Immuno-informatics approach for multi-epitope vaccine designing against SARS-CoV-2

Souvik Banerjee[1*], Kaustav Majumder[2], Gerardo Jose Gutierrez[3], Debkishore Gupta[4] and Bharti Mittal[5]

[1]Department of Microbiology, St. Xavier's College (Autonomous), Kolkata

[2]Department of Biotechnology, Indian Institute of Technology, Bombay

[3]Department of Microbiology and Cell Science, University of Florida

[4]Department of Clinical Microbiology and Infection Control, Calcutta Medical Research Institute and BM Birla Hospital, Kolkata

[5]Immuniteit Labs Pvt Ltd, Bangalore

*Corresponding author (E-mail: souvik97kol@gmail.com)

Abstract

The novel Corona Virus Disease (COVID-19) pandemic has spread a blaze of increasing fatality rates across the world. The dearth of potential vaccines has left the survival of mankind with doubts. The development of multi-epitope vaccine in this current situation could be a possible COVID treatment. We have designed a novel multi-epitope, multi-protein vaccine with different proteins of Severe Acute Respiratory Syndrome - Corona Virus -2 (SARS-CoV-2) with immuno-informatics approaches, which has been validated in silico to be stable and potential. It has been prepared with Cytotoxic T-cell (T_C) and Helper T-cell (T_H) binding epitopes overlapping with B-cell binding epitopes predicted for 6 proteins conserved among 4 different viral strains isolated across the world. Both the humoral and cell-mediated immune responses are ensured due to the presence of T cell and B-cell inducing epitopes along with interferon-gamma inducing epitopes present in the vaccine. The final vaccine construct comprises of an adjuvant at the N terminal, Cytotoxic T Lymphocyte and Helper T Lymphocyte epitopes. The construct showed potential antigenicity and was non-allergic. The molecular docking of the refined, validated tertiary structure model of the vaccine was performed with immune- stimulatory Toll Like Receptors (TLR), TLR-2,3,4. The study of binding energetics of the docked complexes revealed binding interactions of receptor with vaccine. The immune simulation of the vaccine even confirmed the initiation of elevated host immune responses. The efficient translation of the vaccine in an expression vector was confirmed with in silico cloning approach. Certainly, the development of such vaccine candidate could possibly be an effective therapy for COVID-19.

Keywords: SARS-Cov2, immuno-informatics, epitope, interferon - gamma and lymphocytes

Introduction

The entire world has been jeopardized due to the COVID-19 pandemic. The accelerated mortality rate has suffused at its extreme in due course of high infectivity rate of this disease. The World Health Organization (WHO) published situation report-180 on 18th July, 2020 which is the latest till the date of this writing, has recorded 13,876,441 total cases and 593,087 total deaths all over the world[1].The novel SARS-CoV-2 (as termed by the International Committee on Taxonomy of Viruses) was first reported to cause novel pneumonia in Wuhan seafood market, China and an outbreak was observed in city within fifty days of epidemic. The disease spread was set ablaze due to human to human transmission capability of the virus and it emerged to a pandemic. Phylogenetic analysis indicated its relation to SARS like bat viruses. Therefore, the trail signifies bat as one of the primary reservoirs. This data has evinced the dreadful situations ever encountered by mankind. This virus is a single-stranded RNA virus ranging from 26 to 32 kb in length. It belongs to the β group of coronaviruses. It spreads via respiratory droplets and the close contact with infected ones causes acute respiratory distress followed by pulmonary failure in the host. The lineage dates back to Severe Acute Respiratory Syndrome Corona Virus (SARS CoV) of 2003 and Middle East Respiratory Syndrome CoV (MERS CoV) of 2012 which exhibited similar characteristics of respiratory distress and alveolar injury[2]. The mode of infection involves exposure to aerosols carrying virus particles. It was also deciphered that the spike glycoprotein, a structural protein expressed in this virus, has binding affinity to the human Angiotensin Converting Enzyme 2 (ACE2) receptor cells and is responsible for attachment and entry into host cells[3]. There are other proteins encoded by viral genome which assist in viral replication and pathogenesis. The current scenario displays emergence of different strains being isolated from different parts of the world helping out the virus to reach out to more human population. It retards the development of a potential vaccine candidate. The solution to it is not inaccessible since a multi-epitope peptide vaccine could be a possible way to combat COVID. It could be advantageous to elicit immune response in comparison to single protein-target vaccine. The approach of designing such vaccine was entirely in silico. The epitopes predicted for the proteins to induce immune response were membrane glycoprotein, nucleocapsid phosphoprotein, envelope protein, Open Reading Frame (ORF) 6, ORF 7a and ORF 10. The selection of such proteins was due to the following reasons: i) these were found to be 100% conserved among isolated viral strains from India, Italy, USA and China while the spike protein does not show 100% conserved sequences among all the strains. ii) these play an important role in replication, assembly and infection of virus particles. The nucleocapsid phosphoprotein, membrane glycoprotein and envelope protein are structural proteins of the virus while ORF-6, ORF-7a and ORF-10 are the accessory proteins. The nucleocapsid protein is involved in the stability of viral RNA and processes of the viral replication. The envelope protein is a structural protein playing a significant role in maturation of the virus. The membrane glycoprotein forms a part of the outer viral coat and helps in determining the shape of the virus envelope. This protein has the ability to bind to all other structural proteins[4,5]. The ORF 6 protein acts as a signal blocker. It provides blockage to the signals, to immune system, sent from infected cell. The ORF 7a helps in release of more viral particles from infected cell. The ORF 10 though shows homology in different strains but its function still remains unexplored.

Toxicity as well as resistance to inhibition associated with drug-based therapy is often a hurdle to overcome so peptide vaccines designed with immuno-informatics approach are more optimized and could be developed within a short period of time. It also paves the way for cost-effective vaccine preparations. Peptide vaccines are even advantageous than conventional vaccines in providing highly targeted immune response and being devoid of allergenicity. Conventional vaccines might consist of several antigenic epitopes of which only few are required to trigger immune response. The peptide vaccine on the other hand consists of desired T cell as well as B cell induced immune response. It also ensures the vaccine to get rid of allergenic sequences[6,7]. The most important aspect to be addressed with this design of multi-protein and multi-epitope developed with immunodominant different strain conserved proteins includes broad spectrum immune response initiated against multiple strains of the virus. The vaccine construct consists of interferon γ, T-cell and B-cell inducing epitopes which confirms the initiation of both humoral and cell mediated immune response in host. The final vaccine construct includes Cytotoxic T Lymphocyte, Helper T Lymphocyte epitopes and an adjuvant (β subunit of cholera toxin) at the N-terminal. The adjuvant addition aids as an immune response booster.

The entire synthesis of the vaccine commenced from prediction of B-cell and T-cell epitopes in the proteins selected. The proteins were selected on the basis of their homology and sequences conserved among different viral strains selected. This was achieved with multiple sequence alignment. The multi-epitope construct was prepared with combination of T-cell epitopes overlapping with B-cell epitopes. The epitopes

were joined with linkers and this construct was found to be antigenic and non-allergenic. It was subjected to prediction of interferonγ inducing epitopes. This followed the secondary structure prediction and tertiary structure prediction. The tertiary structure was subjected to refinement along with validation. The vaccine interaction with target immune cell receptors is necessary and the Toll like receptors (TLR) play significant role in activation of innate immunity after recognition of pathogen associated molecular patterns (PAMPs) of viruses and other invading pathogens. Moreover, the TLRs play active role in triggering innate immunity and synchronize adaptive immune response. TLR-2 and TLR-4 are involved in recognition of viral structural proteins followed by production of cytokines and even in SARS CoV, they have been found to induce effective immune responses. Certain studies on SARS CoV and MERS CoV even indicated the noteworthy role of TLR 3 in generating protective responses against the virus[8-11]. So, the vaccine was docked with the immunological receptors: TLR-2, TLR-3 and TLR-4. Molecular docking simulations were performed to visualize and evaluate the binding interactions. Finally, the codon optimization through in silico approach was performed for it to ensure its maximal production in host expression system. The ultimate tool included immune simulation to ensure elevated immunological response after vaccine introduction into the body of the host. These advanced aspects of computational tools have surely revolutionized vaccine therapeutic studies.

Protein	B-cell Epitopes	CTL Epitopes	HTL Epitopes
Nucleocapsid Phosphoprotein	PSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQHGKEDL		ASWFTALTQHGKEDL (13)
	NTNSSPDDQIGYYRRATRRIRGGDGKMKDLS		QIGYYRRATRRIRGG (14)
	NNAAIVLQLPQGTTLPKGFYA		NNAAIVLQLPQGTTL (15)
	IRQGTDYKHWPQIAQFAPSASAFF	GTDYKHWPQ (1)	HWPQIAQFAPSASAF (16)
	KKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA		QQTVTLLPAADLDDF (17)
	DQVILLNKHIDAYKTF	LLNKHIDAY (2)	
	KMKDLSPRWYFYYLGT	LSPRWYFYY (3)	
Membrane Glycoprotein	SYYKLGASQRVAGDSG		SYYKLGASQRVAGDS (18)
	TVEELKKLLEQWNLVI	LLEQWNLVI (4)	
	ACLVGLMWLSYFIASF	LVGLMWLSY (5)	
	GDSGFAAYSRYRIGNY	YSRYRIGNY (6)	
Envelope Protein	NVSLVKPSFYVYSRVK	VSLVKPSFY (7)	
ORF6	ILLIIMRTFKVSIWNL		ILLIIMRTFKVSIWN (19)
	KVSIWNLDYIINLIIK	NLDYIINLI (8)	
	IKNLSKSLTENKYSQL	LTENKYSQL (9)	
ORF7a	GVKHVYQLRARSVSPKLFIR		VKHVYQLRARSVSPK (20)
	LALITLATCELYHYQE	ITLATCELY (10)	
	IRQEEVQELYSPIFLIVAAIVFITL	RQEEVQELY (11)	
ORF10	SRNYIAQVDVVNFNLT	QVDVVNFNL (12)	

Table 1. Selected CTL and HTL epitopes overlapping with B-cell epitopes. The numbers in parentheses indicates the sequence of epitopes in the vaccine construct from N terminal to C terminal.

