



OPEN

# Designing a multi-epitope peptide based vaccine against SARS-CoV-2

Abhishek Singh<sup>1,2</sup>, Mukesh Thakur<sup>1</sup>✉, Lalit Kumar Sharma<sup>1</sup> & Kailash Chandra<sup>1</sup>

COVID-19 pandemic has resulted in 16,114,449 cases with 646,641 deaths from the 217 countries, or territories as on July 27th 2020. Due to multifaceted issues and challenges in the implementation of the safety and preventive measures, inconsistent coordination between societies-governments and most importantly lack of specific vaccine to SARS-CoV-2, the spread of the virus that initially emerged at Wuhan is still uprising after taking a heavy toll on human life. In the present study, we mapped immunogenic epitopes present on the four structural proteins of SARS-CoV-2 and we designed a multi-epitope peptide based vaccine that, demonstrated a high immunogenic response with a vast application on world's human population. On codon optimization and in-silico cloning, we found that candidate vaccine showed high expression in *E. coli* and immune simulation resulted in inducing a high level of both B-cell and T-cell mediated immunity. The results predicted that exposure of vaccine by administrating three injections significantly subsidized the antigen growth in the system. The proposed candidate vaccine found promising by yielding desired results and hence, should be validated by practical experimentations for its functioning and efficacy to neutralize SARS-CoV-2.

Coronavirus disease 2019 (COVID-19) is one of the most potent pandemics characterized by respiratory infections in Human<sup>1</sup>. As per the latest report of World Health Organization (WHO) dated July 27th 2020, the worldwide coverage of COVID-19 exceeded 16,114,449 cases with mortality of 646,641 deaths individuals from 217 countries, areas or territories<sup>2</sup>. The outbreak of COVID-19 initially emerged at Wuhan city of Hubei Province of People Republic of China in the early December 2019<sup>3</sup>, and the report of severe pneumonic patients with unknown etiology of symptoms was released by the Chinese Center for Disease Control on December 31st 2019<sup>4</sup>. Later on, the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was confirmed as the causative agent for the clusters of pneumonic cases by the health professionals. Further investigations on the origin of infection possibly linked the SARS-CoV-2 to a seafood market in the Wuhan city, China which got spread to a large extent of human population in a short time span<sup>5</sup>. By the time, scientists/ health professionals paid a sincere note to understand the mode of transmission of SARS-CoV-2, it already infected a large number of people in various countries, by exhibiting a high transmission rate in comparison to the earlier human CoV<sup>6</sup>. This led the WHO to declare COVID-19 as a Public Health Emergency of International Concern (PHEIC) on January 30th 2020<sup>7</sup>.

Until the outbreak of SARS in 2002 and 2003, coronaviruses are considered mild pathogenic virus causing respiratory and intestinal infections in animals and humans<sup>8–12</sup>. With a decade long pandemic of SARS, another potent pathogenic coronavirus, i.e., Middle East Respiratory Syndrome Coronavirus (MERS-CoV) appeared explicitly in the Middle Eastern countries<sup>13</sup>. Extensive studies on these two coronaviruses have led to an understanding that they are genetically diverse but likely to be originated from bats<sup>14,15</sup>. Apart from these two viruses, four other coronaviruses, i.e. HKU1, 229E, OC43 and NL63 are commonly detected in Humans<sup>16,17</sup>. Among all the coronaviruses reported so far, SARS-CoV, MERS-CoV, and recently emerged SARS-CoV-2 have relatively high pathogenicity and fast mutagenic abilities<sup>18</sup>. SARS CoV alone was responsible for 8422 positive cases and 919 probable fatalities over a vast population of 32 countries<sup>19</sup>. With the first mortality case reported in 2012, 2496 cases were found positive for MERS-CoV, and 868 deaths were reported from 27 countries<sup>19</sup>. Both of these pandemics were considered as highly pathogenic in the last two decades until the outbreak of SARS-CoV-2 in December 2019. SARS-CoV-2 has already spread in 217 countries and caused 646,641 deaths since its outbreak in December 2019<sup>2</sup>. As in the current scenario, several countries, e.g. USA, Brazil, India, Russia, South Africa, Mexico, Peru, Chile, Spain, United Kingdom etc. have already experienced or experiencing the most fatal problem of community spread of SARS-CoV-2. Though the estimated mortality rate due to SARS-CoV-2 is about

<sup>1</sup>Zoological Survey of India, New Alipore, Kolkata, West Bengal 700053, India. <sup>2</sup>Gujarat Forensic Sciences University, Gandhinagar, Gujarat 382007, India. ✉email: [thamukesh@gmail.com](mailto:thamukesh@gmail.com)

Structural glycoprotein	Protein ID (NCBI)	Size (aa)	Antigenicity score
Surface	QHD43416.1	1273	0.4646
Membrane	QHD43419.1	222	0.5102
Envelop	QHD43418.1	75	0.6025
Nucleocapsid	QHD43423.2	419	0.5059

**Table 1.** Details of target protein and their antigenicity score.

3.6% in China and about 1.5% outside China<sup>20</sup>, the problem is grave and need to be respond urgently to curb the rapid loss of human life.

Coronaviruses are broadly divided into four genera, Alpha-CoV, Beta-CoV, Gamma-CoV and Delta-CoV, where Alpha-CoV and Beta-CoV are only reported to infect mammals but Gamma-CoV and Delta-CoV are also capable of infecting both mammals and birds<sup>21</sup>. An extensive study of their biology suggests that they are non-segmented enveloped viruses with a positive-sense single-stranded RNA of 26–32 kb size<sup>16</sup>. Some of the recent research indicated that SARS-CoV-2 has an identical genomic organization as of Beta-CoV, and thus the genomic structure of SARS-CoV-2 follows 5'-leader-UTR-replicase- spike-envelope-membrane-nucleocapsid and 3' UTRpoly (A) tail with the subsequent species-specific accessory gene at 3' terminal of the viral genome<sup>22</sup>. The structural glycoprotein includes Spike/Surface (S), Envelop (E), Membrane (M) and Nucleocapsid (N). The S glycoprotein is reported to have a crucial role in the virus transmission as the receptor binding capability and entry to the host cell is regulated by the expression of S glycoprotein<sup>23</sup>. The E and M glycoproteins are responsible for viral assembly and N glycoprotein is necessary for RNA genome synthesis<sup>23</sup>. The complex genetic makeup and high mutation rate of SARS-CoV-2 requires the strategic development of a vaccine by targeting all the structural proteins. Unfortunately, no approved vaccine is registered so far against SARS-CoV-2, but a few candidate vaccines are under trial. A recent study reported the first in human trial of adenovirus type-5 vectored COVID-19 vaccine<sup>24</sup>. Although, some recent studies have proposed peptide-based vaccine targets using the structural glycoproteins<sup>25–31</sup> but no study has covered all the structural proteins for vaccine designing. With the recent development of integrated bioinformatics and immunoinformatics approaches, vaccine design and its specific application have become rapid and cost-effective. Now a days, the targeted immunogenic peptide can be designed and validated for its efficacy to work on biological systems even before having the vaccine in hand for practical exposure and experiments<sup>32</sup>. However, conventional methods have simultaneously proven effective in vaccine design with some limitations, e.g. sacrificing the whole organism or a large protein residue which unnecessarily increases the antigenic load and probability of allergenicity<sup>33</sup>. We believe that an ideal multi-epitope vaccine can enhance immune response and subsequently lower down the risk of re-infection by enhancing the host immunogenicity. In this perspective, here we applied integrated approaches of vaccine design by mapping B-cell, T-cell and IFN-gamma epitopes present on the four structural proteins of the virus, based on several parameters and then the final proposed vaccine construct was validated using the in-silico immune simulations by injecting the vaccine to monitor the immune response.

