodock Vina while for MHC-II 9 binding scores are generated. For some epitopes, the docking structures are unable to generate due to the unavailability of the corresponding structure of the HLA alleles. Furthermore, Ramachandran plot and Z-score are also evaluated for further validation using PyMod 3

Table 5
List of proposed epitopes for SARS-CoV-2 as given in the literature.

Source	Coded Proteins	MHC-I restricted T-cell Epitopes	MHC-II restricted T-cell Epitopes	B-cell Epitopes
Bhattacharya et al. [26]	Spike glycoprotein	SQCVNLTTR YTNSFTRGV GVYYHKNNK	IHVSGTNGT VYYHKNNKS LVRDLPQGF	SQCVNLTTRTQLPPAYTNSFTRGVY FSNVTWFHAIHVSGTNGTKRFDN DPFLGVYYHKNNKSWME
Chen et al. [27]	Spike glycoprotein	LSPRWYFYY RSRNSSRNS IGYYRRATR	IKLDDKDPN RSGARSKQR RIGMEVTPS	EVRQIAPGQTGKIADY GCLIGAEHVNNSYECD FAMQMAYRFNGIGVTQ
Naz et al. [17]	Spike glycoprotein	GVYFASTEK STQDLFLPF KTSVDCTMY	EFVFKNIDGYFKIYS QPYRVVVLSFELLHA MTKTSVDCTMYICGD	YNSASFSTFKCYGVSPTKLNDLCFT
Kar et al. [28]	Spike glycoprotein	QIITTDNTF YQPYRVVVL FTISVTTEI	INITRFQTLLALHRS GINITRFQTLLALHR GWTFGAGAALQIPFA	FSYTESLAGKREMAII HAGPGPGPY KMGPGPGTRFA
Rakib et al. [16]	Spike glycoprotein	WTAGAAAYY CNDPFLGVY GAAAYYVGY	LIVNNATNV IVNNATNVV SKTQSLLIV	RTQLPPAYTNS SGTNGTKRFDN LTPGDSSSGWTAG
Vashi et al. [15]	Spike glycoprotein	RTQLPPAY RTQLPPA LPPAYTNSF	MFVFLVLLPLVSSQC MFVFLVLLPLVSSQCVN QGNFKNLREFVFKNI	PPAYTNSFTRGVYY HVSGTNGTKRFDN YYHKNNKSWMES
Yadav et al. [20]	Spike glycoprotein	GVYFASTEK FEYVSQPFL WTAGAAAYY	NA NA NA	HRSYLTPGDSSSGWTA FPNITNLCPFGEVFNA EVIQIAPGQTGKIADY
Crooke et al. [24]	Membrane glycoprotein	ATSRTLSYY RLFARTRSM YANRNRFLY	TLSYYKLGASQRVAG RTLSYYKLGASQRVA ASFRLFARTRSMWSF	EVTPSGTWL KLDDKDPNFK KTFPPTEPKKDKKKKADETQALPQ
Gupta et al. [23]	Spike glycoprotein	VRFPNITNL YQPYRVVVL PYRVVVLSF	NVTWFHAIHV	GDEVRQIAPGQTGKIADYNYKLP
Bhatnager et al. [29]	Spike glycoprotein	LTDEMIAQY LLTDEMIAQY IPFAMQMAY	VASQSIIAYTMSLGA LTDEMIAQYTSALLA VLNDILSRLDKVEAE	KEEQIGKCSTR ELGKYEQYGPGPGKWP IRAGPGPGGNC
Kwarteng et al. [30]	Nucleocapsid protein	KTFPPTEPK SSPDDQIGY SSPDDQIGYY	AQFAPSASAFFGMSR IAQFAPSASAFFGMS PQIAQFAPSASAFFG	AGLPYGANK SKQLQQSMSSADS RRIRGGDGKMKDL
Baruah et al. [31]	Spike glycoprotein	YLQPRTFLL GVYFASTEK EPVLKGVKL	NA	CVNLTTRTQLPPAYTN NVTWFHAIHVSGTNG SFSTFKCYGVSPTKLND
Bency et al. [32]	Spike glycoprotein	KIADYNYKL CYGVSPTKL VVVLSFELL	VVFLHVTYV IGINITRFQ FNCYFPLQS	MDLEGKQGNFKNL YYVGYLQPR NITNLCPFGE
Singh et al. [33]	Nucleocapsid protein	AQFAPSASA GDAALALLL GMSRIGMEV	AQFAPSASAFFGMSR GDAALALLLLDRLNQ ASAFFGMSRIGMEVT	KEDLKFP IKLDDKDPNFKDQ PPTEPKKDKKKKADETQALPQRQKKQQTVT
Ong et al. [25]	NSP3	STNVTIATY RMYIFFASF AEWFLAYIL	ISNSWLMWLIINLVQ LAYILFTRFFYVLGL AAIMQLFFSYFAVHF	EDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATS EEEQEEDWLDDD VGQQDGSEDNQ

Table 6
List of physico-chemical properties of MHC-I restricted T-cell epitopes.

MHC-I restricted T-cell epitopes	Positively charged	Negatively charged	Polarity	Non Polarity	Aliphaticity	Aromaticity	Acidicity	Basicity	Hydrophobicity	Hydrophilicity	Neutral	Hydroxylic	Sulphur Content
FLKKDAPYI	0.222	0.111	0.111	0.556	0.444	0.222	0.111	0.222	0.556	0.333	0.111	0	0
TAVVIPTKK	0.222	0	0.222	0.556	0.556	0	0	0.222	0.778	0.333	0.222	0.222	0
FLNGSCGSV	0	0	0.333	0.556	0.444	0.111	0	0	0.444	0.111	0.444	0.222	0.111
GSVGFNIDY	0	0.111	0.222	0.556	0.444	0.222	0.111	0	0.333	0.111	0.444	0.111	0
DLKGKYVQI	0.222	0.111	0.222	0.444	0.444	0.111	0.111	0.222	0.333	0.222	0.333	0	0
NPPALQDAY	0	0.111	0.222	0.556	0.556	0.111	0.111	0	0.556	0.333	0.222	0	0
IELKFNPPAL	0.1	0.1	0	0.7	0.6	0.1	0.1	0.1	0.7	0.4	0.1	0	0
VSFLAHIQW	0.111	0	0.222	0.667	0.444	0.222	0	0.111	0.667	0.111	0.222	0.111	0

Table 7
List of physico-chemical properties of MHC-II restricted T-cell epitopes.