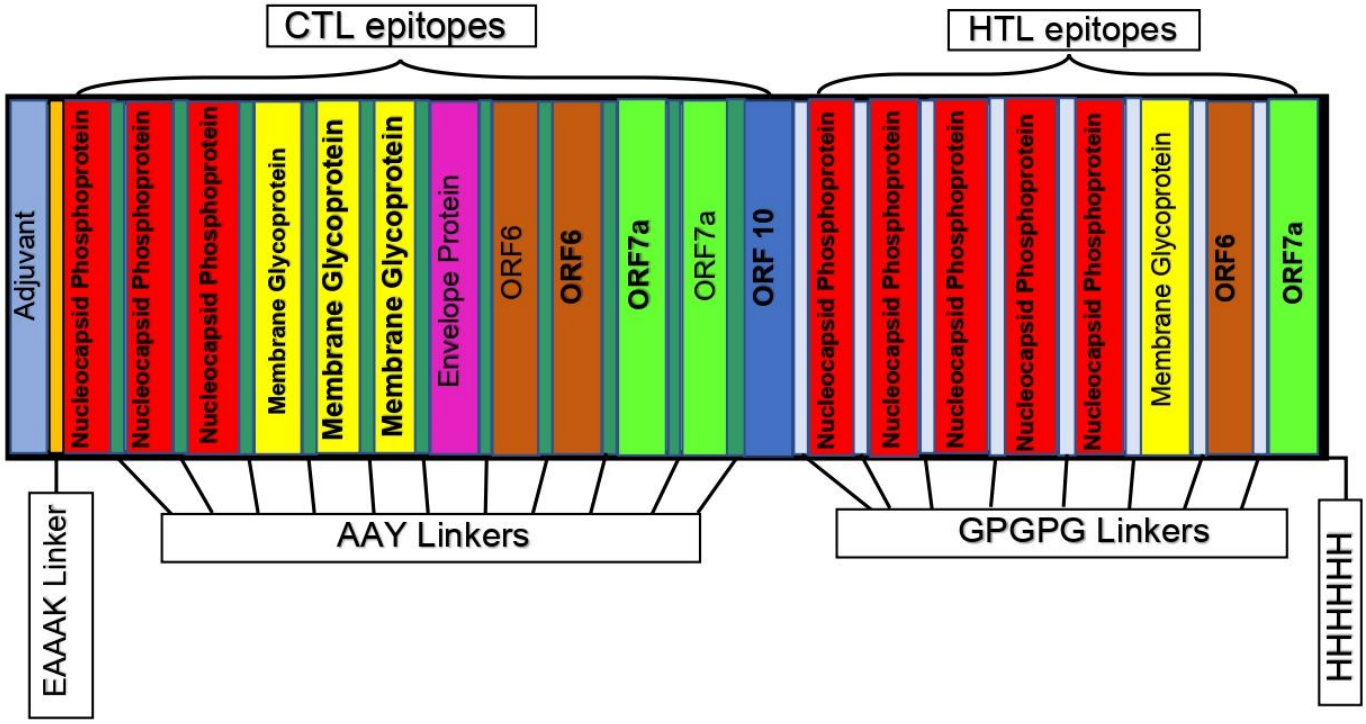

Figure 1. Schematic representation of the final multi-epitope vaccine construct: A 436 amino acid long construct with adjuvant at the N-terminal end linked with EAAAK linker and 6x-His tag added at the Carboxy terminus end. CTL epitopes linked with AAY linkers and HTL epitopes linked with GPGPG linkers.

Results

Selection of viral protein sequences for vaccine construct: Six viral protein sequences were found to be conserved among all the proteins of SARS-CoV 2 strains isolated in India, USA, Italy, Wuhan (Supplementary Fig. 1). Multiple sequence alignments performed for all the proteins helped to achieve such results. These proteins, selected for vaccine construct, were three structural proteins - nucleocapsid phosphoprotein, membrane glycoprotein, envelope protein and three accessory proteins - ORF6, ORF7a and ORF10. The structural proteins are primarily responsible for host cell replication, maturation as well as assembly of virus particles while the accessory proteins assist in viral replication, release of newly formed virus particles. The above-mentioned protein sequences were subjected to further analysis.

Prediction of B-cell epitopes: Linear B-cell epitopes of all the structural and accessory proteins were predicted from BepiPred 2.0, Bcepred, ABCPRED and SVMTriP web servers. The epitope length was set to 16 amino acids in ABCPRED and 20 amino acids in SVMTriP server. 148 epitopes of all the proteins were predicted by all the four servers. 19 epitopes were found to be overlapping with the T-cell epitopes (having good prediction scores) so these were selected (Supplementary table 1). Among the 19 epitopes, 7 of them were from nucleocapsid phosphoprotein, 4 were from membrane glycoprotein, 1 each from envelope protein and ORF10, 3 each from ORF6 and ORF7a (Table 1).

Cytotoxic T Lymphocyte(CTL) prediction: CTL epitopes of A1 supertype were predicted with the NetCTL 1.2 tool with input sequence and 0.15 default weight on C terminal cleavage, 0.05 default weight on TAP transport efficiency along with 0.75 default threshold for epitope identification. 29 epitopes of all the epitopes for 6 proteins predicted by the server were marked with '<-E' identifier in output results. The '<-E' marked were assigned as 'Identified MHC ligands' by the server and they even showed high prediction scores. These epitopes were listed down. 12 of 29 epitopes were found to be overlapping with B-cell epitopes and these were selected for vaccine construct. 3, 3, 1, 2, 1 and 2 epitopes of 12 epitopes were of nucleocapsid phosphoprotein, membrane glycoprotein, envelope protein, ORF6, ORF 10 and ORF7 proteins respectively (Supplementary table 1).

Helper T Lymphocyte(HTL) prediction: 15-mer HTL epitopes having loci in Human Leukocyte Antigen(HLA)-DR, HLA-DP, HLA-DQ were predicted with the NetMHCII 2.3 tool. The input protein sequences were submitted with default peptide length of 15 amino acids along with the default threshold % rank for strong binder (SB) as well as weak binder (SB) set to 2 and 10 respectively in the server. 73 'SB' assigned

high affinity epitopes of all proteins, among all the epitopes predicted by the server for HLA-DR, were selected as probable epitopes. The epitopes predicted by the server for HLA-DQ and HLA-DR were same in sequence and number, so epitopes for any one of them were considered. 8 of these 73 epitopes were found to be overlapping with B-cell epitopes and these were selected for vaccine construct. Among these 8 epitopes, 1, 1, 1 and 5 epitopes were of ORF7a, ORF6, membrane glycoprotein and nucleocapsid phosphoprotein respectively (Supplementary table 1).

Multi-epitope vaccine sequence construction: A linear vaccine construct should be capable of satisfying the following criteria i.e. it should possess overlapping HTL and CTL epitopes, it must be immunogenic, antigenic, not an allergen, show high affinity to HLA alleles. Firstly, the overlapping CTL and HTL epitopes whose sequences in turn were found to be overlapping with the B-cell epitopes were selected for final vaccine construct. A total of 12 CTL epitopes and 8 HTL epitopes were selected and merged by AAY and GPGPG linkers respectively. The β subunit of cholera toxin of 124 amino acids (MIKLKFGVFFTVLLSSAYAHGTPQNITDLCAEYHNTQIHTLNDKIFSYTESLAGKREMAIITFKNGATFQVEV PGSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVWNNKTPHAIAAISMAN) was added as an adjuvant at N-terminal of the vaccine construct with EAAAK linker to boost the immune response. After addition of linkers and adjuvant the final vaccine construct was found to be of length 436 amino acids (Fig. 1).

Interferon γ (IFN- γ) inducing epitope prediction: The IFN-γ inducing epitopes were predicted from the IFN epitope server in the vaccine construct. A total of 428 potential epitopes were predicted by this tool among which 152 showed positive results. This server executes such prediction with the usage of MERCI software.

Allergenicity and antigenicity prediction of the vaccine candidate: The antigenicity of the final vaccine construct (including the adjuvant) predicted by the VaxiJen 2.0 server was found to be 0.5448 with a virus model at a threshold of 0.4 and with ANTIGENpro it was found to be 0.777. The VaxiJen score>0.4 indicates it to be probable antigen and even the predicted antigen probability score (0.777) in ANTIGENpro indicates good antigenicity of construct. The vaccine candidate was even predicted to be non-allergenic on the AllergenFP as well as AllerTOP v. 2.0 servers.

Physicochemical analysis of vaccine constructs: The ProtParam server was used to calculate the physicochemical properties of the vaccine construct. The molecular weight of construct was calculated as 48.14kDa. The theoretical pI was found to be 9.11 which indicates it to be slightly basic in nature. The instability index was found to be 37.77 which confirms it as a stable protein since a score greater than 40 denotes unstable protein. The total no. of positively charged residues and negatively charged residues present in vaccine were 37 and 28 respectively. The estimated half-life of the construct was found to be 30 hours, greater than 20 hours and 10 hours in mammalian reticulocytes (in vitro), yeast (in vivo) and Escherichia coli (in vivo) respectively. The aliphatic index was found to be 84.50 and Grand average of hydropathicity (GRAVY) was found to be -0.172. The aliphatic index score ensures its thermo-stability while the negative GRAVY score confirms its hydrophilicity and it even casts light on its solubility. The Recombinant protein solubility prediction tool predicted 100% solubility when overexpressed in *Escherichia coli.*

Secondary structure prediction: The multiepitope chimeric peptide vaccine's secondary structure was predicted with PSIPRED tool. It involved 172, 85 and179 amino acids in formation of alpha-helix, beta strand and coil respectively. So, the peptide contains 39% alpha helix, 19% beta strand and remaining 41% as coil (Fig. 2) (Supplementary Fig. 2). The RaptorX Property server was even utilized to explore solvent accessibility and disordered domain residues in the protein. 42%, 27%, 30% of the residues were found to be exposed, medium exposed and buried respectively. 8% of the residues were found to be in disordered domains.

Tertiary structure modelling: The tertiary structure of vaccine was modelled with the I-TASSER server. It predicted five tertiary structure models of the chimeric protein based on 10 threading templates of which 1ltrA, 4kxfK, 3chbD, 4l6t and 6b0n were of top five Protein Data Bank(PDB) hits. All the 10 templates (PDB hits) showed good alignment as per their Z-score values (ranging from 1.82 to 5.40). The five predicted models had C-score values in the range -4.08 to -0.64. Since higher score indicates higher confidence the model with higher C-score was selected for further model refinements. Moreover, this model had an estimated TM-score of 0.63±0.13 which is >0.5 and indicates a model of correct topology. These threshold values are independent of length of protein. The TM-score of <0.17 denotes random similarity. The protein model had an estimated Root Mean Square Deviation (RMSD) value of 8.4±4.5Å. The RMSD value is often sensitive to local error so TM-score has been estimated (Fig. 3A).