## Results

**Target glycoproteins and their antigenicity.** We scrutinized target structural protein sequences, i.e. S, M, E and N glycoproteins from the whole-genome of SARS-CoV-2, (GenBank: MN908947.3). Antigenicity prediction for the sequences of structural proteins revealed that E glycoproteins have the highest antigenicity score 0.6025, followed by M, N and S glycoproteins (Table 1).

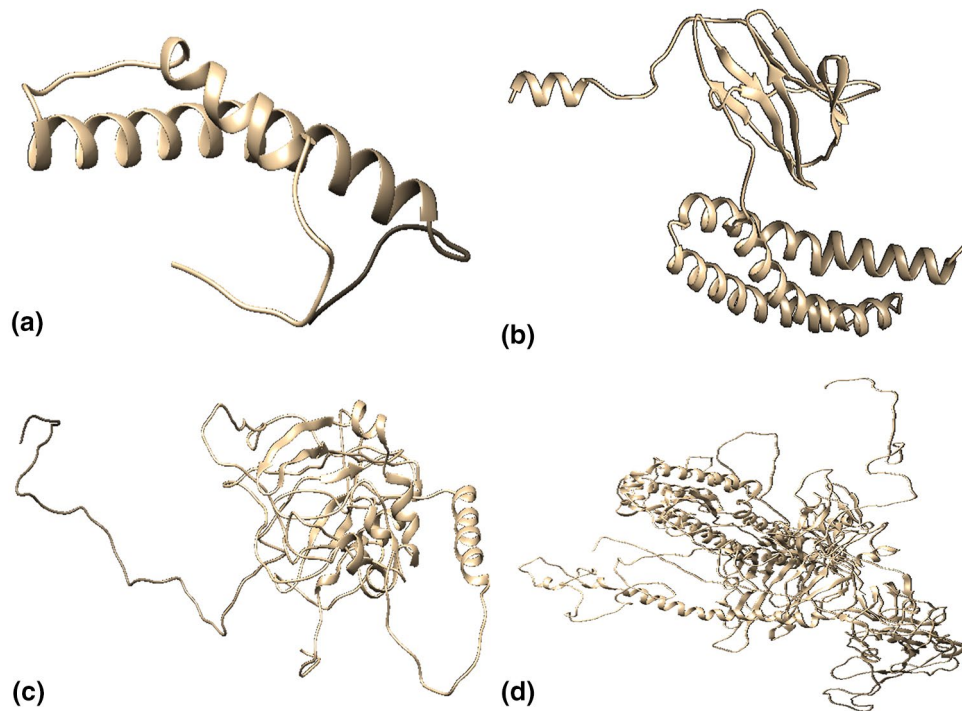
**Physicochemical and secondary structural properties.** Theoretical estimates of isoelectric point (pI) suggested that all the structural proteins were basic and their predicted instability indices were below 40 indicating their stable nature except the N glycoprotein with an instability index of 55.09 (Table S1). The predicted secondary structure characteristics of all the structural proteins showed variable percentage of alpha helix, extended strand, beta turn and random coil (Table S2).

**Three dimensional (3D) structure of target glycoproteins.** The refined finalized 3D structures obtained from Raptor-X showed a very low percentage of outliers and found better than the models generated from I-Tasser and Phyre-2 (Fig. 1). Details of various models of the structural glycoproteins through Ramachandran plot analysis are given in Table S3.

**Prediction of CTL and HTL epitopes.** A total of 29 CTL epitopes in E glycoprotein, 77 in N, 340 in S and 89 epitopes in M glycoprotein were predicted with strong binding affinity for multiple alleles. Similarly, 35 HTL epitopes in E glycoprotein, 51 in N, 252 in S and 23 epitopes in M glycoprotein were predicted. All these epitopes were strong binders and predicted percentile rank of  $\leq 2$ .

**Identification of overlapping T cell epitopes.** The CTL epitopes overlapping with HTL epitopes were scrutinized and subjected to immunogenicity and allergenicity prediction. In total, 16 CTL overlapping epitopes were predicted in E glycoprotein, 8 in N glycoprotein, 25 in S glycoprotein and 9 epitopes in M glycoprotein (Tables S4–S11). All these selected epitopes were predicted to be non-allergic with high antigenic scores.

## List of Figures



**Figure 1.** The 3D models of structural proteins of SARS-CoV-2. **(a)** Envelop protein; **(b)** Membrane protein; **(c)** Nucleocapsid protein; **(d)** Surface protein.

**Identification of B-cell epitopes.** The identified continuous B cell epitopes present on the surface of the structural proteins were selected based on their antigenic, non-allergic nature coinciding with the CTL and HTL epitopes. With these criteria, five epitopes were obtained in S glycoprotein, three in N and one in M glycoprotein (Table S12). No epitope with such criteria was identified in E glycoprotein. Further, one discontinuous epitope in E glycoprotein, seven in N, nine in S and four discontinuous epitopes were predicted in M glycoprotein (Table S13).

**Identification of IFN-gamma epitopes.** The server IFN epitope predicted eight epitopes in E glycoprotein, 39 in M, 66 in N and 281 in S glycoprotein. Based on antigenicity and non-allergenicity, eight epitopes were scrutinized in E glycoprotein, 13 in M, 43 in N and 81 in S glycoprotein. To obtain the highest immunogenicity and reduce the overload of epitopes in vaccine construct, we prioritized one epitope from each glycoprotein with the highest antigenicity score and non-allergenicity (Table S14).

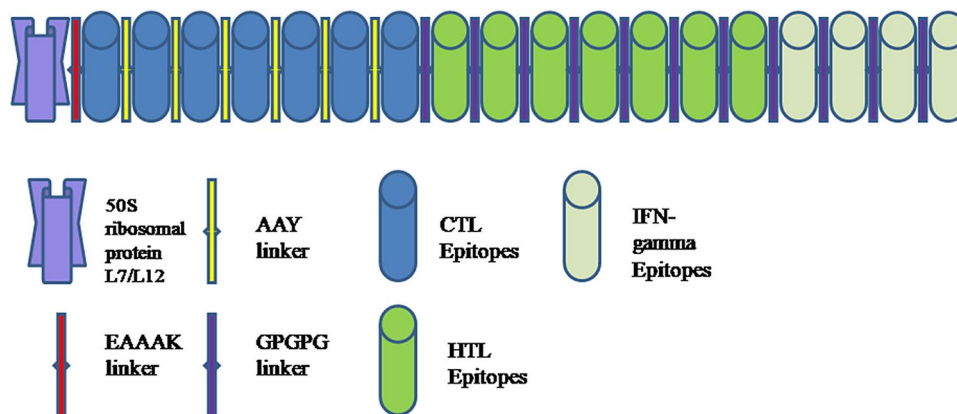
**Characterization of predicted epitopes.** *Conservation analysis, population coverage, and autoimmunity identification.* Conservation prediction resulted that all selected T-cell and B-cell epitopes were 100% conserved among the target structural proteins and population coverage prediction of T-cell epitopes showed that five epitopes in E glycoprotein, two in M, two in N and six epitopes in S glycoprotein covered more than 50% of the worldwide population. No selected epitope with >50% population coverage showed homology with any human protein. Thus, altogether 30 epitopes, 15 each of CTL and HTL, met the above criteria were processed for the interaction analysis with the commonly occurring HLA alleles in the Human population (Table S15).