MHC-II restricted T-cell epitopes	Positively charged	Negatively charged	Polarity	Non Polarity	Aliphaticity	Aromaticity	Acidicity	Basicity	Hydrophobicity	Hydrophilicity	Neutral	Hydroxylic	Sulphur Content
ITFLKKDAPYIVGDV	0.133	0.133	0.133	0.6	0.533	0.133	0.133	0.133	0.6	0.2	0.267	0.067	0
IDITFLKKDAPYIVG	0.133	0.133	0.133	0.6	0.533	0.133	0.133	0.133	0.6	0.2	0.267	0.067	0
CGSVGFNIDYDCVSF	0	0.133	0.333	0.467	0.333	0.2	0.133	0	0.467	0.067	0.4	0.133	0.133
KGKYVQIPTTCANDP	0.133	0.067	0.333	0.4	0.4	0.067	0.067	0.133	0.533	0.333	0.333	0.133	0.067
DLKGKYVQIPTTCAN	0.133	0.067	0.333	0.4	0.4	0.067	0.067	0.133	0.533	0.267	0.333	0.133	0.067
QIELKFNPPALQDAY	0.067	0.133	0.2	0.533	0.467	0.133	0.133	0.067	0.533	0.267	0.267	0	0
IELKFNPPALQDAYY	0.067	0.133	0.2	0.533	0.467	0.2	0.133	0.067	0.533	0.267	0.2	0	0
GVYSVIYLYLTFYLT	0	0	0.467	0.533	0.467	0.333	0	0	0.6	0	0.267	0.2	0

Table 8
List of physico-chemical properties of B-cell epitopes.

B-cell epitopes	Positively charged	Negatively charged	Polarity	Non Polarity	Aliphaticity	Aromaticity	Acidicity	Basicity	Hydrophobicity	Hydrophilicity	Neutral	Hydroxylic	Sulphur Content
TLVSDIDITFLKKDAP	0.125	0.188	0.188	0.500	0.438	0.062	0.188	0.125	0.625	0.188	0.375	0.188	0
LHPDSATLVSDIDITF	0.062	0.188	0.250	0.500	0.438	0.062	0.188	0.062	0.625	0.125	0.438	0.250	0
CGSVGFNIDYDCVSFC	0	0.125	0.375	0.438	0.312	0.188	0.125	0	0.500	0.062	0.375	0.125	0.188
TTCANDPVGFTLKNTV	0.062	0.062	0.312	0.438	0.375	0.062	0.062	0.062	0.688	0.250	0.375	0.250	0.062
LQQIELKFNPPALQDA	0.062	0.125	0.188	0.562	0.500	0.062	0.125	0.062	0.562	0.250	0.312	0	0
YSVIYLYLTFYLTNDV	0	0.062	0.438	0.438	0.375	0.312	0.062	0	0.562	0.062	0.25	0.188	0

Table 9
Docking and Z-scores of MHC-I and MHC-II restricted T-cell epitopes for the top 5 ranked CCnRs.

MHC-I restricted T-cell epitopes	Score from Autodock Vina	Z Score	MHC-II restricted T-cell epitopes	Score from Autodock Vina	Z Score
FLKKDAPYI	−8.2	−9.81	ITFLKKDAPYIVGDV	−9	−5.53
TAVVIPTKK	−8.1	−5.9	IDITFLKKDAPYIVG	−8.8	−5.59
FLNGSCGSV	Not Generated	Not Generated	CGSVGFNIDYDCVSF	Not Generated	Not Generated
GSVGFNIDY	−7.1	−5.4			
DLKGKYVQI	−8.1	−8.81	KGKYVQIPTTCANDP	Not Generated	Not Generated
			DLKGKYVQIPTTCAN	Not Generated	Not Generated
NPPALQDAY	Not Generated	Not Generated	QIELKFNPPALQDAY	Not Generated	Not Generated
IELKFNPPAL	Not Generated	Not Generated	IELKFNPPALQDAYY	Not Generated	Not Generated
VSFLAHIQW	−8.8	−9.26	GVYSVIYLYLTFYLT	−8	−5.02