Tertiary structure refinement: The predicted tertiary (3D) structure of crude vaccine construct was further subjected to model refinement using GalaxyRefine server. GalaxyRefine generated 5 models after refinement, out of which model 1 (Fig. 3B) was selected for having best scores compared to others depending on various parameters comprising GDT-HA (0.948), RMSD (0.433), MolProbity (2.492), Clash score (19.8), Poor rotamers (1.5) and 88.9% favoured region in Ramachandran plot.

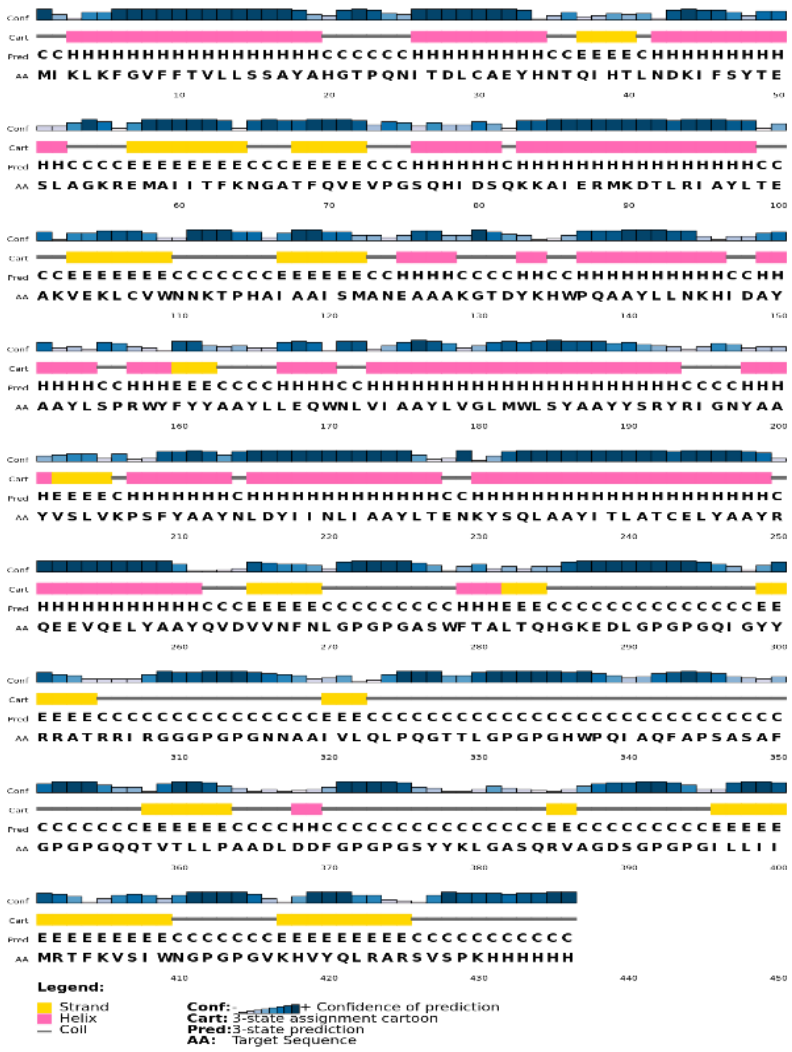

Figure 2. Schematic representation of the secondary structure prediction of the multi-epitope vaccine construct. The Secondary structure prediction results depict the respective contents of α-helix (39%), β-strands (19%) and coils (41%).

Tertiary structure validation: The validations of the predicted tertiary structures were performed with RAMPAGE server performing Ramachandran plot analysis of the modelled vaccine construct. The results showed 90.3%, 7.1% and 2.5% of the residues of the protein in favored, allowed and outlier regions respectively (Fig. 3C). The favoured region score is almost consistent with the score of refined models selected in GalaxyRefine. The quality and potential errors' determination was inevitable so the ProSA-web and ERRAT servers were utilized respectively. The chosen refined model showed up an overall quality factor of 80.73% in ERRAT server (Fig. 3E) while ProSA-web gave a Z-score of −5.14 for the input vaccine model. This Z score lies within the range of commonly found native proteins of comparable size so the purpose of validation gets fulfilled with the obtained results (Fig. 3D).

Prediction of discontinuous B-cell epitope: ElliPro server which predicts epitopes on the basis of tertiary structures was the medium of prediction of Discontinuous B-cell epitopes. The PDB file of vaccine was provided as input and the epitope prediction parameters, minimum score and maximum distance were set to default score 0.5 and 6 respectively. Ten epitopes were predicted from the server from which the epitope (of

75 residues) with maximum score 0.755 was selected as a discontinuous epitope (Table2). 14 linear B-cell epitopes were even predicted by the server (Table 3).

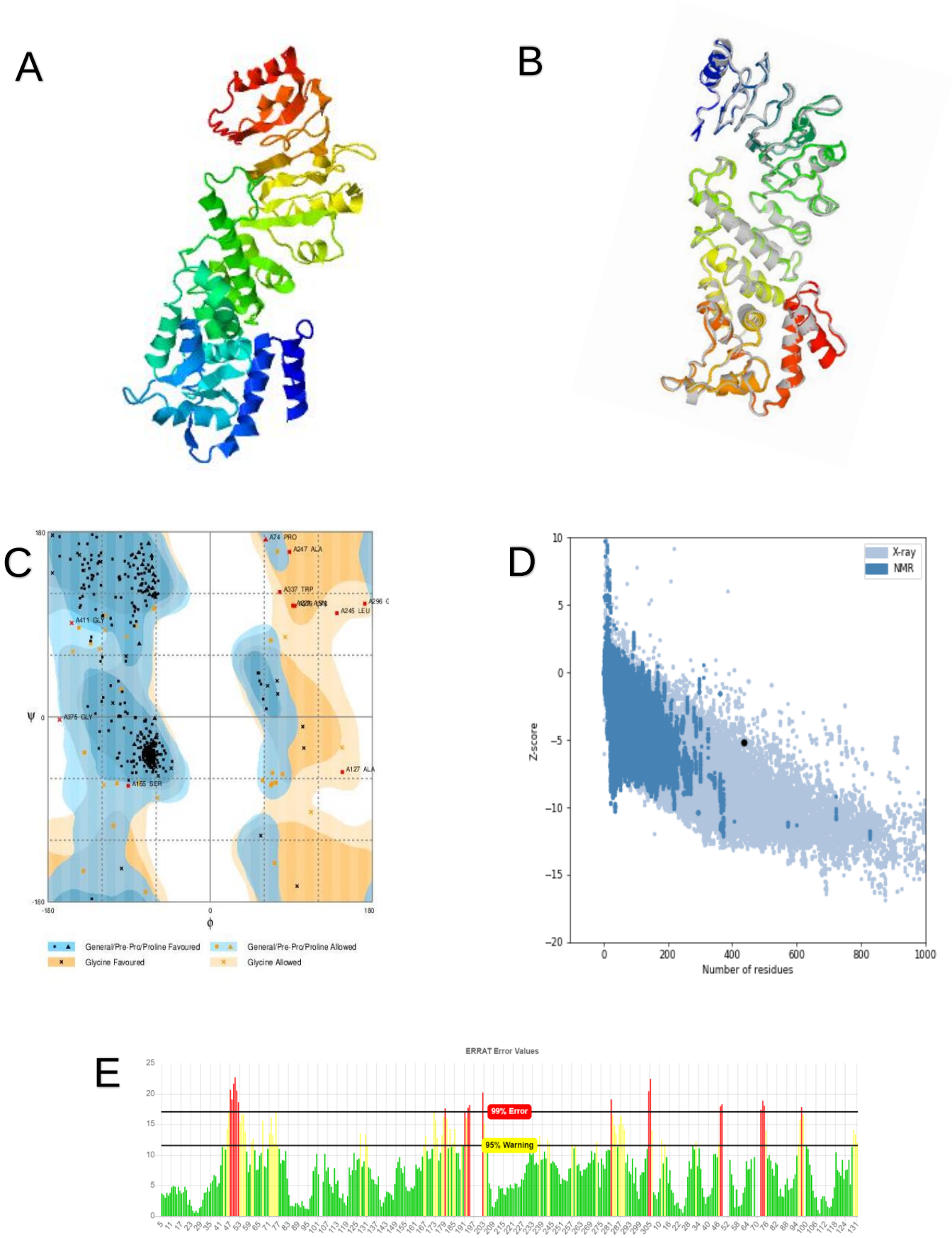

Figure 3. Modelling, Refinement and model validation of the multi-epitope vaccine construct **(A)** The 3D model of the vaccine construct. **(B)** Tertiary structure of the vaccine after model refinement. **(C)** Validation of multi-epitope vaccine tertiary structure by Ramachandran plot analysis which depicts 90.3% residues in

favored region, 7.1% residues in allowed region and 2.5% residues in outlier region. **(D)** Validation of model with ProSA-web providing a z score of -5.14. **(E)** Validation by ERRAT with a score 80.73.

Discontinuous Epitopes	Number of Residues	Score
VF (8-9), TVLLSSAY (11-18), E (32), NTQIHTLNDKIFSYTESLA (35-53) K (55), NGATFQVEVPGSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVW (65-109)	76	0.755
AHGTPQN (19-25)	7	0.719
NLGPG (269-273), GYYRRATRRIR (298-308), PQGTTLGPGPGHWPQ (325-339), Q (342), PAADLDDF (363-370), YYKLGASQRVAGDSGPGPGILLIIMRTFKVSIWNGPGPGVKHVYQLRARSVSPKHHHHHH (377-436)	100	0.702
AKGTDYKHWPQ (128-138), AY (140-141), NK (144-145), L (167), Q (169)	17	0.684
PSASAFGPGPG (345-355)	11	0.631
RIG (194-196)	3	0.615
YSQLA (231-235)	5	0.566
MIK (1-3)	3	0.531
E (252), VQE (254-256), PGA (274-276), FT (279-280), TQHGK (283-287)	14	0.518
SLVK (203-206)	4	0.505

Table 2. Discontinuous B-cell epitopes provided with prediction scores along with number of residues.