*Interaction analysis of epitopes and their HLA alleles.* Molecular interaction analysis of each selected CTL and HTL epitopes with their respective HLA alleles resulted in 10 predicted docked complexes. Each complex showed different global energy value. Of the 30 characterized epitopes, those epitopes showing global energy value threshold of  $-40$  or above were considered as effectively interacting epitopes with the corresponding HLA allele, and thus selected for final vaccine construct (Table 2). This resulted in two CTL and HTL epitopes in E glycoprotein, one in M, two in N and two in S glycoprotein. For an ideal candidate vaccine, the prioritized epitopes included should have a potential affinity towards their respective alleles. Therefore, we selected epitopes showing high interaction with their alleles.

*Construction of final multi-epitope vaccine.* Finally, we selected 18 epitopes, i.e. seven CTL, seven HTL and four IFN-gamma for the development of the multi-epitope vaccine (Figure S1). The adjuvant, i.e. 50S ribosomal protein L7/L12 was coupled by the EAAAK linker with CTL epitope and subsequently, AAY linker was used to

S. no	Protein	HTL	Global energy	CTL	Global energy
1	Envelop	LLFLAFVVFLVTLA	-71.44	LLFLAFVVF	-41.45
2	Envelop	LAFVVFLVTLAILT	-64.39	FLLVTLAIL	-43.31
3	Membrane	TLACFVLAAYRINW	-53.12	TLACFVLAA	-54.85
4	Nucleocapsid	GDAALALLLDRLNQ	-87.95	GDAALALLL	-40.49
5	Nucleocapsid	AQFAPSASAFFGMSR	-58.74	AQFAPSASA	-43.29
6	Surface	IPFAMQMAYRFNGIG	-46.33	FAMQMAYRF	-41.37
7	Surface	FVFLVLLPLVSSQCV	-74.92	FVFLVLLPL	-55.15

**Table 2.** Final epitopes selected for multi-epitope vaccine construct after docking analysis with respective HLA alleles.



**Figure 2.** Schematic diagram of final vaccine construct.

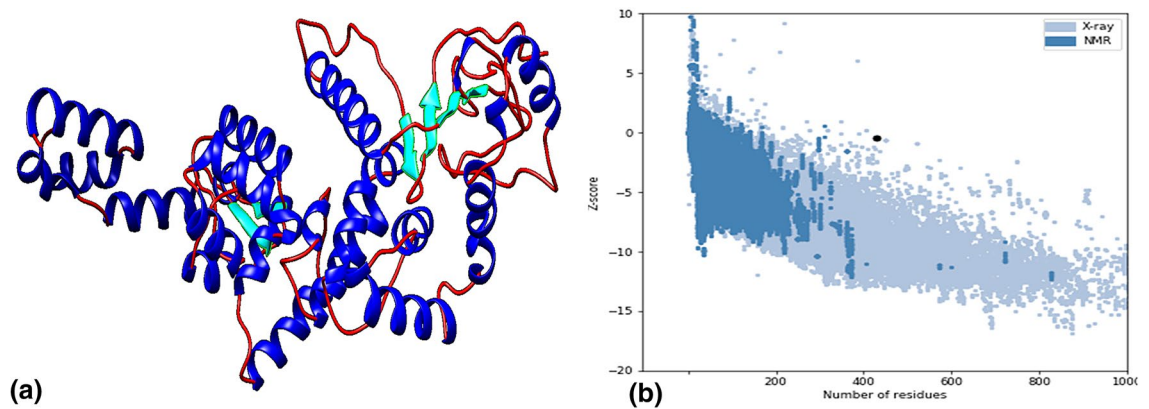
couple CTL epitopes and GPGPG linker was used to couple HTL and IFN-gamma epitopes (Fig. 2). The final multi-epitope vaccine construct was composed of 430 amino acid residues which were then validated for antigenic, allergenic and physicochemical properties.

*Evaluation of antigenicity, allergenicity, physicochemical properties and secondary structure of the vaccine construct.* The multi-epitope vaccine construct found to be immunogenic (Ag score- 0.627), non-allergic with 45.131 KDa molecular weight. The prediction suggested that vaccine construct was stable and basic with instability index 23.46 and theoretical pI 9.01. The half-life of the vaccine in mammalian reticulocytes (in vitro) was 30 h, while it was about 20 h in yeast (in-vivo) and 10 h in *E. coli* (in vivo). The predicted secondary structure of the vaccine construct consisted of 42.33% alpha-helix, 22.09% extended strand, 4.42% beta-turn and 31.16% random coil.

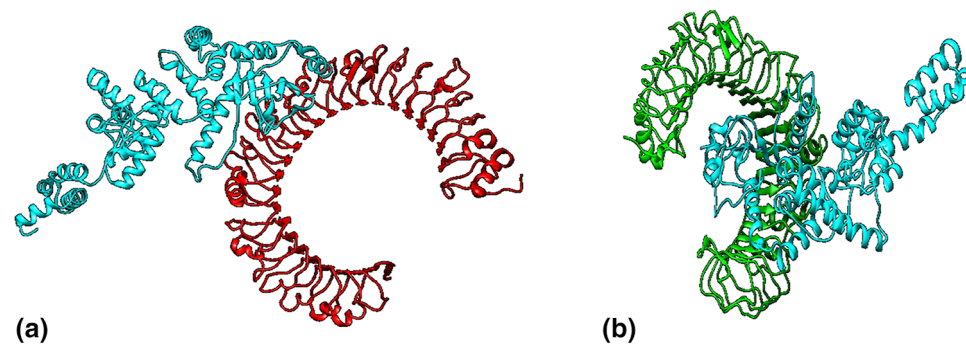
*Modelling, refinement, and validation of vaccine construct.* Of the top 10 predicted models, the model with the Z-score of 7.61 with template 1dd3A was selected for improvement. The ProSA-web tool assessed the overall and local quality of the refined model with a Z-score of -0.47, found close to the range of native protein of similar size (Fig. 3). Further, the Ramachandran plot analysis revealed 92% of residues in the favoured region, 6.1% in the allowed region and 1.9% outliers.

*Prediction of continuous and discontinuous B-cell epitopes.* We found four continuous epitopes, - AKILKEKYGLD, EILDKSKEKTSFD, LKESKDLV and VPKHLKKGLSKEEAESLKKQLEEV on the surface of predicted vaccine following prediction made by BCpred 2.0 server (Figure S2), while Ellipro server predicted six discontinuous epitopes with a score higher than 0.8 (Figure S3; Table S16).

*Molecular docking of vaccine construct with TLR3 and TLR4 receptor.* Cluspro v.2 predicted 30 models each of vaccine receptor TLR3 complex and TLR4 complex with their corresponding cluster scores (Table S17–S18). Among these models, the model 2 in TLR3 complex and model 1 in TLR4 complex was selected as a best-docked complex with the lowest energy score of -1199.1 with 46 members (TLR3) and lowest energy score of -1229.9 with 79 members (TLR4). This signifies potential molecular interaction between predicted vaccine construct with TLR3 and TLR 4 receptors (Fig. 4).