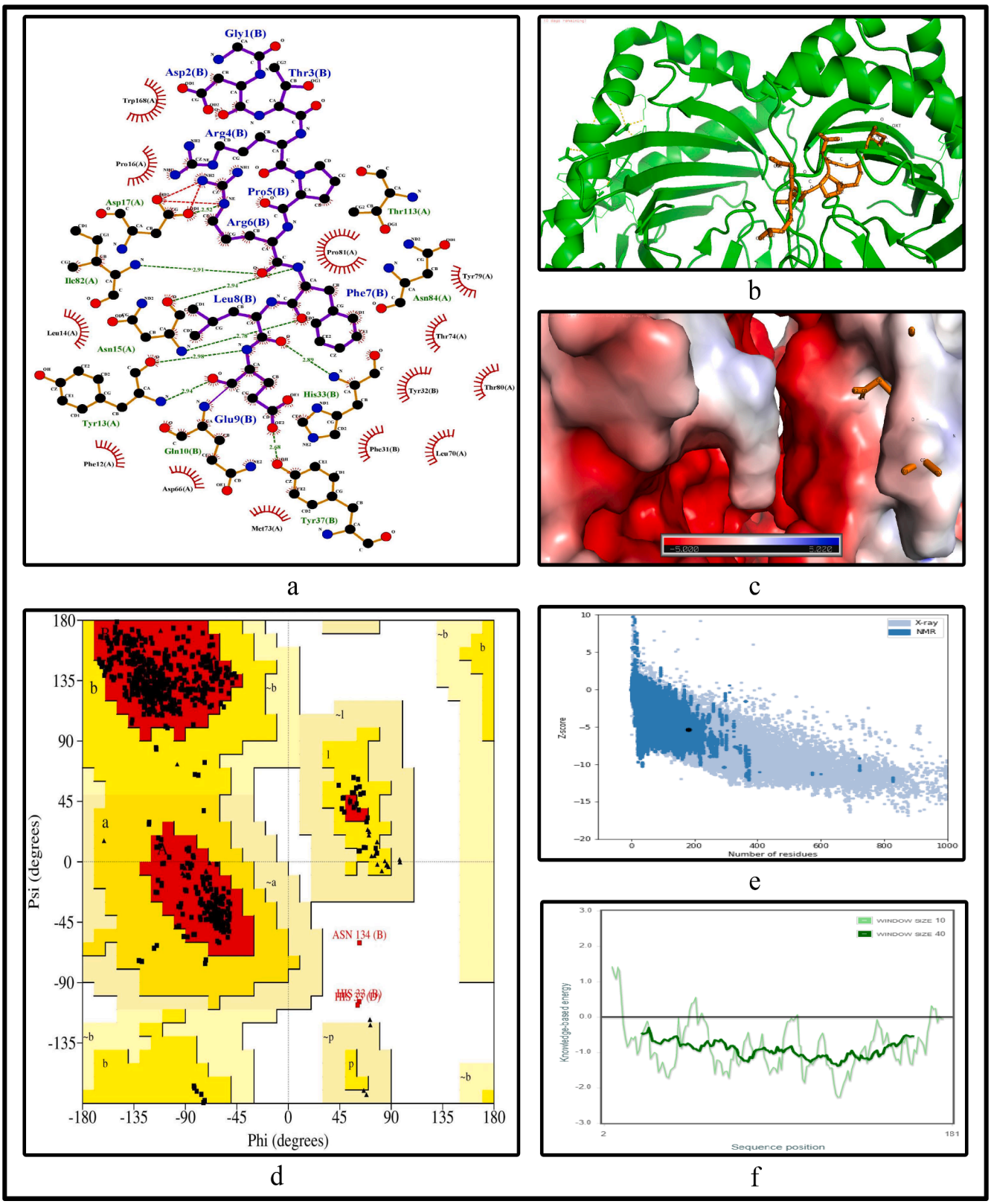

Fig. 5. Structural analysis for the highly immunogenic MHC-I restricted T-cell epitope "FLKKDAPYI" for NSP3 coded protein (a) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (b) Docking structure of MHC-I restricted T-cell epitope (c) The surface interaction between the allele and epitopes showing the fitting sites in binding grooves (d) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frame (e) Z-score plot and (f) all residue energy.

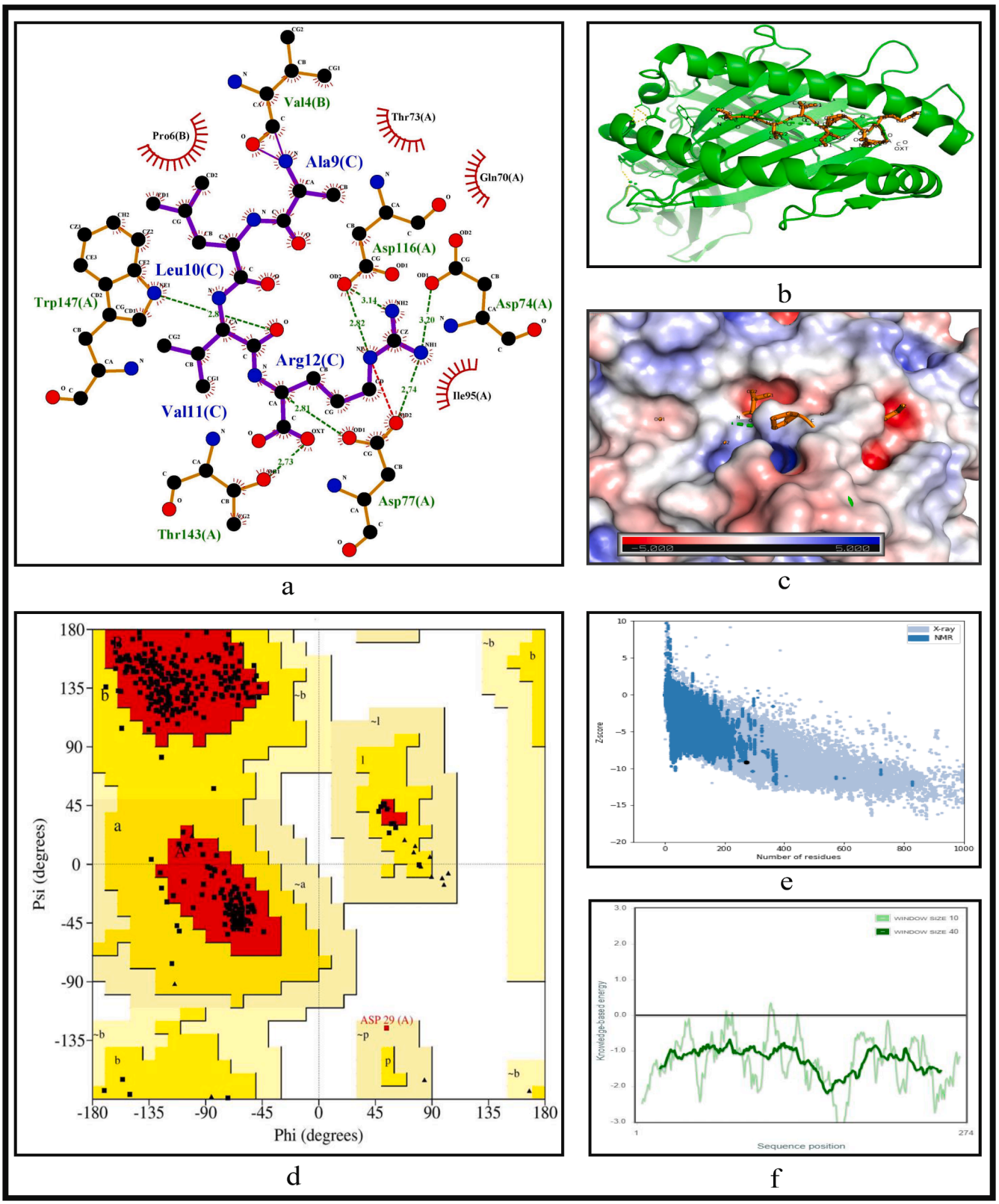

Fig. 6. Structural analysis for the highly antigenic MHC-I restricted T-cell epitope "TAVVIPTKK" for NSP3 coded protein (a) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (b) Docking structure of MHC-I restricted T-cell epitope (c) The surface interaction between the allele and epitopes showing the fitting sites in binding grooves (d) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frame (e) Z-score plot and (f) all residue energy.