Linear B-cell Epitope	Number of Residues	Score
VFFTVLLSSAYAHGTPQN (8-25)	18	0.763
KLGASQRVAGDSGPGPGILLIIMRTFKVSIWNGPGPGVKHVYQLRARSVSPKHHHHHH (379-436)	58	0.764
NGATFQVEVPGSQHIDSQKKAIERMKDTLRIAYLTEAKVEKLCVWNN (65-111)	47	0.762
AKGTDYKHWPQAAY (128-141)	14	0.725
YYRRATRRIR (299-308)	10	0.726
NTQIHTLNDKIFSYTESL (35-52)	18	0.704
PQGTTLGPGPGHWPQ (325-339)	15	0.67
PSASAFGPGPG (345-355)	11	0.631
SQLA (232-235)	4	0.584
PAADLDDFG (363-371)	9	0.55
MIKLK (1-5)	5	0.54

TQHGK (283-287)	5	0.529
YRIG (193-196)	4	0.519
SLVK (203-206)	4	0.505

Table 3. Linear B-cell epitopes predicted by ElliPro with prediction scores along with number of residues

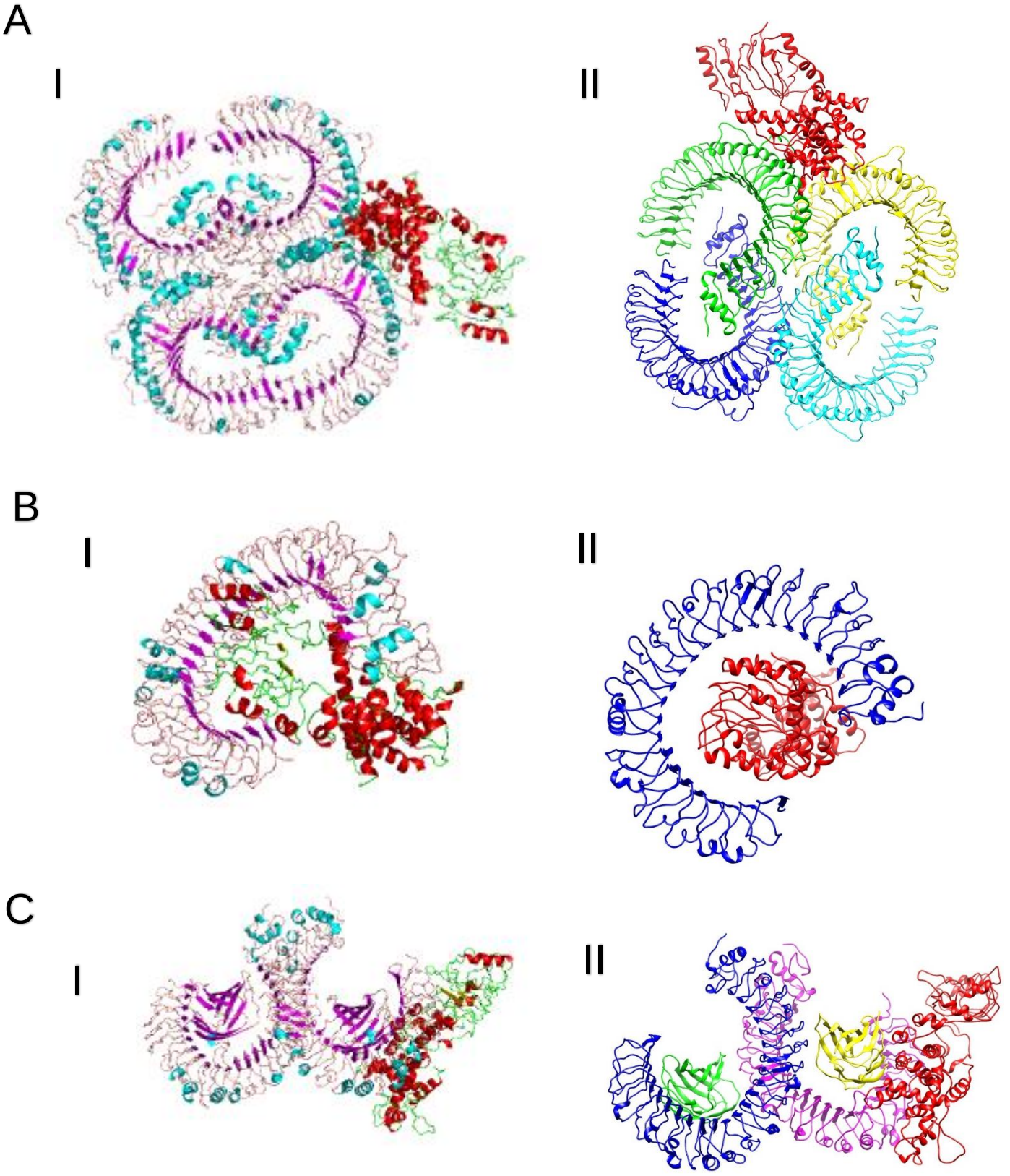

Figure 4. 3D visualization of **(A)** TLR2-vaccine docked complex **(Ai**) in ClusPro **(Aii)** in UCSF Chimera, **(B)** TLR3-vaccine docked complex **(Bi)** in ClusPro **(Bii)** in UCSF Chimera and **(C)** TLR4-vaccine docked complex **(Ci)** in ClusPro **(Cii)** in UCSF Chimera

Molecular docking of final vaccine construct with immunological receptor: ClusPro online protein-protein docking server was used to perform molecular docking between refined model and immune receptors TLR-2 (PDB ID- 6NIG), TLR 3 (PDB ID- 2A0Z) and TLR 4 (PDB ID- 4G8A). The output results displayed 10 models of docked complexes and 30 clusters for each docked complex ranked according to cluster members. The weighted scores of energies of the clusters are also provided. Cluster 1 of TLR-2 – vaccine docked complex (Fig. 4A) with 32 members having lowest energy of -1197.2 and Cluster 0 of TLR-3 – vaccine docked complex (Fig. 4B) with 97 members (highest no. of members) having lowest energy of -1524.4 were selected for further analysis. Cluster 22 of TLR-4 docked complex showed best weighted score of lowest energy but it had only 12 members so cluster 2 (Fig. 4C) with 48 members and having lowest energy of -1207.7 was selected. The 3 docked complexes were further visualized with UCSF Chimera.

Analysis of binding affinity of docked complexes: The study of binding energetics is inevitable for biological complexes. The binding affinity of a complex, or the Gibbs free energy (ΔG) i.e. binding free energy determines the probability of occurrence of interactions at specific cellular conditions. So, the PRODIGY web server was used to analyse the binding affinity of the 3 docked complexes. The input accepts PDB (Protein Data Bank) file of the docked complexes along with the interactor chains of receptor and ligand. The temperature was set to default value of 25°C. The ΔG values for the vaccine-TLR2, vaccine-TLR3, vaccine-TLR4 complex were found to be − 18.9 kcal mol^{-1}, − 19.1 kcal mol^{-1}, − 18.0 kcal mol^{-1}. The results evinced energetically feasible docking, as depicted by the negative values of Gibbs free energy. The dissociation constants of the complexes, number of interfacial contacts per property and non-interacting surface per property are shown in Table 4.

Docked Complexes	**Gibb's Free Energy (kcal/mol)**	**K_d (M) at 25° C**	**Number of Interfacial Contacts (ICs) per property**						**Non- Interacting Surface (NIS) per property**	
			ICs charged - charged	ICs charged-polar	ICs charged -apolar	ICs polar-polar	ICs polar-apolar	ICs apolar-apolar	NIS charged (%)	NIS apolar (%)
Vaccine-TLR2	-18.9	1.3E-14	10	18	49	4	33	40	26.42	31.62
Vaccine-TLR3	-19.1	1.0E-14	28	25	67	2	17	19	21.29	36.13
Vaccine-TLR4	-18.0	6.1E-14	7	6	34	9	42	33	23.20	35.03

Table 4. Predicted results of binding affinity (Gibbs free energy), dissociation constant, number of interfacial contacts per property, non-interacting surface per property by PRODIGY web server

Codon optimization and in silico cloning of the vaccine construct: In silico cloning of the vaccine construct is of immense noteworthy for its expression in *Escherichia Coli* expression system. As a result, the codon optimization of construct is inexorable as per usage in expression system in order to assure efficient translation in chosen expression system. The Java Codon Adaptation Tool (JCat) was utilized for codon optimization of the final vaccine construct for maximal protein expression in *Escherichia Coli* (K-12 strain). The length of the generated cDNA sequence after codon optimization was of 1308 nucleotides. The Codon Adaptation Index (CAI) of the optimized nucleotide sequence was found to be 0.96 which indicates high expression of gene since the optimum range is 0.8-1.0 and the average GC content of the adapted sequence

was 53.98% which also indicates the possibility of good expression of the vaccine candidate in the host system since the optimal percentage of GC content lies in the range of 30-70%. Finally, the design of the recombinant plasmid was accomplished in silico by inserting the adapted codon sequences into pET-28a (+) vector using SnapGene software. This study ensured effective cloning strategy of the multi-epitope vaccine construct (Fig. 5).