**Figure 3.** (a) 3 dimensional structure of final vaccine construct, (b) ProSA validation of predicted structure with Z score of  $-0.47$ .



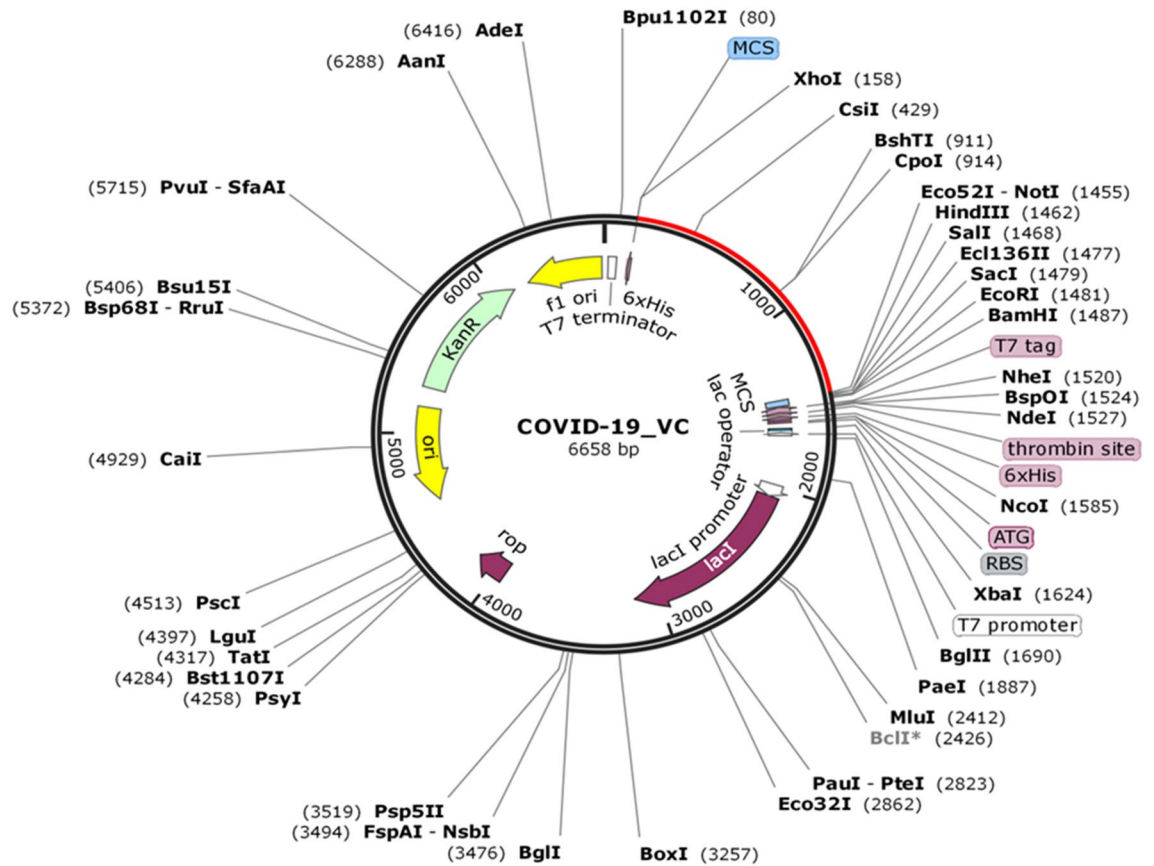
**Figure 4.** The interaction pattern of designed vaccine with TLR3 and TLR4. (a) Vaccine (Cyan) docked with receptor TLR3 (Red). (b) Vaccine (Cyan) docked with receptor TLR4 (Green).

**In-silico cloning and vaccine optimization.** The Java Codon Adaptation Tool optimized the codon usage of the vaccine and produced an optimized codon sequence of 1290 nucleotides. The Codon Adaptation Index (CAI) of the optimized sequences was 0.95, and the GC content was 54.41%, which ensured the potential, expression of vaccine construct in the host cell, *E. coli*. Later on, the restriction sequence of Xho I and Not I restriction enzymes were added to N and C terminal of adapted codon sequence. Further, the adapted sequence was also cloned in pET-28(+) vector using SnapGene tool (Fig. 5).

**Immune simulations of vaccine construct.** The simulations result suggested that administration of the vaccine by three injections was good enough to potentially induce various immunoglobulins. The primary response appeared by the increase in the level of IgM and the secondary response was characterized by an increased level of IgM + IgG, IgG1 + IgG2, IgG1, IgG2, and B-cell populations. On the subsequent exposure of three injections of the vaccine, a decrease in the level of antigens was observed. Both T cell populations, i.e. CTL and HTL, developed increased response corresponding to the memory cells signifying the immunogenicity of T cell epitopes included in the vaccine construct. Increased activity of macrophages was also observed at each exposure, whereas the activity of NK cells was observed consistently throughout the period. A significant increase in the level of IFN-gamma, IL-10, IL-23, and IL-12 was also found at subsequent exposure (Fig. 6). After the repeated exposure of vaccine construct through 12 injections, on a regular interval, the level of antigen attained a similar peak (figure S4-a), but showed a remarkable increase in the level of IgM + IgG, IgG1 + IgG2. A surge in the memory corresponding to B cell and T cells were observed throughout the exposure while the level of IFN-gamma was consistently high from first to last exposure (Figure S4). This signifies that the vaccine generated a robust immune response in case of short exposure and immunity increases even on subsequent repeated exposure.