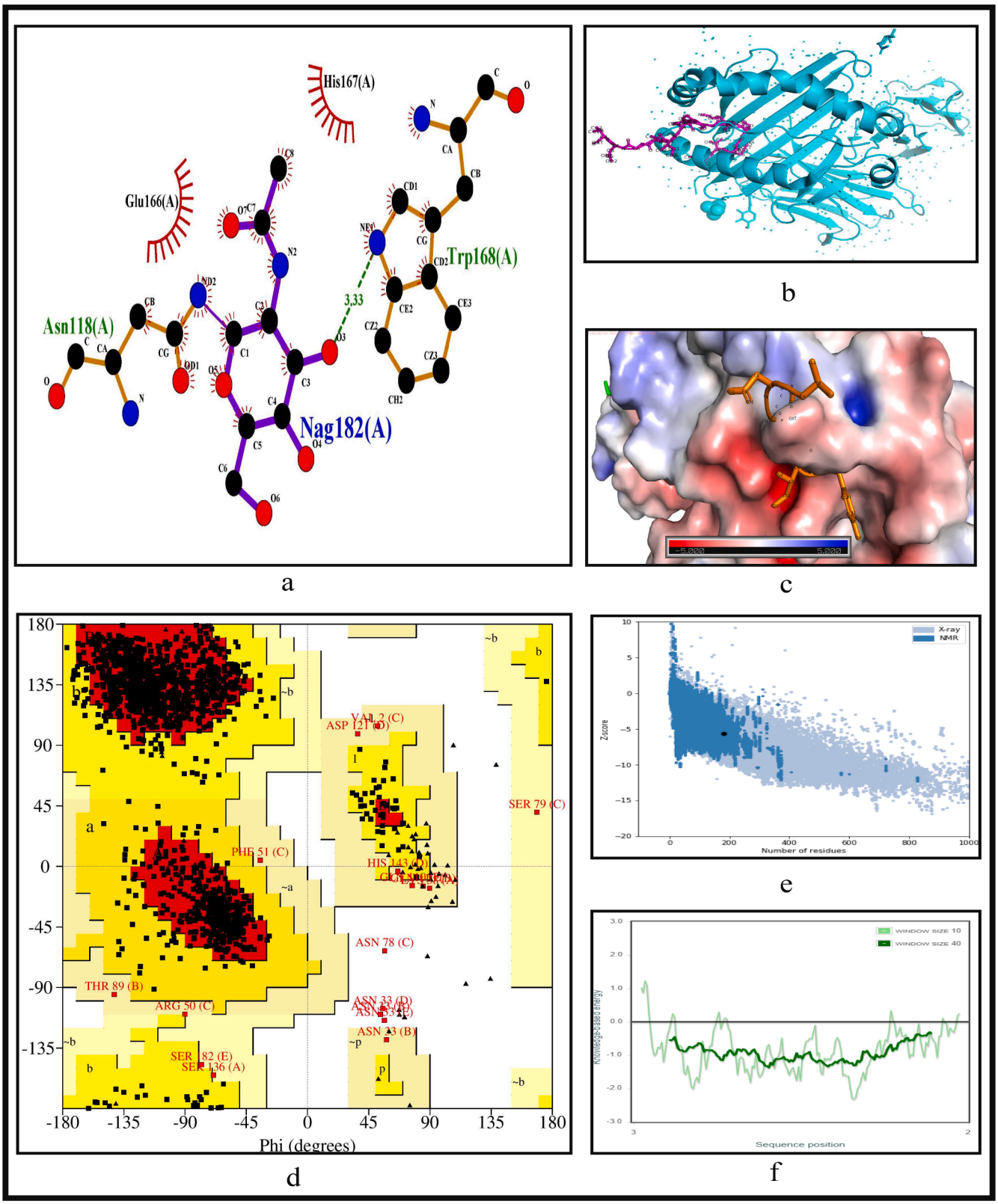

Fig. 7. Structural analysis for the highly immunogenic MHC-II restricted T-cell epitope "ITFLKKDAPYIVGDV" for NSP3 coded protein (a) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (b) Docking structure of MHC-II restricted T-cell epitope (c) The surface interaction between the allele and epitopes showing the fitting sites in binding grooves (d) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frame (e) Z-score plot and (f) all residue energy.