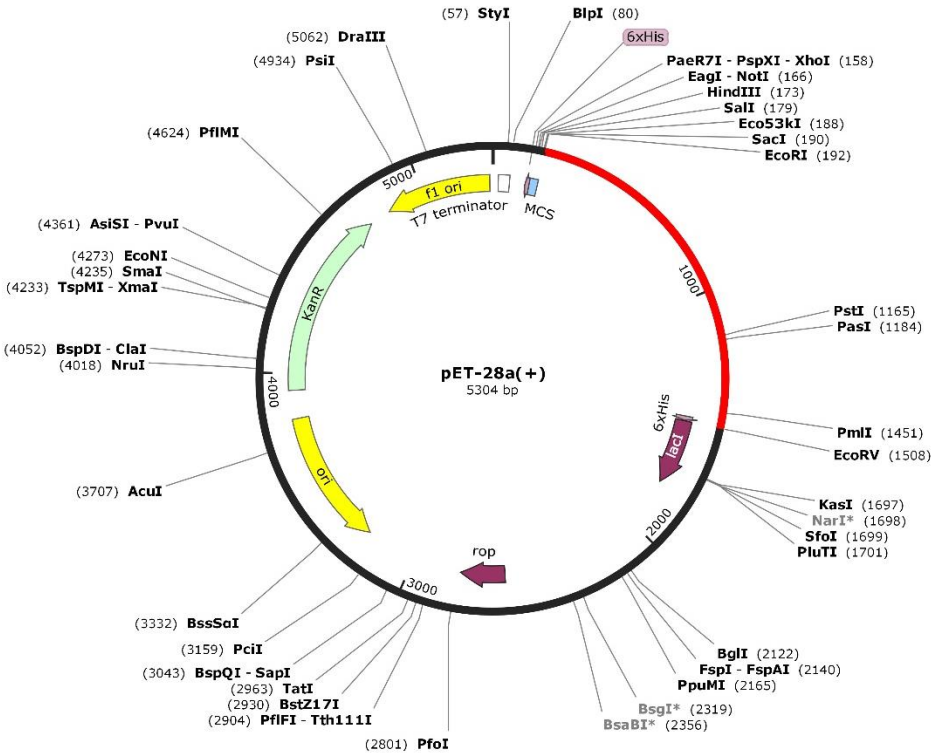

Figure 5. In silico restriction cloning of the final vaccine construct sequence into the pET28a (+) expression vector. The red part represents the gene encoding for the vaccine and the black circle represents the vector backbone.

Immune Simulation: The immune simulation of the vaccine was performed with C ImmSim server. The results depicted secondary and tertiary immune response (IgG1, IgG2, IgG + IgM) to be greater than primary immune response (IgM). The antigen concentration decreases and immunoglobulin concentration (IgM, IgG1+ IgG2, IgG + IgM) increases after vaccine injection. Long lasting B cells exhibit isotype switching ability and development of memory cells. T_H and T_C cell responses are found high with corresponding memory development. The pre-activation of T_C cell response is found during vaccination. Natural Killer cells and Dendritic Cells show consistent response throughout. High levels of macrophage activity are also indicated. 12 doses of injections consistently given 4 weeks apart show high levels of IFN-gamma and Interleukin (IL)-2 elicited which are consistent with the prediction of IFN-gamma epitopes in the vaccine. Two aspects of input were implemented for immune simulation. One method incorporated was that after vaccination, the live replicating virus was simulated at day 366. No antigenic surge was present and the antigen was contained immediately. It reveals the presence of protective antibodies. The second aspect was that without prior vaccination, live replicating virus was simulated at around similar day (366). The antigenic surge present in this case indicates the failure to contain the virus in spite of presence of a mild immune response initiated (Fig. 6).

Discussion

The development of a multi-epitope vaccine candidate in silico plays a more significant role in therapeutics as compared to single epitope since it facilitates to identify antigenic epitopes which could initiate targeted immune response. Moreover, the immune-informatics approach of developing such candidate is a cost effective and time-saving method in order to predict potential antigenic epitopes for the vaccine construct[12-13]. Our aim was to develop a multi-epitope, multi-protein vaccine construct which could trigger both humoral response as well as cell-mediated immune response when injected in the host since the construct contains both T_C-cell and T_H-cell epitopes overlapping with B-cell epitopes. This designed vaccine construct would surely be able to provide double protection in targeted immunological aspects. The vaccine candidate

consists of epitopes for multiple viral proteins which could definitely provide noteworthy defence against infection. The design of this construct commenced with selection of 6 proteins with conserved sequences among 4 different SARS-CoV-2 strains. The T-cell epitopes and B-cell epitopes were predicted for these proteins. The CTL epitopes and HTL epitopes which were found to be overlapping with B-cell epitopes predicted for these viral proteins were joined by linkers and utilized to construct this vaccine. An Adjuvant was even added to the N terminal of this construct to boost the immune response. The 6x-Histidine(His) tag was added at the C-terminal for purification purpose. The vaccine was found to be non-allergic and antigenic. Interferon-γ epitopes were even identified in the vaccine which confirmed the potential of vaccine to initiate functioning of Interferon-γ exhibiting both immunomodulatory and immunostimulatory actions. The secondary structure of the vaccine was predicted and even the tertiary structure was modelled. This 3D model was further refined and validated with different servers. The vaccine was docked with TLR-2, TLR-3 and TLR-4 receptors. These receptors are able to trigger innate immune response. The binding affinity related parameters of docked complexes were further analysed. The binding free energy of TLR3-vaccine complex was found to be the most negative of all docked complexes which confirmed most stable binding and even the dissociation constant was found to be lowest of all complexes. The in-silico cloning approach was performed to ensure maximal expression of the vaccine in an expression vector. The immune simulation was finally performed for the evaluation of immunogenic profile of the vaccine. Antigen is contained immediately after infection with live virus. Antigen concentration is found to be decreasing immediately due to the presence of memory cells since prior vaccination helped in memory cell development. The memory B cells, memory helper T and memory cytotoxic T cells are triggered in presence of the live virus which ultimately contain the virus. On the other hand, without prior vaccination memory cells being absent caused antigenic surge for a much longer amount of time after the injection of the live virus.

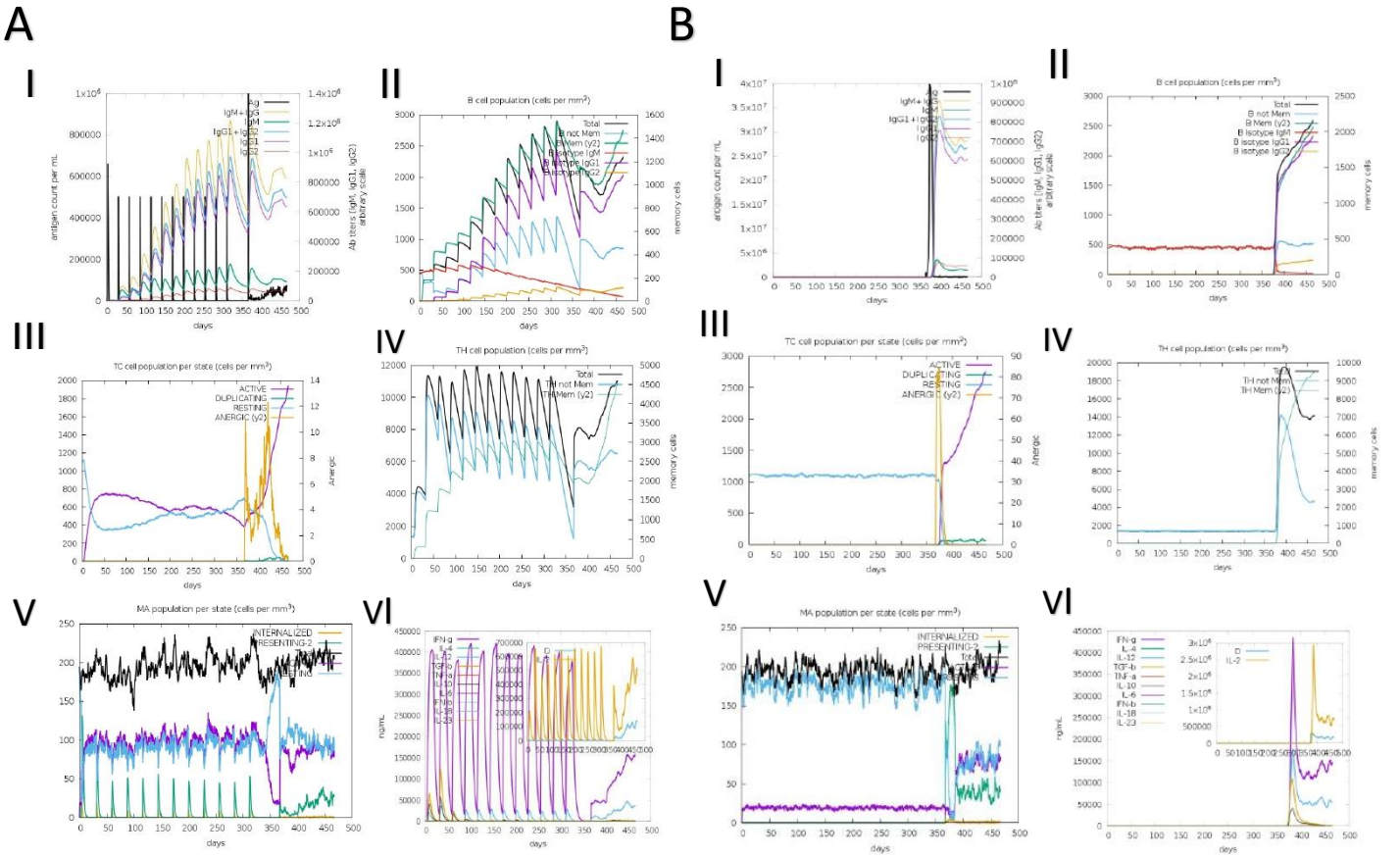

Figure 6. (A) 12 doses of vaccine injections were given for almost 15 months and a live replicating virus was injected at around day 366 **(Ai)** indicates the increase of antigen concentration and relative antibodies responses. Live-replicating virus, injected two months after last vaccine dose is contained immediately due to the production of protective immunoglobulins highlighting the efficiency of the vaccination. **(Aii)** represents the count of B-cell population. **(Aiii-iv)** indicating the activation of cytotoxic T-cells and helper T-cells. Increased number of memory helper T-cells proves the efficacy of the vaccine. (Av) Macrophage activation is shown **(Avi)** High levels of IFN-gamma, Tumor Necrosis Factor(TNF)-b, Interleukin(IL)-12 and IL-2

indicates good immune response. **(B)** comparison by a control experiment where the live virus is injected at the same time period (at around day 366) but without prior vaccination in this case. **(Bi)** Antigenic surge for a longer period of time indicates the absence of any memory cells without prior vaccination. **(Bii-vi)** absence of strong immune response due to lack of vaccination could not contain the antigenic load which further proves the efficiency of the vaccine construct

Methodology:

Accession to viral protein sequences: The sequences of all the proteins of SARS-CoV 2 strains isolated from India, Italy, USA and Wuhan were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/genbank/) with Reference Sequence IDs MT050493.1, MT066156.1, MN985325.1 and NC_045512.2 respectively[14].