## Discussion

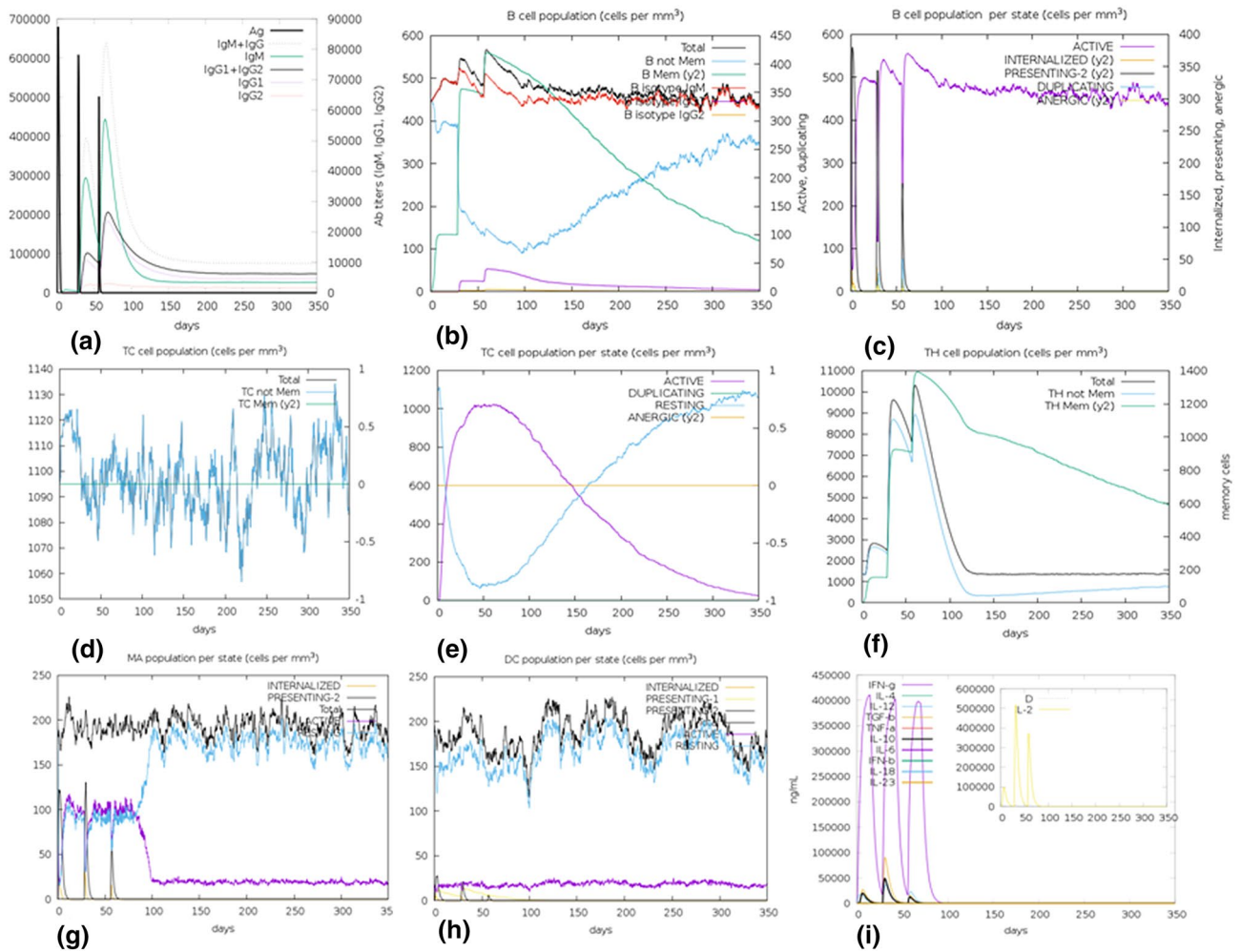
The recent outbreak of SARS-CoV-2 in Wuhan city of China raised several questions regarding the susceptibility of humans against this novel pathogen. As predicted earlier about the re-emergence of unique variants of SARS-CoV<sup>34</sup>, SARS-CoV-2 has spread more widely in a short period as compared to the previous variants. This escalated rate of transmission initiated a pursuit for vaccine development against SARS-CoV-2, which already caused 646,641 deaths as on July 27th 2020<sup>2</sup>. With strict and comprehensive measures, vaccine development



**Figure 5.** In silico cloning of vaccine. The segment represented in red is the multi-epitope vaccine insert in pET-28(+) expression vector.

and its application could play a key role in eliminating the virus from the human population or to restrain the community spread. Thus several efforts are being made to address the challenge that appeared in the current scenario with appreciable advancements in the understanding of virus biology and its etiology<sup>35</sup>. The lack of knowledge regarding the response of the immune system against viral infection is one of the major limitations in the path of vaccine development for SARS-CoV-2.

This study is a novel attempt in describing the potential immunogenic target using the all four structural glycoproteins of the virus and proposes a novel multi-epitope vaccine construct, by providing detailed insights in the initial phase of vaccine development. Since, immunity against any antigen is prominently dependent on, how it gets recognized by B cells and T cells. We identified epitopes corresponding to B cells and T cells in each structural protein so that both innate and adaptive immunity can be induced with the exposure of vaccine construct. Further, IFN-gamma epitopes were identified which are equally effective in immune regulatory, antiviral and antitumor activities. The final vaccine construct was designed with due consideration of the overlapping T-cell epitopes and IFN-gamma epitopes to increase the possibilities of immune recognition ability. The epitopes were prioritized following various parameters, e.g. nature of epitopes being antigenic, non-allergic, conservation efficiency among the target proteins, their affinity for multiple alleles, homology search with any of the human proteins, their vast coverage on human populations and effective molecular interaction with their respective HLA alleles. The highest antigenicity score for E glycoprotein showed the most potent candidate to generate immune response. Considering the functional role of E glycoprotein in the life history traits of the virus, Schoeman and Fielding<sup>23</sup> recently reported that those CoVs lacking E protein make promising candidate vaccine. The designed vaccine of 430 amino acid residues with the molecular weight of 45.131 KDa, also found well fit within the defined range (40–50 KDa) of average molecular weight of a multi-epitope vaccine<sup>36</sup>. The instability index and theoretical pI also suggested that the vaccine is stable, basic in nature and the estimated half-life suggested that the recognized peptide does not possess a short half-life and would remain viable for a span adequate to generate a potent immune response. Vaccine model refinement and validation indicated that the quality of the predicted model was good as more than 90% residues were in the favoured region<sup>37</sup>. The B-cell epitopes were predicted on the surface of the vaccine, which showed its effectiveness to be recognized by the B-cells. Since TLR 3 and TLR 4 have proven recognition capability in both SARS-CoV and MERS-CoV<sup>38,39</sup> and considering a similar genome organization of SARS-CoV-2 like SARS-CoV, the molecular interaction of vaccine through docking analysis suggested that candidate vaccine has significant affinity to TLR 3 and TLR 4 to act as a sensor for recognizing molecular patterns of pathogen and initiating immune response. Thus, the vaccine TLR complex is capable of generating an effective innate and adaptive immune response against SARS-CoV-2. Further, Codon adaptation



**Figure 6.** In silico simulation of immune response using vaccine as an antigen after subsequent three injections. (a) Antigen and Immunoglobins. (b) B-cell population. (c) B-cell population per state. (d) Cytotoxic T-cell population. (e) Cytotoxic T-cell population per state. (f) Helper T-cell population. (g) Macrophages population per state. (h) Dendritic cell population per state. (i) Cytokine production.

improved the expression of the candidate vaccine in *E. coli* strain K12 with significant codon adaptation index and GC content indicating elevated expression level. After successful cloning of the candidate vaccine in the pET-28(+) vector, the simulation-based generated immune response, suggested that the primary and secondary immune response will be coherent with the actually expected response. Subsequent exposure for three injections was found adequate in generating effective immunogenic response. The consistently high level of IFN-gamma also supported the activation of innate immunity. However, repeated exposure of the vaccine by administering 12 periodic injections suggested an obvious more stringent immune response. All the immunoinformatics approaches demonstrated that vaccine designed against SARS-CoV-2 yielded promising results by reducing antigen growth and inducing immune response.

Further, to seek the convenient proof and reliability of the immunoinformatic approaches used in designing the potential candidate vaccine, we compared the predicted epitopes with the experimentally determined SARS-CoV-2-derived B cell and T cell epitopes<sup>30</sup>. We observed thirteen epitopes i.e. six T-cell nucleocapsid epitopes, one T-cell surface epitope, two linear B cell epitopes and four discontinuous B-cell epitopes, were either identical and/or overlapping with Ahmed et al.<sup>30</sup> with high antigenicity and allergenicity scores (Table S19). Since, we have considered all four structural proteins to screen T-cell, B-cell and IFN gamma epitopes for designing the potential candidate vaccine, the present study provides a comprehensive and better insights for vaccine design against SARS-CoV-2.