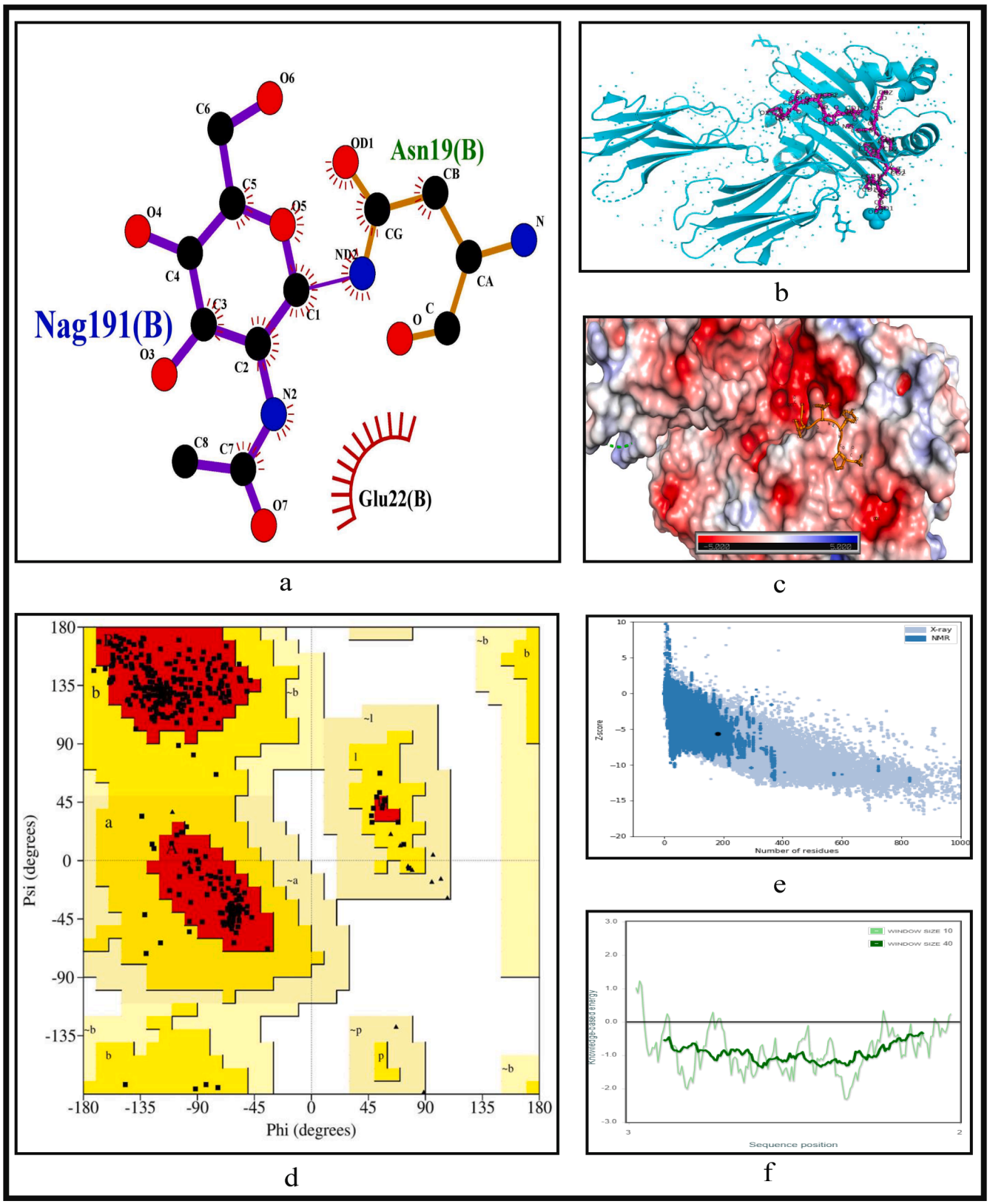

Fig. 8. Structural analysis for the highly antigenic MHC-II restricted T-cell epitope "IDITFLKKDAPYIVG" for NSP3 coded protein (a) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (b) Docking structure of MHC-II restricted T-cell epitope (c) The surface interaction between the allele and epitopes showing the fitting sites in binding grooves (d) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frame (e) Z-score plot and (f) all residue energy.

and ProSA server respectively. The results of docking along with Z-scores are reported in Table 9. The results for FLKKDAPYI and TAVVIPTKK which are the most highly immunogenic and antigenic MHC-I restricted T-cell epitopes are shown in Figs. 5 and 6 while ITFLKKDAPYIVGDV and IDITFLKKDAPYIVG which are the most highly immunogenic and antigenic MHC-II restricted T-cell epitopes are shown in Figs. 7 and 8 respectively. In these four figures, (a) shows the binding pose of the molecules of the two epitopes, (b) shows the exact binding position of the epitopes in the binding grooves of the alleles obtained from Autodock Vina with docking scores of −8.2 and −8.1 for MHC-I and −9 and −8.8 for MHC-II for both immunogenic and antigenic epitopes respectively and (c) depicts the surface interaction between the alleles and the identified epitopes showing the fitting sites in binding grooves. Further, quality of the residues inside the epitopes are evaluated on the basis of rotational spin of the atoms around bonds. This is depicted in (d) of Figs. 5 and 6 for MHC-I and Figs. 7 and 8 for MHC-II through Ramachandran plot in which points lying in the red region represents much more stable state of their bond orientations inside a molecule. This is followed by the Z-Score evaluation in (e) where the negative values of Z-score which are −9.81 and −5.9 for MHC-I and −5.53 and −5.59 for MHC-II as shown in Table 9 and Figs. 5–8 verify the stability of the structures and (f) shows the overall negative energy values of the entire residues inside the whole structures which confirm the molecular stability of the identified epitopes. The results for docking along with Z-scores for all the 23 CCnRs are reported in Supplementary Table S10 while the corresponding structural analysis are given in Supplementary Figs. S3 and S4.

Due to the worldwide pandemic caused by SARS-CoV-2, development of safe and effective vaccines is the need of the hour. This study has identified T-cell and B-cell epitopes using computational methods which can be used for probable vaccine design. The main advantages of this work can be summarised as (a) whole genome analysis of 566 Indian SARS-CoV-2 genomes in order to consider the genetic mutations to understand and target the virus proteins, (b) finding consensus conserved regions from four alignment techniques viz. ClustalW, MUSCLE, ClustalO and MAFFT and (c) using latest tools like NetMHCpan EL 4.1 (published in September 2020), PyMod 3 and BepiPred 2.0 for computational purposes. Furthermore, we have used our own developed tool ABCpred to predict the B-cell epitopes.