Identification of proteins for vaccine development: All the protein sequences of different strain isolates were subjected to Multiple Sequence Alignment using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) which incorporates progressive approach, seeded guide trees as well as HHalign package for generating both swift and accurate alignments of three or more input sequences[15]. Alignment results showed 100 % conserved sequences of the proteins, Nucleocapsid phosphoprotein, Membrane glycoprotein, Envelope protein, ORF6, ORF7a and ORF10. All these proteins play active role in viral infection and pathogenesis.

B-cell epitope prediction: The B-cell epitopes were predicted for selected proteins utilizing BepiPred2.0 (http://www.cbs.dtu.dk/services/BepiPred/) which relies on random forest algorithm for discerning epitope database annotated from various antigen-antibody complexes[16].This server predicts by taking only FASTA sequence input of the protein. Apart from this server, the epitopes were predicted from three more servers viz. Bcepred (https://webs.iiitd.edu.in/raghava/bcepred/bcepred_submission.html) which is solely based on physico-chemical properties of dataset created with epitope sequences acquired from Bcipep and SWISS-PROT database[17]. This server takes amino acid sequence of the protein as input. The default threshold is set for all the parameters, hydrophilicity (2), accessibility (2), exposed surface (2.4), antigenic propensity (1.8), flexibility (1.9), turns (1.9), polarity (2.3) and combined (1.9) in the input. The next server is ABCPRED(https://webs.iiitd.edu.in/raghava/abcpred/ABC_submission.html) which predicts on the basis of artificial neural network methodology using fixed length patterns[18]. The 16mer epitope length with default threshold 0.51 was set and the default overlapping filter was set on. This server also accepts only the protein amino acid sequence as input. The final one was SVMTriP (http://sysbio.unl.edu/SVMTriP/prediction.php) where a new method to predict antigenic epitope with sequence input from IEDB database is implemented. This server utilizes Support Vector Machine (SVM) combining the Tri-peptide similarity and Propensity scores (SVMTriP) in order to achieve the better prediction performance[19]. This server takes FASTA sequence of the protein as input. The default epitope length was set to 20 aminoacids.

Cytotoxic T lymphocytes (CTL) epitope prediction: The prediction of CTL for peptide vaccine design is inevitable so the epitopes were acquired from a freely accessible NetCTL 1.2 **(**http://www.cbs.dtu.dk/services/NetCTL/). This tool incorporates prediction of MHC class I binding peptides, proteasomal C-terminal cleavage, and TAP (Transporter Associated with Antigen Processing) transport efficiency. Although CTL epitopes remain restricted to 12 MHC class I supertypes in the server, only the A1 supertype was utilized. MHC class I binding and proteasomal cleavage are performed using artificial neural networks. A weight matrix is the mode to determine TAP transport efficiency[20]. The default threshold value for epitope identification set at 0.75 was used for the prediction of CTL epitopes.

Helper T-cell (HTL) epitope prediction: HTL epitopes of 15-mer length for human alleles were predicted by the means of the Net MHCII 2.3 Server (http://www.cbs.dtu.dk/services/NetMHCII/). Its server helps to predict binding of peptides to HLA-DR, HLA-DQ, and HLA-DP alleles with the aid of artificial neuron networks. Each epitope is assigned a percentile rank as well as IC50 value. These two characteristics aid to deduce receptor affinity which in-turn forms the basis of this prediction[21,22]. High-affinity peptides should possess IC50 values <50 nM. IC50 value lesser than 500 nM indicates intermediate affinity and values less than 5000 nM indicate low affinity. Thus, it could be inferred that the percentile rank may therefore be inversely related to the affinity of the epitope on one hand while directly related to the IC50 on the other hand.

Construction of a multi-epitope vaccine sequence: The various HTL and CTL epitopes predicted were subjected to analysis of overlapping sequences with linear B-cell epitopes. Effective functioning of each epitope is ensured with insertion of linkers between two epitopes. GPGPG and AAY linkers were used to link all the overlapped epitopes of HTL and CTL respectively[23]. The amino acid sequence of β-subunit of cholera toxin was added as an adjuvant at the N-terminal of the construct with the help of EAAAK linker[23]. It elicits host immune response.

IFN-γ inducing epitope prediction: Interferon gamma (IFN-γ) provides an amplified response to MHC antigens, stimulates macrophages and natural killer cells playing an active role in adaptive and innate immune response. IFN-γ epitopes were predicted from IFNepitope server(http://crdd.osdd.net/raghava/ifnepitope/predict.php). The server essentially constructs overlapped sequences utilized for IFN-γ epitope prediction. The server compiles all MHC Class II binders from IEDB database and categorizes them into inducing and non-inducing binders. The algorithm of this server runs with three models- Motif based model, Support Vector Machine based model and Hybrid approach (both SVM and Motif based) for an input query peptide[24]. The IFN-γ epitopes were predicted with Hybrid approach.

Antigenicity and allergenicity prediction: The antigenicity was predicted by online server ANTIGENpro (http://www.scratch.proteomics.ics.uci.edu/). It is free from any alignment and does not depend on any pathogen identity for antigenecity prediction. A two-step process is implemented in the prediction which is based on five algorithms and multiple representations of the sequence. SVM classifier generates a brief result of prediction informing about the probability of a peptide possessing characteristics of an antigen[25]. Validation of antigenicity has been performed with VaxiJen v. 2.0 server (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html). It assists antigen prediction with alignment-free approach with a predefined threshold of 0.5 and target organism as a virus. It revolves around auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties[26]. The allergenicity of multi-epitope vaccine construct was predicted by using a freely accessible online tool AllerTOP 2.0 (https://www.ddg-pharmfac.net/AllerTOP/). This method is based on auto cross covariance (ACC) transformation of protein sequences into uniform equal-length vectors. The principal properties of the amino acids were represented by five E descriptors: amino acid hydrophobicity, molecular size, helix-forming propensity, relative abundance of amino acids, and β-strand forming propensity. The proteins are categorized using k-nearest neighbor algorithm (kNN, k=1) based on training set which consists of 2427 known allergens and 2427 non-allergens[27]. It was further validated with AllergenFP v.1.0 (http://ddg-pharmfac.net/AllergenFP/). This approach incorporated a four-step algorithm. Firstly, the amino acids in the protein sequences in data sets were described by above mentioned five E-descriptors and the generated strings were transformed into uniform vectors by auto-cross covariance (ACC) transformation. The vectors were transformed into binary fingerprints and compared in terms of Tanimoto coefficient[28].

Prediction of various physicochemical properties: The physicochemical properties of the multi-epitope vaccine which includes molecular weight, extinction coefficient, amino-acid composition, estimated half-life, no. of amino acids, aliphatic index, instability index, theoretical pI and grand average of hydropathicity were explored using ProtParam tool (https://web.expasy.org/protparam/). ProtParam implements the N-end rule to predict half-life, assigns weight value of instability to dipeptides for instability index, mole-percent as well as volumes occupied by aliphatic amino acid side chains for aliphatic index and average of hydropathy values for GRAVY score[29]. The Recombinant protein solubility prediction tool (https://biotech.ou.edu/) was used to predict protein solubility considering the assumption that it is over expressed in *Escherichia coli.* A statistical model has been incorporated in this tool which was made using logistic regression of 32 possible parameters[30].

Secondary structure prediction: The secondary structure predictions of the multi-epitope vaccine construct was performed with PSIPRED server(http://bioinf.cs.ucl.ac.uk/psipred/). The execution of secondary structure prediction with accuracy in PSIPRED is accomplished with incorporation of two feed-forward neural networks which perform synchronized analysis of output obtained from first network followed by output analysis ,of the input as initial network prediction output, generated from PSI-BLAST (Position Specific Iterated - BLAST)[31]. PSIPRED achieves an average Q_3 score of 81.6% with the usage of a very stringent cross validation method to evaluate the method's performance. The RaptorX Property Prediction tool (http://raptorx.uchicago.edu/StructurePropertyPred/predict/)was utilized to analyse solvent accessibility of the vaccine construct. The secondary structure (SS), solvent accessibility (ACC) and disorder regions (DISO) are predicted with a strong method of in-house deep learning model DeepCNF (Deep Convolutional Neural

Fields) incorporated in the server. On one hand, the DeepCNF models complex sequence–structure relationship by a deep hierarchical architecture and on the other hand models interdependency between adjacent property labels. The following characteristics of approx. 84% Q3 accuracy for 3-state SS, around 72% Q8 accuracy for 8-state SS, around 66% Q3 accuracy for 3-state solvent accessibility are attained by the server and approx. 0.89 area under the ROC curve (AUC) for disorder prediction is even accomplished by the server[32-35].

Tertiary Structure prediction: The vaccine construct's tertiary structure prediction was performed with I-TASSER On-lineserver **(**https://zhanglab.ccmb.med.umich.edu/I-TASSER/). I-TASSER (Iterative Threading Assembly Refinement) utilizes hierarchial approach to cast light on protein structure as well as its function. It accepts input protein sequence and proceeds with identification of structural templates from the PDB by multiple threading approach LOMETS, along with iterative template-based fragment assembly simulations generated full-length atomic models. Function predictions of the target are then executed by re-threading the 3D models by means of protein function database BioLiP[36]. The output results display five predicted tertiary structure models of the input sequence based on the first 10 PDB hits as the threading templates. The five models are also provided with C-scores and TM-scores.

Tertiary Structure Refinement: The refinement of the template-based protein tertiary structure is unavoidable for achieving precision in model. This refinement was performed with the freely accessible GalaxyRefine server (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE). This server accepts predicted tertiary structure PDB model as input. The refinement method possesses the ability for improving the global as well as local structural qualities. This server operates firstly via rebuilding all side-chains, beginning from the core and extending to surface layer by layer, by placing the highest-probability rotamers. The next rotamers of highest probabilities are attached on having encounter with steric clashes. The new model with rebuilt sidechains is subjected to refinement with following two relaxation methods – mild and aggressive. The model 1 with lowest energy is generated by mild relaxation and the other four models are generated by aggressive relaxation. The triaxial loop closure method is applied in order to avoid breaks in model structures caused by perturbations to internal torsion angles. The GalaxyRefine displays an output of five generated structure models and their respective GDT-HA, MolProbity, Clash, RMSD, Poor rotamers' scores along with % favoured regions in Ramachandran plot[37].