Without following the standard operating procedures and guidelines issued time to time by WHO, it is difficult to curb coronavirus pandemic given the situation when no vaccine is available. However, a few therapeutics have been recently tested against SARS-CoV-2<sup>5,40</sup>, but none of them showed complete effectiveness to cure the disease. In this study, a vaccinomics approach was carried out to design a multi-epitope vaccine against SARS-CoV-2 using several in-silico and immunoinformatic methods. We propose that the designed vaccine has all the potential to induce both the innate and adaptive immune systems and can neutralize the SARS-CoV-2. Therefore, we appeal for an experimental validation of the designed vaccine must be undertaken by the professionals/clinicians for reaching a conclusion for its safety, efficacy and success.

**Methodology.** *Retrieval of SARS-CoV-2 protein sequences and antigenicity prediction.* Sequences of all the four structural glycoproteins of SARS-CoV-2 were retrieved from NCBI (GenBank: MN908947.3). We predicted antigenicity from each protein using the online server VaxiJen v2.0<sup>41</sup> and proteins that showed antigenicity  $\geq 0.4$  in the virus category were subjected for further analysis.

*Prediction of physicochemical and secondary structural properties of the target proteins.* The physicochemical properties of target proteins were examined using ExPASy ProtParam online server<sup>42</sup> and the conformational states, e.g. Helix, sheet, turn and coil was predicted using the online Secondary structure analysis tool (SOPMA)<sup>43</sup> with default parameters.

*3D homology modelling and validation.* The 3D structure of target proteins was modelled using three homology modelling tools, i.e. I-Tasser, Raptor-X and Phyre2<sup>44–46</sup> and post processing to reduce any distortion in the modelled structure was done using the Galaxy refine server<sup>47</sup>. The modelled structures were subjected to RAM-PAGE server for Ramachandran plot analysis for quality check and validation<sup>48</sup> and visualized using Chimera 1.10.1 visualization system<sup>49</sup>.

**Prediction of T-cell epitopes.** *Cytotoxic T-cell (CTL) epitopes.* Nine residues long CTL epitopes recognized by HLA class-I supertypes, i.e. A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, B62 were predicted using NetCTL.1.2 server<sup>50</sup> considering the default values of weight on C terminal cleavage, weight on TAP transport efficiency and the threshold for epitope identification. Further, another set of CTL epitopes recognized by HLA class I alleles like A\*02:02, B\*07:02, C\*04:01, E\*01:01, etc. were identified by consensus method of Immune Epitope Database (IEDB) tool<sup>51</sup>. The identified epitopes were scrutinized based on the consensus score, i.e.  $\leq 2$  and only those epitopes were considered as strong binders and selected for further analysis, which were predicted by more than one allele by both the methods.

*Helper T-cell (HTL) epitopes.* Fifteen residue long HTL epitopes recognized by HLA Class II DRB1 alleles were predicted using Net MHC II pan 3.2 servers<sup>52</sup>. The threshold for strong binder and weak binder were set as default. Further another set of HTL epitopes of 15 residues length recognized by HLA DR alleles were identified using the consensus method as implemented in the IEDB server<sup>51</sup>. The epitopes with a percentile rank of  $\leq 2$  and predicted by more than one allele by both methods were considered as strong binders and scrutinized for further analysis.

*Identification of overlapping T-cell epitopes.* Considering the fact that epitopes with an affinity for multiple HLA alleles tend to induce relatively more immune response in the host cell, we scrutinized overlapping epitopes with affinity to both HLA class I and class II alleles for immunogenicity and allergen prediction using VaxiJen v2.0 server<sup>41</sup> and AllerTop v.2.0<sup>53</sup> tool with an immunogenic threshold of 0.4. These epitopes were assumed to show high potential to activate both CTL and HTL cells.

*Identification of continuous and discontinuous B-cell epitopes.* B-cell epitopes are short amino acid sequences recognized by the surface-bound receptors of B-lymphocyte. Their identification plays a major role in vaccine design, and thus two types-continuous and discontinuous B-cell epitopes were predicted over the corresponding structural protein sequences. The continuous B-cell epitopes, present on the surface of proteins, were predicted using BCPred 2.0<sup>54</sup> server and the discontinuous epitopes which were reasonably large were predicted using Ellipro server<sup>55</sup>. The prediction parameters like Minimum score and Maximum distance (Angstrom) were set as 0.8 and 6, respectively.

*Identification of IFN-gamma epitopes.* Since IFN-gamma is the signature cytokine of both the innate and adaptive immune system, therefore epitopes with potency to induce IFN-gamma could boost the immunogenic capacity of any vaccine. Thus, IFN-gamma epitopes were predicted for each structural protein by the IFN epitope server<sup>56</sup>.

**Characterization of predicted epitopes.** *Conservation analysis.* Predicted T-cell and B-cell epitopes were submitted to the IEDB conservancy analysis tool<sup>57</sup> to identify the degree of conservancy in the structural protein sequences, and the epitopes with 100% conservancy were selected for further analysis.

*Population coverage and autoimmunity identification.* The immunogenic response to human population towards the selected overlapping CTL and HTL epitopes against their respective HLA genotype frequencies was predicted using IEDB population coverage analysis tool<sup>58</sup>. Those epitopes exhibiting 50% or more population coverage were scrutinized. To reduce the probability of autoimmunity, all the selected epitopes were subjected to Blastp search analysis against the Human proteome. Epitopes showing similarity to any human protein were excluded from further analysis.

*Interaction analysis of epitopes and their HLA alleles.* The sequence of scrutinized epitopes was submitted to an online server PEPFOLD 3<sup>59</sup> for 3D structure modelling. The structure of the most common HLA alleles, i.e. HLA-DRB1 \*01:01 (HLA class II) and HLA-A\* 02:01 (HLA class I) in human population were retrieved from Protein data bank (PDB) with a PDB ID of 2G9H and 1QEWE. Any ligand associated with the HLA allele's struc-



ture was removed, and energy-minimization was carried out. After that, the modelled epitopes were docked with the corresponding HLA allele using a web server Patchdock<sup>60</sup>, to study the interaction pattern of receptor and ligand. The parameters like clustering RMSD value and the complex type was kept as 0.4 and default. The result obtained from Patchdock was forwarded to the Firedock server<sup>61</sup> for the refinement of the best 10 models. Based on Global energy value, the complex with the lowest global energy was scrutinized, and the corresponding CTL and HTL epitopes were selected for the final vaccine construct.

**Construction and quality control of final multi-epitope vaccine.** For the construction of a multi-epitope vaccine against SARS-CoV-2, we followed Chauhan et al.<sup>33</sup> with some modifications. The epitopes were prioritized for selection based on meeting the following criteria: (1) epitopes must be immunogenic, non-allergic and possessing overlapping affinity to both HLA class I and class II alleles; (2) epitopes must be capable of activating both CTL and HTL cells and must have a minimum 50% of the population coverage; (3) epitopes should not be overlapping with any human gene, and the predicted B-Cell epitopes should be present on the surface of the target protein. Based on meeting these criteria, the HTL, CTL and IFN-gamma epitopes were selected in the final construct of the multi-epitope vaccine. The HTL and IFN-gamma epitopes were linked by GPGPG linkers, whereas the CTL epitopes were linked by AAY linkers<sup>33</sup>. An adjuvant, 50S ribosomal protein L7/L12 (Locus RL7\_MYCTU) with an NCBI accession no: P9WHE3 was added to the N terminal to enhance the immunogenicity of the vaccine construct.