4. Conclusion

In this work, genome-wide analysis of 566 Indian SARS-CoV-2 genomes have been performed to extract the potential conserved regions for epitope-based synthetic vaccine design which show high immunogenicity and antigenicity. In this regard, 125 CCnRs have been identified after extracting the conserved regions from the aligned sequences of the four multiple sequence alignment techniques. These CCnRs are then filtered based on three major criteria of length greater than or equal to 60nt, no stop codons in the proteins and percentage of BLAST specificity score as query coverage equal to 100%. Such filtering resulted in 23 CCnRs covering NSP1, NSP2, NSP3, NSP4, 3CL-Proteinase, NSP10, RNA-directed RNA polymerase, Helicase, Spike glycoprotein and Nucleocapsid protein. This ranking also resulted in 34 MHC-I and 37 MHC-II restricted T-cell epitopes with 16 and 19 unique HLA alleles and 29 B-cell epitopes for the 23 CCnRs. These CCnRs are then ranked based on their immunogenic and antigenic scores to identify the MHC-I and MHC-II restricted T-cell and B-cell epitopes. This ranking identified CCnR from NSP3 coded protein to be highly immunogenic and antigenic, providing MHC-I and MHC-II restricted T-cell and B-cell epitopes, FLKKDAPYI, ITFLKKDAPYIVGDV, TLVSDIDITFLKKDAP as most immunogenic and TAVVIPTKK, IDITFLKKDAPYIVG, LHPDSATLVSDIDITF as most antigenic respectively. These epitopes can be considered for designing of synthetic vaccines. Furthermore, to validate the relevance of these epitopes, their binding confirmation and physico-chemical properties are also shown with respect to HLA alleles. This study thus provides the potential MHC-I and MHC-II restricted T-cell and B-cell epitopes to design epitope-based synthetic vaccines.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Availability of data and materials

The aligned 566 Indian SARS-CoV-2 genomes with reference as well as consensus sequences and the final results of this work are available at "http://www.nitttrkol.ac.in/indrajit/projects/COVID-EpitopeVaccine-India/". Moreover, Indian SARS-CoV-2 genomes used in this work are publicly available at GISAID database.

Consent for publication

Not applicable.

Funding

This work has been partially supported by CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India.

Author contributions

Nimisha Ghosh: Formal analysis; Methodology, Coding; Visualization; Writing - original draft & editing, **Nikhil Sharma**: Methodology; Coding; Visualization; Writing - review & editing, **Indrajit Saha**: Conceptualization; Data curation; Supervision; Funding acquisition; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing - review & editing, **Sudipto Saha**: Conceptualization; Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgement

We thank all those who have contributed sequences to GISAID database.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.intimp.2020.107276.

References

[1] P. Zhou, X.L. Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, H.R. Si, Y. Zhu, B. Li, C. Huang, H. Chen, J. Chen, Y. Luo, H. Guo, R. Jiang, M. Liu, Y. Chen, X. Shen, X. Wang, X. Zheng, K. Zhao, Q. Chen, L.L.F. Deng, B. Yan, F. Zhan, Y. Wang, G. Xiao, Z. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273, https://doi.org/10.1038/s41586-020-2012-7.

[2] T. Ksiazek, D. Erdman, C. Goldsmith, S. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J. Comer, W. Lim, P. Rollin, S. Dowell, A. Ling, C. Humphrey, W.J. Shieh, J. Guarner, C. Paddock, P. Rota, B. Fields, L. Anderson, A novel coronavirus associated with severe acute respiratory syndrome, New Engl. J. Med. 348 (2003) 1953–1966, https://doi.org/10.1056/NEJMoa030781.

[3] K. Holmes, L. Enjuanes, The sars coronavirus: A postgenomic era, Science (New York, N.Y.) 300 (2003) 1377–1378, https://doi.org/10.1126/science.1086418.

[4] R.D. Groot, S. Baker, R. Baric, C. Brown, C. Drosten, L. Enjuanes, R. Fouchier, M. Galiano, A. Gorbalenya, Z. Memish, S. Perlman, L. Poon, E. Snijder, G. Stephens, P. Woo, A. Zaki, M. Zambon, J. Ziebuhr, Middle east respiratory syndrome coronavirus (mers-cov): Announcement of the coronavirus study group, J. Virol. 87 (2013), https://doi.org/10.1128/JVI.01244-13.

[5] J. Xu, S. Zhao, T. Teng, A. Abdalla, W. Zhu, L. Xie, Y. Wang, X. Guo, Systematic comparison of two animal-to-human transmitted human coronaviruses: Sars-cov-2 and sars-cov, Viruses 12 (2020) 244, https://doi.org/10.3390/v12020244.

[6] Worldometer, Coronavirus disease 2019 (covid-19) cases in india, https://www.worldometers.info/coronavirus/country/india/, accessed: 2020-10-21 (2020).

[7] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, Lancet 395 (2020), https://doi.org/10.1016/S0140-6736(20)30183-5.

[8] M. Skwarczynski, I. Toth, Peptide-based synthetic vaccines, Chem. Sci. 7 (2015), https://doi.org/10.1039/C5SC03892H.

[9] A. Nandy, S. Basak, Bioinformatics in Design of Antiviral Vaccines, Encycl. Biomed. Eng. (2018), https://doi.org/10.1016/B978-0-12-801238-3.10878-5.

[10] A. Patronov, I. Doytchinova, T-cell epitope vaccine design by immunoinformatics, Open Biol. 3 (2013) 120139, https://doi.org/10.1098/rsob.120139.

[11] T. Ahmad, A. Ewida, S. Sheweita, B-cell epitope mapping for the design of vaccines and effective diagnostics, Trials Vaccinol. 5 (2016) 71–83, https://doi.org/10.1016/j.trivac.2016.04.003.

[12] D. Wrapp, N. Wang, K. Corbett, J. Goldsmith, C. Hsieh, O. Abiona, B. Graham, J. Mclellan, Cryo-em structure of the 2019-ncov spike in the prefusion conformation, bioRxiv: the preprint server for biology (2020). doi:10.1101/2020.02.11.944462.

[13] F. Amanat, F. Krammer, Sars-cov-2 vaccines: Status report, Immunity 52 (2020), https://doi.org/10.1016/j.immuni.2020.03.007.

[14] R. Ling, Y. Dai, B. Huang, W. Huang, X. Lu, Y. Jiang, In silico design of antiviral peptides targeting the spike protein of sars-cov-2, Peptides 130 (2020) 170328, https://doi.org/10.1016/j.peptides.2020.170328.

[15] Y. Vashi, V. Jagrit, S. Kumar, Understanding the b and t cells epitopes of spike protein of severe respiratory syndrome coronavirus-2: A computational way to predict the immunogens, Infect. Genet. Evol. 84 (2020) 104382, https://doi.org/10.1016/j.meegid.2020.104382.