Tertiary Structure Validation: The tertiary structure validations of refined models were performed with RAMPAGE, ERRAT and ProSA-web servers. The RAMPAGE (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) server accepts PDB input of the model and displays favoured, allowed and outlier region residues in Ramachandran plot along with diagrammatic representation. This server generates the plot with energetically favourable dihedral angles of the amino acids which are calculated on the basis of the Van der Waal radius of the sidechain[38]. The ERRAT (https://servicesn.mbi.ucla.edu/ERRAT/) performs statistical analysis of the non-bonded atomic interactions and verifies their pattern. It involves a quadratic error function for characterization of pair-wise atomic interactions from nine-residue sliding windows accumulated in a database of relied protein structures. The error prone regions are then recognized with pattern of interactions from each window[39]. The ProSA-web server (https://prosa.services.came.sbg.ac.at/prosa.php) involves a computational engine which incorporates knowledge-based potentials of mean force for evaluation of model accuracy. All calculations are carried out with C^{α} potentials which enables its applications to low-resolution structures. The calculation of the energy of the structure is executed with a distance-based pair potential along with a potential capturing the solvent accessibility of protein residues. The output displays model *z*-score and a plot of residue energies. The generated *z*-score indicates overall model quality and measures its deviation from the total energy of the structure with respect to an energy distribution derived from random conformations. *Z*-score of the model outside a range specific for native proteins denote erroneous structures[39-40].

Prediction of discontinuous B-cell epitopes: B-cell epitopes are categorized to linear and discontinuous on the basis of their conformations. Discontinuous epitopes of the vaccine candidate were predicted with freely accessible ElliPro web server **(**http://tools.iedb.org/ellipro/). The server takes PDB file input and tries to find the protein or its homologues in PDB with the protein BLAST. It executes three algorithms for performing the following tasks, firstly protein shape is approximated as an ellipsoid, secondly the residue protrusion index is calculated and thirdly the neighbouring residues are clustered based on their PI values. The output results generate a score for each epitope predicted which is based on Protrusion Index value averaged over epitope residues. The results even display scores, provide access to 3D structure visualization of predicted linear epitopes as well as the discontinuous ones followed by 2D score chart of residues in input.

This method involves detection of protein's 3D shape by means of quantitative approximation of no. of ellipsoids[41].

Molecular Docking of vaccine construct with immunological receptor: The immunoreceptor-antigen interaction is highly significant to elicit the immune response. So, it was inexorable to perform molecular docking of the vaccine construct with immunological receptor molecules, TLR-2, TLR-3 and TLR-4. The docking was executed with ClusPro 2.0 protein-protein docking online server (https://cluspro.bu.edu/login.php). It accepts PDB file input of the two proteins to be docked. The server incorporates the following three computational steps, firstly it performs rigid body docking by means of sampling billions of conformations, secondly root-mean-square deviation (RMSD) based clustering of the 1000 lowest energy structures generated in quest of the largest clusters which are expected to represent the most likely models of the complex and thirdly, it executes refinement of selected structures with the help of energy minimization. The rigid body docking runs with PIPER docking program. This program utilizes Fast Fourier Transform correlation approach. PIPER indicates the interaction energy between two proteins. The second step involves clustering of lowest energy 1000 docked structures by means of interface root mean square deviation (IRMSD) as a form of distance measurement. The third step eliminates steric overlaps with energy minimization. The output models thus obtained represent the models of the 10 most populated clusters and the centre, lowest energy weighted scores of 30 clusters along with cluster members are also displayed[43-45]. The selection of clusters for molecular dynamics simulation was performed based on both the no. of cluster members and weighted score of lowest energy. The 3 docked complexes were visualized with UCSF Chimera software with resources (http://www.rbvi.ucsf.edu/chimera/). The UCSF Chimera implements extensive visualization of the molecular structures[46].

Binding affinity analysis: The binding affinity of the docked complexes were explored with the PROtein BinDIng enerGY prediction PRODIGY web server **(**https://bianca.science.uu.nl/prodigy/). It predicts the binding free energy of the biological complex. It incorporates predictive model solely based on intermolecular contacts along with non-interface surface properties. This server prediction methods have been compiled with Python scripts and Perl. The model predicts with accuracy on a huge heterogenous data showing a Pearson's Correlation of 7.3, with p-value < 0.0001, between the predicted and experimental values and a Root Mean Square Error of 1.89 kcal mol^{-1}.This server takes PDB input of docked complex, chain identifiers of the biomolecules involved in interaction and temperature at which dissociation constant is measured. The output results show the predicted binding free energy of complex (in kcal mol^{-1}), dissociation constant (M), the number as well as type of intermolecular contacts within 5.5 Å distance threshold and the percentages of charged along with polar amino-acids on the non-interacting surface[47,48].

Codon optimization and in silico cloning of the vaccine: Codon optimization is a pre-requisite for effective cloning strategy of the final vaccine construct. The JCat tool (http://www.jcat.de/) was utilized for codon optimization and reverse translation to ensure expression of the construct in an expression vector as per the usage of the expression system. The tool takes protein sequence and the organism for vaccine expression, as input. The output incorporates sequence evaluation and provides Codon Adaptation Index (CAI), GC content of the sequence, adapted sequence. CAI score>0.8 and GC content in the range 30-70% assures efficient translation of the protein in the expression system. This tool is amalgamated to the PRODORIC database which harbours all related data of different organisms[49,50]. The SnapGene software was finally used to design the recombinant plasmid. The following was the methodology of cloning. At the 5′ and 3′ end of the construct, the restriction sites EcoRI and EcoRV were included respectively. This optimized the vaccine construct sequence comprising restriction sites, was cloned in pET- 28a (+) vector using the software to assure expression.

Immune Simulation: The immune simulation of the final multi-epitope vaccine construct is significant for the assessment of its immunogenic profile. The C ImmSim web server (http://kraken.iac.rm.cnr.it/C-IMMSIM/index.php?page=1) was used for such evaluation and simulation of the immune response. This server incorporates prediction methods along with Miyazawa and Jernigan protein-protein potential measurements for the purpose of assessment of molecular binding in reference to immune complexes. A classical immunization experiment (entirely simulated for the model generated) reproduces development of immunological memory. At the end, the aspect of chronic exposure to the same immunogenic molecule leading to the emergence of one or more dominating clones of lymphocytes is even explored and the results manifest high affinity clones undergoing proliferation more than any other. The prediction methods employ algorithms which delineate biological complexes in the form of bit strings and certain Position Specific Scoring Matrix based methods are even included for performance prediction. The server for simulation accepts, antigen sequence as input along with the matrices which define the binding motifs for the haplotype (four

matrices for class I, two HLA-A and two HLA-B, as well as two matrices for class II) and other variables as well as parameters used to modulate the system[51]. The total time step of injection was set to 1400 (1 time step corresponds to 8 hour). The vaccine was injected at an interval of four weeks at following time steps 10, 94, 178, 262, 346, 430, 514, 598, 682, 766, 850, 934. After vaccination, a live replicating virus was injected at time step 1,100 with multiplication factor (0.2) and infectivity (0.6). As a control experiment, a live replicating virus was injected at the same time-step 1,100 without prior vaccination.

References:

1) Who.int. 2020.https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200718-covid-19-sitrep-180.pdf?sfvrsn=39b31718_2 (accessed 19 Jul 2020).

2) Shereen M, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research* 2020; 24: 91-98.

3) Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S 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. *Cellular & Molecular Immunology* 2020; 17: 613-620.

4) Astuti I, Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 2020; 14: 407-412.

5) Schoeman D, Fielding B. Coronavirus envelope protein: current knowledge. *Virology Journal* 2019; 16. doi:10.1186/s12985-019-1182-0.

6) Pettersson M, Crews C. PROteolysis TArgeting Chimeras (PROTACs) — Past, present and future. *Drug Discovery Today: Technologies* 2019; 31: 15-27.

7) Li W, Joshi M, Singhania S, Ramsey K, Murthy A. Peptide Vaccine: Progress and Challenges. *Vaccines* 2014; 2: 515-536.

8) Lester, S. N. & Li, K. Toll-like receptors in antiviral innate immunity. J. Mol. Biol. 426, 1246–1264 (2014).

9) Birra D, Benucci M, Landolfi L, Merchionda A, Loi G, Amato P et al. COVID 19: a clue from innate immunity. *Immunologic Research* 2020; 68: 161-168.

10) Hu, W., Yen, Y.-T., Singh, S., Kao, C.-L. & Wu-Hsieh, B. A. SARS-CoV regulates immune function-related gene expression in human monocytic cells. Viral Immunol. 25, 277–288 (2012).

11) Dosch, S. F., Mahajan, S. D. & Collins, A. R. SARS coronavirus spike protein-induced innate immune response occurs via activation of the NF-κB pathway in human monocyte macrophages in vitro. Virus Res. 142, 19–27 (2009).

12) Zhang L. Multi-epitope vaccines: a promising strategy against tumors and viral infections. *Cellular & Molecular Immunology* 2017; 15: 182-184.

13) Tosta S, Passos M, Kato R, Salgado Á, Xavier J, Jaiswal A et al. Multi-epitope based vaccine against yellow fever virus applying immunoinformatics approaches. *Journal of Biomolecular Structure and Dynamics* 2020; : 1-17.

14) National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2020 Jul 19].

15) Sievers F, Wilm A, Dineen D, Gibson T, Karplus K, Li W et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. 2020.

16) Jespersen M, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research* 2017; 45: W24-W29.

17) Saha S, Raghava G. BcePred: Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties. *Lecture Notes in Computer Science* 2004; : 197-204.

18) Saha S, Raghava G. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function, and Bioinformatics* 2006; 65: 40-48.

19) Yao B, Zhang L, Liang S, Zhang C. SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity. *PLoS ONE* 2012; 7: e45152.

20) Larsen M, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics 2007; 8. doi:10.1186/1471-2105-8-424.

21) Peters, B., Bulik, S., Tampe, R., Van Endert, P. M. & Holzhutter, H. G. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 171, 1741–9 (2003).

22) Jensen K, Andreatta M, Marcatili P, Buus S, Greenbaum J, Yan Z et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 2018; 154: 394-406.