**Antigenicity, allergenicity and physicochemical properties of the vaccine.** The antigenicity of the proposed vaccine was predicted using Vaxijen v2.0 server<sup>41</sup>, and allergenicity was predicted using Allertop v2.0 prediction tool<sup>53</sup>. The physicochemical property of the designed vaccine was predicted by submitting the final sequence of the vaccine to the Expasy Protpram online server<sup>42</sup>.

**Prediction of secondary structure.** The secondary structure of the vaccine construct was predicted using the SOPMA tool<sup>43</sup>. With the input of vaccine sequence and output width 70, the parameters like the number of conformational states, similarity threshold, and window width were set as 4 (Helix, Sheet, Turn, Coil), 8 and 17.

**Modelling, refinement, and validation of vaccine construct.** The final model of the vaccine construct was prepared using a homology modelling tool SPARKS-X<sup>62</sup>. Among the top 10 predicted models, the model with the highest Z-score was selected and subjected for refinement using the Galaxy refine server<sup>47</sup>. Then after, we utilized ProSA-webtool<sup>63</sup> to assess the overall and local quality of the predicted model based on the Z-score. If the Z-score lies outside the range of native protein of similar size, a chance of error increases in the predicted structure. Further to check the overall quality of the refined structure of vaccine, Ramachandran plot analysis was carried out using the RAMPAGE server<sup>48</sup>. The final model of the vaccine construct was visualized using Chimera 1.10.1 visualization system<sup>49</sup>.

**Prediction of continuous and discontinuous epitopes.** The continuous and discontinuous B-cell epitopes in the vaccine construct were predicted using BCPred 2.0<sup>54</sup> and Ellipro server<sup>55</sup>, respectively. For the effective design of vaccine, the vaccine construct should possess B-cell epitopes for potential enhancement of immunogenic reactions.

**Molecular docking of vaccine construct with TLR3 and TLR4 receptor.** The 3D structures of human TLR3 and TLR4 were retrieved from protein data bank (PDB ID: 2A0Z and 2z63). Any ligand attached to the retrieved structures was removed, and the receptor was subjected for model refinement using Galaxy refine server<sup>47</sup>. Molecular docking analysis was performed using the Cluspro v.2 protein–protein docking server<sup>64</sup> to analyze the interaction pattern of vaccine with TLR3 and TLR4. The server provides cluster scores based on rigid docking by sampling billions of conformation, pairwise RMSD energy minimization. Based on the lowest energy weight score and members, the final vaccine-TLR3 complex and TLR4 complex model was selected and visualized using Chimera 1.10.1 visualization system<sup>49</sup>.

**In-silico cloning and vaccine optimization.** The efficiency of the multi-epitope vaccine construct in cloning and expression is of the utmost importance for vaccine design. The final multi-epitope vaccine construct was submitted to the Java Codon Adaptation Tool (JCat)<sup>65</sup>, where codon optimization was performed in *E. coli* strain K12. In addition to this, parameters like Avoid rho-independent transcription terminators, Avoid prokaryotic ribosome binding sites, Avoid Cleavage Sites of Restriction Enzymes were selected for the final submission. The output of the JCat tool provided CAI value and GC content of the optimized sequence. Ideally, the CAI value should be greater than 0.8 or nearer to 1<sup>66</sup>, and the GC content should be in a range of 30 to 70%<sup>67</sup> to ensure that the vaccine has a potential translation, stability and transcription efficiency. Further, an *in-silico* cloning of a multi-epitope vaccine construct in *E. coli* pET-28(+) vector was performed using the SnapGene 4.2 tool<sup>68</sup>.

**Immune simulations of vaccine construct.** To characterize the real-life immunogenic profiles and immune response of the multi-epitope vaccine, C-ImmSim server<sup>69</sup> was utilized. Position-specific scoring matrix (PSSM) and machine learning are the two methods, based on which C-ImmSim predicts immune epitopes and immune interactions. It simultaneously simulates three different anatomical regions of mammals, i.e. Bone marrow, Thymus and tertiary lymphatic organs. To conduct the immune simulation, three injections at an interval of four weeks were administered with 1000 vaccine molecules per injection. The parameters like random seed, simula-

tion volume, and simulation step were kept as 12,345, 10  $\mu$ l and 1050, respectively. Several studies suggested that the minimum interval between two injections should be kept four weeks and therefore, the time steps followed for three injections were 1, 84 and 168, where each time step is equal to 8 h of real-life<sup>70</sup>. Since, several patients worldwide got recurrent infection of SARS-CoV-2, 12 consecutive injections were administered four weeks apart for the assessment of after-effects of vaccine exposure on SARS-CoV-2.

## Data availability

All relevant data is included in the manuscript.

Received: 9 April 2020; Accepted: 16 September 2020

Published online: 01 October 2020

## References

- Ceraolo, C. & Giorgi, F. M. Genomic variance of the 2019-nCoV coronavirus. *J. Med. Virol.* **92**(5), 522–528 (2020).
- World Coronavirus disease 2019 (COVID-19) Situation report—189, 27th July 2020.
- Zhang, Y. Z. & Holmes, E. C. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* **181**(2), 223–227 (2020).
- World-Health-Organization Coronavirus disease (COVID-19) outbreak. Available online: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> (accessed on 18th June 2020).
- Guo, Y. *et al.* The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—an update on the status. *Military Med. Res.* **7**, 11 (2020).
- Tai, W. *et al.* Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: Implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol. Immunol.* **17**, 613–620 (2020).
- World-Health-Organization Statement on the second meeting of the International Health Regulations(2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). Available online: [https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-\(2005\)-emergency-committee-regarding-the-outbreak-of-novelcoronavirus-\(2019-ncov\)](https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novelcoronavirus-(2019-ncov)) (accessed on 29 March 2020).
- Masters, P. S., Perlman, S. In *Fields Virology* Vol. 2 (eds Knipe, D. M. &Howley, P. M.) 825–858 (Lippincott Williams & Wilkins, 2013).
- Zhong, N. S. *et al.* Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet* **362**, 1353–1358 (2003).
- Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
- Fouchier, R. A. *et al.* Aetiology: Koch's postulates fulfilled for SARS virus. *Nature* **423**, 240 (2003).
- Ksiazek, T. G. *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1953–1966 (2003).
- Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. & Fouchier, R. A. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
- Lau, S. K. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl Acad. Sci.* **102**, 14040–14045 (2005).
- Ithete, N. L. *et al.* Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. *Emerg. Infect. Dis.* **19**, 1697–1699 (2013).
- Su, S. *et al.* Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.* **24**, 490–502 (2016).
- Forni, D., Cagliani, R., Clerici, M. & Sironi, M. Molecular evolution of human coronavirus genomes. *Trends Microbiol.* **25**, 35–48 (2017).
- Jia, Y. *et al.* Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. Preprint at <https://doi.org/10.1101/2020.04.09.034942v1> (2020).
- Meo, S. A. *et al.* Novel coronavirus 2019-nCoV: Prevalence, biological and clinical characteristics comparison with SARS-CoV and MERS-CoV. *Eur. Rev. Med. Pharmacol. Sci.* **24**(4), 2012–2019 (2020).
- Baud, D. *et al.* Real estimates of mortality following COVID-19 infection. *Lancet Infect. Dis.* S1473–3099(20)30195-X (2020).
- Woo, P. C. *et al.* Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J. Virol.* **86**, 3995–4008 (2012).
- Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**(8), 727–733 (2020).
- Schoeman, D. & Fielding, B. C. Coronavirus envelope protein: current knowledge. *Virol. J.* **16**, 69 (2019).
- Zhu, F. C. *et al.* Safety, tolerability, and immunogenicity of a recombinant adenovirus type-5 vectored COVID-19 vaccine: A dose-escalation, open-label, non-randomised, first-in-human trial. *Lancet* **395**(10240), 1845–1854 (2020).
- Abdelmageed, M. I. *et al.* Design of multi epitope-based peptide vaccine against E protein of human 2019-nCoV: An immunoinformatics approach. *Biomed. Res. Int.* **2020**, 2683286 (2020).
- Bojin, F., Gavriluc, O., Margineanu, M., Paunescu, V. Design of an Epitope-Based Synthetic Long Peptide Vaccine to Counteract the Novel China Coronavirus (2019-nCoV). Preprint at <https://www.preprints.org/manuscript/202002.0102/v1> (2020).
- Li, L. *et al.* Epitope-based peptide vaccine design and target site characterization against novel coronavirus disease caused by SARS-CoV-2. Preprint at <https://doi.org/10.1101/2020.02.25.965434v2> (2020).
- Enayatkhani, M. *et al.* Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: An in silico study. *J. Biomol. Struct. Dyn.* **2**, 1–16 (2020).
- Lee, C. H. & Koohy, H. In silico identification of vaccine targets for 2019-nCoV. *F1000Res.* **9**, 145 (2020).
- Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses.* **12**(3), 254 (2020).
- Grifoni, A. *et al.* A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe.* **27**(4), 671–680 (2020).
- Kozlova, E. E. G. *et al.* Computational B-cell epitope identification and production of neutralizing murine antibodies against Atoxlysin-I. *Sci. Rep.* **8**(1), 14904 (2018).
- Chauhan, V. *et al.* Designing a multi-epitope based vaccine to combat Kaposi Sarcoma utilizing immunoinformatics approach. *Sci. Rep.* **9**, 2517 (2019).
- Cui, J., Li, F. & Shi, Z. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
- Zhang, J. *et al.* Progress and prospects on vaccine development against SARS-CoV-2. *Vaccines (Basel).* **8**(2), E153 (2020).
- Pandey, R. K., Ojha, R., Mishra, A. & Prajapati, V. K. Designing B- and T-cell multi-epitope based subunit vaccine using immunoinformatics approach to control Zika virus infection. *J. Cell Biochem.* **119**(9), 7631–7642 (2018).
- Wlodawer, A. Stereochemistry and validation of macromolecular structures. *Methods Mol. Biol.* **1607**, 595–610 (2017).