[16] A. Rakib, A. Saad, S. Sami, N.J. Mimi, M. Chowdhury, T. Eva, F. Nainu, A. Paul, A. Shahriar, A. Tareq, N.U. Laam, S. Chakraborty, S. Shil, D. Mily, T.B. Hadda, F. Almalki, T. Emran, Immunoinformatics-guided design of an epitope-based vaccine against severe acute respiratory syndrome coronavirus 2 spike glycoprotein, Comput. Biol. Med. 124 (2020) 103967, https://doi.org/10.1016/j.compbiomed.2020.103967.

[17] A. Naz, F. Shahid, T. Butt, F. Awan, A. Ali, D. Malik, Designing multi-epitope vaccines to combat emerging coronavirus disease 2019 (covid-19) by employing immuno-informatics approach, Front. Immunol 11 (2020) 1663, https://doi.org/10.3389/fimmu.2020.01663.

[18] A. Grifoni, D. Weiskopf, S. Ramirez, J. Mateus, J. Dan, C. Moderbacher, S. Rawlings, A. Sutherland, L. Premkumar, R. Jadi, D. Marrama, A. Silva, A. Frazier, A. Carlin, J. Greenbaum, B. Peters, F. Krammer, D. Smith, S. Crotty, A. Sette, Targets of t cell responses to sars-cov-2 coronavirus in humans with covid-19 disease and unexposed individuals, Cell 181 (2020), https://doi.org/10.1016/j.cell.2020.05.015.

[19] Z. Noorimotlagh, C. Karami, S.A. Mirzaee, M. Kaffashian, S. Mami, M. Azizi, Immune and bioinformatics identification of t cell and b cell epitopes in the protein structure of sars-cov-2: A systematic review, Int. Immunopharmacol. 86 (2020) 106738, https://doi.org/10.1016/j.intimp.2020.106738.

[20] P.D. Yadav, V. Potdar, M.L. Choudhary, D.A. Nyayanit, M. Agrawal, S.M. Jadhav, T.D. Majumdar, A.S. Aich, A. Basu, P. Abraham, S.S. Cherian, Full-genome sequences of the first two sars-cov-2 viruses from india, Indian J Med Res 151 (2020), https://doi.org/10.4103/ijmr.IJMR_663_20.

[21] I. Saha, N. Ghosh, D. Maity, N. Sharma, J. Sarkar, K. Mitra, Genome-wide analysis of indian sars-cov-2 genomes for the identification of genetic mutation and snp, Infect. Genet. Evol. 85 (2020) 104457, https://doi.org/10.1016/j.meegid.2020.104457.

[22] W. Zhu, C. Wang, B.Z. Wang, From variation of influenza viral proteins to vaccine development, Int. J. Mol. Sci. 18 (2017), https://doi.org/10.3390/ijms18071554.

[23] A.K. Gupta, M.S. Khan, S. Choudhury, A. Mukhopadhyay, Sakshi, A. Rastogi, A. Thakur, P. Kumari, M. Kaur, Shalu, C. Saini, V. Sapehia, Barkha, P.K. Patel, K. T. Bhamare, M. Kumar, Coronavr: A computational resource and analysis of epitopes and therapeutics for severe acute respiratory syndrome coronavirus-2, Frontiers in Microbiology 11 (2020) 1858, https://doi.org/10.3389/fmicb.2020.01858.

[24] S.N. Crooke, I.G. Ovsyannikova, R.B. Kennedy, G.A. Poland, Immunoinformatic identification of b cell and t cell epitopes in the sars-cov-2 proteome, Sci. Rep. 10 (2020) 14179, https://doi.org/10.1038/s41598-020-70864-8.

[25] E. Ong, M.U. Wong, A. Huffman, Y. He, Covid-19 coronavirus vaccine design using reverse vaccinology and machine learning, Front. Immunol. 11 (2020) 1581, https://doi.org/10.3389/fimmu.2020.01581.

[26] M. Bhattacharya, A. Sharma, P. Patra, P. Ghosh, G. Sharma, B. Patra, S.S. Lee, C. Chakraborty, Development of epitope-based peptide vaccine against novel coronavirus 2019 (sars-cov-2): Immunoinformatics approach, J. Med. Virol. 92 (2020), https://doi.org/10.1002/jmv.25736.

[27] H.Z. Chen, L.L. Tang, X.L. Yu, J. Zhou, Y.F. Chang, X. Wu, Bioinformatics analysis of epitope-based vaccine design against the novel sars-cov-2, Infect. Dis. Poverty 9 (2020), https://doi.org/10.1186/s40249-020-00713-3.

[28] T. Kar, U. Narsaria, S. Basak, D. Deb, F. Castiglione, D. Mueller, A. Srivastava, A candidate multi-epitope vaccine against sars-cov-2, Sci. Rep. 10 (2020) 10895, https://doi.org/10.1038/s41598-020-67749-1.

[29] R. Bhatnager, M. Bhasin, J. Arora, A. Dang, Epitope based peptide vaccine against sars-cov2: an immune-informatics approach, J. Biomol. Struct. Dyn. (2020) 1–16, https://doi.org/10.1080/07391102.2020.1787227.

[30] A. Kwarteng, E. Asiedu, S.A. Sakyi, S.O. Asiedu, Targeting the sars-cov2 nucleocapsid protein for potential therapeutics using immuno-informatics and structure-based drug discovery techniques, Biomed. Pharmacotherapy (2020) 132, https://doi.org/10.1016/j.biopha.2020.110914.

[31] V. Baruah, S. Bose, Immunoinformatics-aided identification of t cell and b cell epitopes in the surface glycoprotein of 2019-ncov, J. Med. Virol. 92 (2020), https://doi.org/10.1002/jmv.25698.

[32] J. Bency, M. Helen, Novel epitope based peptides for vaccine against sars-cov-2 virus: immunoinformatics with docking approach, Int. J. Res. Med. Sci. 8 (2020) 2385, https://doi.org/10.18203/2320-6012.ijrms20202875.

[33] A. Singh, M. Thakur, L. Sharma, K. Chandra, Designing a multi-epitope peptide based vaccine against sars-cov-2, Sci. Rep. 10 (2020), https://doi.org/10.1038/s41598-020-73371-y.