23) Hajighahramani N, Nezafat N, Eslami M, Negahdaripour M, Rahmatabadi S, Ghasemi Y. Immunoinformatics analysis and in silico designing of a novel multi-epitope peptide vaccine against Staphylococcus aureus. *Infection, Genetics and Evolution* 2017; 48: 83-94.

24) Dhanda S, Vir P, Raghava G. Designing of interferon-gamma inducing MHC class-II binders. *Biology Direct* 2013; 8. doi:10.1186/1745-6150-8-30.

25) Magnan C, Zeller M, Kayala M, Vigil A, Randall A, Felgner P et al. High-throughput prediction of protein antigenicity using protein microarray data. *Bioinformatics* 2010; 26: 2936-2943.

26) Doytchinova I, Flower D. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007; 8. doi:10.1186/1471-2105-8-4.

27) Dimitrov I, Bangov I, Flower D, Doytchinova I. AllerTOP v.2—a server for in silico prediction of allergens. *Journal of Molecular Modeling* 2014; 20. doi:10.1007/s00894-014-2278-5.

28) Dimitrov I, Naneva L, Doytchinova I, Bangov I. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* 2013; 30: 846-851.

29) Gasteiger, E. et al. Protein Identifcation and Analysis Tools on the ExPASy Server. In the Proteomics Protocols Handbook (ed. Walker, J. M.) 571–607 (Humana Press, Totowa, NJ, 2005).

30) Diaz A, Tomba E, Lennarson R, Richard R, Bagajewicz M, Harrison R. Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnology and Bioengineering* 2010; 105: 374-383.

31) Jones D. Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. *Journal of Molecular Biology* 1999; 292: 195-202.

32) Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientifc Reports 6, 18962 (2016).

33) Yang, Y. et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform (2016)

34) Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Research 2016; 44: W430-W435.

35) Wang S, Ma J, Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics 2016; 32: i672-i679.

36) Yang J, Zhang Y. Protein Structure and Function Prediction Using I-TASSER. *Current Protocols in Bioinformatics* 2015; 52. doi:10.1002/0471250953.bi0508s52.

37) Heo L, Park H, Seok C. GalaxyRefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Research* 2013; 41: W384-W388.

38) Lovell S, Davis I, Arendall W, de Bakker P, Word J, Prisant M et al. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* 2003; 50: 437-450.

39) Colovos C, Yeates T. Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science* 1993; 2: 1511-1519.

40) Wiederstein M, Sippl M. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 2007; 35: W407-W410.

41) Sippl M. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Genetics* 1993; 17: 355-362.

42) Ponomarenko J, Bui H, Li W, Fusseder N, Bourne P, Sette A et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* 2008; 9. doi:10.1186/1471-2105-9-514.

43) Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella S, Xia B et al. New additions to the ClusPro server motivated by CAPRI. *Proteins: Structure, Function, and Bioinformatics* 2017; 85: 435-444.

44) Kozakov D, Hall D, Xia B, Porter K, Padhorny D, Yueh C et al. The ClusPro web server for protein–protein docking. *Nature Protocols* 2017; 12: 255-278.

45) Kozakov D, Beglov D, Bohnuud T, Mottarella S, Xia B, Hall D et al. How good is automated protein docking? *Proteins: Structure, Function, and Bioinformatics* 2013; 81: 2159-2166.

46) Pettersen E, Goddard T, Huang C, Couch G, Greenblatt D, Meng E et al. UCSF Chimera?A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 2004; 25: 1605-1612.

47) Vangone A, Bonvin A. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife* 2015; 4. doi:10.7554/elife.07454.

48) Xue L, Rodrigues J, Kastritis P, Bonvin A, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. Bioinformatics 2016; : btw514.

49) Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel D et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research* 2005; 33: W526-W531.

50) Morla, S., Makhija, A. & Kumar, S. Synonymous codon usage pattern in glycoprotein gene of rabies virus. Gene 584, 1–6 (2016).

51) 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* 2010; 5: e9862.

Acknowledgement

Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311

Author contribution

SB and KM designed the protocols and the experiments. Multiple sequence alignments and binding analysis were done by BM. Tertiary structure prediction and docking experiments were done by KM. SB performed the physicochemical properties prediction, secondary structure prediction, tertiary structure refinement, codon optimization and cloning experiments. Immune simulation was done by BM, SB and KM. Potential epitope prediction, antigenicity prediction, allergenicity prediction and manuscript reviewing were done by GJG, DG and KM. Original manuscript has been written by KM. SB and BM have supervised the experiments. Final manuscript reviewing was done by all the authors.

Competing interests

The authors declare no competing interests.

Supplementary files

Protein	CTL Epitopes	Combined Score	HTL Epitopes	Affinity (nM)
Nucleocapsid Phosphoprotein	GTDYKHWPQ	1.21	ASWFTALTQHGKEDL	30.1
	LLNKHIDAY	1.39	QIGYYRRATRRIRGG	7.9
	LSPRWYFYY	2.34	NNAAIVLQLPQGTTL	10.6
			HWPQIAQFAPSASAF	17.2
			QQTVTLLPAADLDDF	7
Membrane Glycoprotein	LLEQWNLVI	0.77	SYYKLGASQRVAGDS	3.3
	LVGLMWLSY	1.40		
	YSRYRIGNY	1.66		
Envelope Protein	VSLVKPSFY	1.71		
ORF6	NLDYIINLI	0.79	ILLIIMRTFKVSIWN	5.8
	LTENKYSQL	0.95		
ORF7a	ITLATCELY	1.42	VKHVYQLRARSVSPK	5.6
	RQEEVQELY	1.60		
ORF10	QVDVVNFNL	0.85		

Table 1. Prediction scores of selected CTL epitopes and affinities of selected HTL epitopes

A

```
CLUSTAL O(1.2.4) multiple sequence alignment

lcl|MT050493.1_prot_QIA98586.1_5          MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPV 60
lcl|MT066156.1_prot_QIA98557.1_5          MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPV 60
lcl|MN985325.1_prot_QHO60597.1_5          MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPV 60
lcl|NC_045512.2_prot_YP_009724393.1_6     MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPV 60
                                          ************************************************************

lcl|MT050493.1_prot_QIA98586.1_5          TLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILL 120
lcl|MT066156.1_prot_QIA98557.1_5          TLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILL 120
lcl|MN985325.1_prot_QHO60597.1_5          TLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILL 120
lcl|NC_045512.2_prot_YP_009724393.1_6     TLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILL 120
                                          ************************************************************

lcl|MT050493.1_prot_QIA98586.1_5          NVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYK 180
lcl|MT066156.1_prot_QIA98557.1_5          NVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYK 180
lcl|MN985325.1_prot_QHO60597.1_5          NVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYK 180
lcl|NC_045512.2_prot_YP_009724393.1_6     NVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYK 180
                                          ************************************************************

lcl|MT050493.1_prot_QIA98586.1_5          LGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ   222
lcl|MT066156.1_prot_QIA98557.1_5          LGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ   222
lcl|MN985325.1_prot_QHO60597.1_5          LGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ   222
lcl|NC_045512.2_prot_YP_009724393.1_6     LGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ   222
                                          ****************************************
```

B

```
CLUSTAL O(1.2.4) multiple sequence alignment

QIA98588.1          MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNSPFHPLADNKFALTCFS        60
QIA98559.1          MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNSPFHPLADNKFALTCFS        60
QHO60599.1          MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNSPFHPLADNKFALTCFS        60
YP_009724395.1      MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNSPFHPLADNKFALTCFS        60
                    ************************************************************

QIA98588.1          TQFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKT        120
QIA98559.1          TQFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKT        120
QHO60599.1          TQFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKT        120
YP_009724395.1      TQFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKT        120
                    ************************************************************

QIA98588.1          E   121
QIA98559.1          E   121
QHO60599.1          E   121
YP_009724395.1      E   121
                    *
```

C

```
CLUSTAL O(1.2.4) multiple sequence alignment

lcl|MT050493.1_prot_QIA98585.1_4          MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
lcl|MT066156.1_prot_QIA98556.1_4          MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
lcl|MN985325.1_prot_QHO60596.1_4          MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
lcl|NC_045512.2_prot_YP_009724392.1_5     MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
                                          ************************************************************

lcl|MT050493.1_prot_QIA98585.1_4          RVKNLNSSRVPDLLV      75
lcl|MT066156.1_prot_QIA98556.1_4          RVKNLNSSRVPDLLV      75
lcl|MN985325.1_prot_QHO60596.1_4          RVKNLNSSRVPDLLV      75
lcl|NC_045512.2_prot_YP_009724392.1_5     RVKNLNSSRVPDLLV      75
                                          **************
```

D

```
CLUSTAL O(1.2.4) multiple sequence alignment

QIA98587.1          MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEI        60
QIA98558.1          MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEI        60
QHO60598.1          MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEI        60
YP_009724394.1      MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEI        60
                    ************************************************************

QIA98587.1          D   61
QIA98558.1          D   61
QHO60598.1          D   61
YP_009724394.1      D   61
                    *
```

E

```
CLUSTAL O(1.2.4) multiple sequence alignment

lcl|MT050493.1_prot_QIA98585.1_4          MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
lcl|MT066156.1_prot_QIA98556.1_4          MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
lcl|MN985325.1_prot_QHO60596.1_4          MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
lcl|NC_045512.2_prot_YP_009724392.1_5     MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYS 60
                                          ************************************************************

lcl|MT050493.1_prot_QIA98585.1_4          RVKNLNSSRVPDLLV      75
lcl|MT066156.1_prot_QIA98556.1_4          RVKNLNSSRVPDLLV      75
lcl|MN985325.1_prot_QHO60596.1_4          RVKNLNSSRVPDLLV      75
lcl|NC_045512.2_prot_YP_009724392.1_5     RVKNLNSSRVPDLLV      75
                                          **************
```

F

```
CLUSTAL O(1.2.4) multiple sequence alignment

QIA98591.1          MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT        38
QIA98562.1          MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT        38
QHO60602.1          MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT        38
YP_009725255.1      MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT        38
                    **************************************
```

Figure 1. Multiple sequence alignment results of **(A)** Nucleocapsid phosphoprotein **(B)** ORF7a **(C)** Envelope protein **(D)** ORF6 **(E)** Membrane glycoprotein **(F)** ORF 10

Figure 2. (A) Results of secondary structure predicted by PSIPRED. **(B)** Chemical properties of each amino acid residue of the vaccine construct predicted by PSIPRED.