38. Gralinski, L. E. *et al.* Allelic variation in the toll-like receptor adaptor protein Ticam2 contributes to SARS-coronavirus pathogenesis in mice. *G3* **7**(6), 1653–1663 (2017).
39. Mubarak, A., Alturaiki, W. & Hemida, M. G. Middle East respiratory syndrome coronavirus (MERS-CoV): Infection, immunological response, and vaccine development. *J. Immunol. Res.* **2**, 1–11 (2019).
40. Chen, L., Xiong, J., Bao, L. & Shi, Y. Convalescent plasma as a potential therapy for COVID-19. *Lancet. Infect. Dis* **20**(4), 398–400 (2020).
41. Doytchinova, I. A. & Flower, D. R. VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* **8**, 4 (2007).
42. Wilkins, M. R. *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112**, 531–552 (1999).
43. Geourjon, C. & Deleage, G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics* **11**, 681–684 (1995).
44. Yang, J. *et al.* The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **12**, 7 (2015).
45. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **7**, 1511 (2012).
46. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845 (2015).
47. Heo, L., Park, H. & Seok, C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, 384–388 (2013).
48. Lovell, S. C. *et al.* Structure validation by Ca geometry:  $\phi$ ,  $\psi$  and C $\beta$  deviation. *Proteins.* **50**, 437–450 (2003).
49. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **13**, 1605–1612 (2004).
50. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* **8**, 424 (2007).
51. Kim, Y. *et al.* Immune epitope database analysis resource. *Nucleic Acids Res.* **40**, 525–530 (2012).
52. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**(3), 394–406 (2018).
53. Bangov, I., Doytchinova, I., Dimitrov, I. & Flower, D. AllerTOP vol 2—A server for in silico prediction of allergens. *J. Mol. Model.* **20**, 2278–2284 (2014).
54. Manzalawy, Y. E. L., Dobbs, D. & Honavar, V. Predicting linear B-cell epitopes using string kernels. *J. Mol. Recogn. Interdiscip. J.* **21**, 243–255 (2008).
55. Ponomarenko, J. *et al.* ElliPro: A new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* **9**, 514 (2008).
56. Dhandra, S. K., Vir, P. & Raghava, G. P. Designing of interferon-gamma inducing MHC class-II binders. *Biol. Direct.* **8**, 30 (2013).
57. Bui, H. 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 Bioinformatics* **8**, 361 (2007).
58. Bui, H. H. *et al.* Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics* **7**, 153 (2006).
59. Lamiabile, A. *et al.* PEP-FOLD3: Faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Res.* **44**, 449–454 (2016).
60. Schneidman, D., Inbar, Y., Nussinov, R. & Wolfson, H. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* **33**, 363–367 (2005).
61. Mashiaev, E., Schneidman, D., Andrusier, N., Nussinov, R. & Wolfson, H. FireDock: A web server for fast interaction refinement in molecular docking. *Nucleic acids Res.* **36**, 229–232 (2008).
62. Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **27**(15), 2076–2082 (2011).
63. Wiederstein, M. & Sippl, M. J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35**, 407–410 (2007).
64. Kozakov, D. The ClusPro web server for protein-protein docking. *Nat. Protoc.* **12**, 255–278 (2017).
65. Grote, A. *et al.* JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.* **33**, 526–531 (2005).
66. Morla, S., Makhija, A. & Kumar, S. Synonymous codon usage pattern in glycoprotein gene of rabies virus. *Gene* **584**, 1–6 (2016).
67. Ali, M. *et al.* Exploring dengue genome to construct a multi-epitope based subunit vaccine by utilizing immunoinformatics approach to battle against dengue infection. *Sci. Rep.* **7**, 9232 (2017).
68. SnapGene software (from Insightful Science; available at [snapgene.com](http://snapgene.com)).
69. Rapin, N., Lund, O., Bernaschi, M. & Castiglione, F. Computational immunology meets bioinformatics: The use of prediction tools for molecular binding in the simulation of the immune system. *PLoS ONE* **5**, e9862 (2010).
70. Castiglione, F., Mantile, F., De Berardinis, P. & Prisco, A. How the interval between prime and boost injection affects the immune response in a computational model of the immune system. *Comput. Math. Methods Med.* **2012**(10), 1–9 (2012).

## Acknowledgements

We thank Zoological Survey of India, Kolkata for providing access to the computational support, working space and facilities to undertake the present study.

## Author contributions

A.S. and M.T. conceived the primary idea, drafted the skeleton and conceptualized it with the assistance of L.K.S. and K.C. A.S. undertook all primary analysis and wrote the first draft of the manuscript. M.T. validated and checked data quality. M.T. and L.K.S. revised the manuscript. All authors read and agreed to the contents of the manuscript. K.C. provided all the logistic support and administrative approval.

## Funding

There was no special funding to undertake this study.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-73371-y>.

**Correspondence** and requests for materials should be addressed to M.T.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020