[34] J.D. Thompson, D.G. Higgins, T.J. Gibson, Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (22) (1994) 4673–4680, https://doi.org/10.1093/nar/22.22.4673.

[35] R. Edgar, Muscle: Multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797, https://doi.org/10.1093/nar/gkh340.

[36] F. Sievers, A. Wilm, D. Dineen, T. Gibson, K. Karplus, W. Li, R. Lopez, H. Mcwilliam, M. Remmert, J. Söding, J. Thompson, D. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega, Mol. Syst. Biol. 7 (2011) 539, https://doi.org/10.1038/msb.2011.75.

[37] F. Sievers, D. Higgins, Clustal omega, Curr. Protocols Bioinformat. 48 (2014), https://doi.org/10.1002/0471250953.bi0313s48, 3.13.1–3.13.16.

[38] K. Katoh, K.I. Kuma, T. Miyata, H. Toh, Improvement in the accuracy of multiple sequence alignment program mafft, Genome informatics, Int. Conf. Genome Informat. 16 (2005) 22–33.

[39] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T. Madden, Ncb blast: a better web interface, Nucleic Acids Res. 36 (2008) W5–9, https://doi.org/10.1093/nar/gkn201.

[40] J. Sidney, C. Dow, B. Mothé, A. Sette, B. Peters, A systematic assessment of mhc class ii peptide binding predictions and evaluation of a consensus approach, PLoS Comput. Biol. 4 (2008) e1000048, https://doi.org/10.1371/journal.pcbi.1000048.

[41] A.C. Wallace, A.R. Laskowski, J.M. Thornton, Ligplot: a program to generate schematic diagrams of protein-ligand interactions, Protein Eng. Des. Select. 8 (2) (1995) 127–134, https://doi.org/10.1093/protein/8.2.127.

[42] M. Jespersen, B. Peters, M. Nielsen, P. Marcatili, Bepipred-2.0: Improving sequence-based b-cell epitope prediction using conformational epitopes, Nucleic Acids Res. 45 (2017), https://doi.org/10.1093/nar/gkx346.

[43] S. Yuan, H.S. Chan, Z. Hu, Using pymol as a platform for computational drug design, WIREs Comput. Mol. Sci. 7 (2) (2017) e1298, https://doi.org/10.1002/wcms.1298.

[44] M.A. Rauf, Ligand docking and binding site analysis with pymol and autodock/vina, Int. J. Basic Appl. Sci. 4 (2015) 168–177, https://doi.org/10.14419/ijbas.v4i2.4123.

[45] G. Janson, A. Paiardini, Pymod 3: a complete suite for structural bioinformatics in pymol, Bioinformatics (2020) 1367–4803, https://doi.org/10.1093/bioinformatics/btaa849.

[46] M. Wiederstein, M. Sippl, Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, Nucleic Acids Res. 35 (2007) W407–10, https://doi.org/10.1093/nar/gkm290.

[47] J. Salat, K. Mikulasek, O. Larralde, P.P. Formanova, A. Chrdle, J. Haviernik, J. Elsterova, D. Teislerova, M. Palus, L. Eyer, Z. Zdrahal, J. Petrik, D. Ruzek, Tick-borne encephalitis virus vaccines contain non-structural protein 1 antigen and may elicit ns1-specific antibody responses in vaccinated individuals, Vaccines 8 (2020), https://doi.org/10.3390/vaccines8010081.

[48] C.A. Gibson, J.J. Schlesinger, A.D.T. Barrett, Prospects for a virus non-structural protein as a subunit vaccine, Vaccine 6 (1) (1988) 7–9, https://doi.org/10.1016/0264-410X(88)90004-7.

[49] H.R. Chen, Y.C. Lai, T.M. Yeh, Dengue virus non-structural protein 1: a pathogenic factor, therapeutic target, and vaccine candidate, J. Biomed. Sci. 25 (58) (2018), https://doi.org/10.1186/s12929-018-0462-0.

[50] J. Lan, J.M. Zhou, Y. Yin, D.Y. Fang, Y.X. Tang, L.F. Jiang, Selection and identification of b-cell epitope on ns1 protein of dengue virus type 2, Virus Res. 150 (2010) 49–55, https://doi.org/10.1016/j.virusres.2010.02.012.

[51] P.P. Ip, A. Boerma, J. Regts, T. Meijerhof, J. Wilschut, H.W. Nijman, T. Daemen, Alphavirus-based vaccines encoding nonstructural proteins of hepatitis c virus induce robust and protective t-cell responses, Mol. Therapy 22 (4) (2014), https://doi.org/10.1038/mt.2013.287.

[52] R. Satyam, E. Janahi, T. Bhardwaj, P. Somvanshi, S. Haque, M. Najm, In silico identification of immunodominant b-cell and t-cell epitopes of non-structural proteins of usutu virus, Microbial Pathogen. 125 (2018), https://doi.org/10.1016/j.micpath.2018.09.019.

[53] A. Cafaro, A. Tripiciano, O. Picconi, C. Sgadari, S. Moretti, S. Buttò, P. Monini, B. Ensoli, Anti-tat immunity in hiv-1 infection: Effects of naturally occurring and vaccine-induced antibodies against tat on the course of the disease, Mol. Ther. 7 (3) (2019), https://doi.org/10.3390/vaccines7030099.

[54] R. Vita, S. Mahajan, J. Overton, S. Dhanda, S. Martini, J. Cantrell, D. Wheeler, A. Sette, B. Peters, The immune epitope database (iedb): 2018 update, Nucleic Acids Res. (2018) gky1006, https://doi.org/10.1093/nar/gky1006.
[55] I. Doytchinova, D. Flower, Vaxijen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. bmc bioinformatics 8:4, BMC Bioinformat. 8 (2007) 4, https://doi.org/10.1186/1471-2105-8-4.
[56] S. Saha, G. Raghava, Prediction methods for b-cell epitopes, Methods Mol. Biol. (Clifton, N.J.) 409 (2007) 387–394, https://doi.org/10.1007/978-1-60327-118-9_29.