bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identification of potential vaccine candidates against *SARS-CoV-2*,

A step forward to fight COVID-19: A Reverse Vaccinology Approach

Ekta Gupta[1], Rupesh Kumar Mishra[2], Ravi Ranjan Kumar Niraj[2#]

1. Dr. B. Lal Institute of Biotechnology, Jaipur (INDIA) - 302017

2. Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur (INDIA)-303002

 # Correspondence Author:

Dr. Ravi Ranjan Kumar Niraj

Assistant Professor

Amity Institute of Biotechnology,

Amity University Rajasthan, Jaipur (INDIA)-303002

Email: rrkniraj@gmail.com

ORCID: 0000-0002-8547-6037

Mobile: +91-9729559580

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract: The recent Coronavirus Disease 2019 (COVID-19) causes an immense health crisis to global public health. The COVID-19 is the etiologic agent of a recently arose disease caused by the severe acute respiratory syndrome coronavirus 2 (*SARS-CoV-2*). Presently, there is no vaccine available against this emerged viral disease. Therefore, it is indeed a need of the hour to develop an effectual and safe vaccine against this decidedly pandemic disease. In the current study, we collected *SARS-CoV-2* genome which is prominent in India against human host, further more using reverse vaccinology here we claim effective vaccine candidates that can be mile stone in battle against COVID19. This novel study divulged one promising antigenic peptide GVYFASTEK from surface glycoprotein (protein accession no. - QIA98583.1) of *SARS-CoV-2*, which was predicated to be interacted with MHC alleles and showed up to 90% conservancy and high value of antigenicity. Subsequently, the molecular docking and simulation studies were verified molecular interaction of this prime antigenic peptide with the residues of HLA-A*11-01 allele for MHC Class I. After vigorous analysis, this peptide was predicted to be suitable epitope which is capable to induce the strong cell-mediated immune response against the *SARS-CoV-2*. Consequences from the current study could facilitate selecting *SARS-CoV-2* epitopes for vaccine production pipelines in the immediate future. This novel research will certainly pave the way for a fast, reliable and virtuous platform to provide timely countermeasure of this dangerous pandemic disease, COVID-19.

Keywords: COVID-19, *SARS-CoV-2*, Reverse vaccinology, Molecular docking, Molecular dynamics simulation, Vaccine candidates.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

The Coronavirus Disease 2019 (COVID-19) begin in December 2019, like a viral outbreak in Wuhan city of China (Holshue et al., 2020). It gained rapid foothold across the world resulting in WHO declared it as a pandemic (https://www.cnbc.com/2020/03/11/who-declares-the-coronavirus-outbreak-a-global-pandemic.html). As on October, 9 2020 worldwide total 36,799,479 cases and 1,067,537 deaths were reported by WHO(https://www.worldometers.info/coronavirus/).Coronaviruses have long been recognized as crucial veterinary pathogens, causing enteric and respiratory diseases in mammals as well as in birds(Adeoye, Oso, Olaoye, Tijjani, & Adebayo, 2020). The SARS-CoV-2 virus spreads primarily through saliva, droplets, or discharges from the nose of an infected person after coughing or sneezing (Pant, Singh, Ravichandiran, Murty, & Srivastava, 2020). The coronaviruses are enveloped RNA viruses having the largest genome among all RNA (Sinha et al., 2020). As no drug or vaccine is available against SARS-CoV-2 many countries, including India has announced lockdown to keep away from further spread of the virus(Umesh, Kundu, Selvaraj, Singh, & Dubey, 2020).

In this period of time when the continuous transmission of the virus across borders and health burden on the global scale is rapidly increasing, more urgent studies are required and in the absence of effective cures majorly drugs, vaccination or immunization therapy is imperative in order to target whole population. Likewise to move forward vaccine development pipeline, immunoinfomatics tools have been proved crucial(S. Mishra & Sinha, 2009). There is less knowledge about the pathogenesis of the virus; therefore, an immunoinformatics-based approach to investigate the immunogenic epitopes is required(Enayatkhani et al., 2020).

Since the Covid19 has affected almost all of the world's population, promiscuous epitopes binding to a variety of HLA alleles for larger dissemination is vital. For that, *in silico* approaches will be remarkably useful in helping to develop a cure in as fast manner as possible(D. Mishra, 2020). The antibody generation by activation of B-cell as well as acute viral clearance by T-cells along with virus-specific memory generation by CD8+ T-cells are analogously important to develop immunity against the virus(Enjuanes et al., 2016). The S protein is considered to be highly antigenic and thereby can evoke strong immune responses and generate neutralizing antibodies that can block the attachment of virus to the host cells(Du et al., 2009).

In reverse vaccinology, various tools of *in silico* biology are used to discover the novel antigens by studying the genetic makeup of a pathogen and the genes that could lead to good epitopes are determined. The reverse vaccinology approach provide fast and cost effective vaccine discovery plateform(Ullah, Sarkar, & Islam, 2020). In

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reverse vaccinology approach a novel antigen is identified using omics analysis of target organism. In-silico as well as reverse vaccinology approach facilitates an easier, time and labor saving process of antigen discovery(Gupta, Gupta, & Niraj, 2019).

Herein, we explored the proteome of SARS-CoV-2, prominent in Indian geographical origin against human host to identify potential antigenic proteins and epitopes that can effectively elicit cellular mediated immune response against COVID-19. This significant research disclosed promising antigenic peptides from surface glycoprotein of *SARS-CoV-2*. The outcomes from this very significant analysis could help selecting *SARS-CoV-2* epitopes for vaccine production pipelines soon.

2. Methodology

2.1 Selection of strain

The highly virulent strain *SARS-CoV-2* was selected for *in-silico* study. The complete genome of *SARS-CoV-2* is available on the National Center for Biotechnology Information or NCBI(https://www.ncbi.nlm.nih.gov/) with *RefSeq NC_045512.2*

2.2 Protein Identification and Retrieval

Twelve viral protein sequences of *SARS-CoV2* against (Host: Human, Country: India) were retrieved from ViPR database(Pickett et al., 2012). These proteins were: Orf10 protein (**QIA98591.1)**, Orf8 protein **(QIA98589.1)**, Orf7a protein **(QIA98588.1)**, Orf6 protein **(QIA98587.1)**, Orf3a protein **(QIA98584.1)**, Membrane glycoprotein **(QIA98586.1)**, Envelope protein **(QIA98585.1)**, Surface glycoprotein **(QIA98583.1)**, Surface glycoprotein **(QHS34546.1)**, Nucleocapsid protein **(QII87776.1)**, Nucleocapsid protein **(QII87775.1)** and Nucleocapsidphosphoprotein **(QIA98590.1)**.

2.3 Physicochemical Property Prediction

The ExPASy's online tool ProtParam (Walker, 2005) was used to predict various physicochemical properties of the selected protein sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.4 Protein Antigenicity

VaxiJen v2.0 (Doytchinova & Flower, 2007) was used to predict antigenicity of proteins. This software uses FASTA file format of amino acid sequences as an input and on the basis of physicochemical properties of proteins, antigenicity is predicted by the tool. The output is denoted according to an antigenic score (Meunier et al., 2016). During analysis, the threshold was kept at 0.4.

2.5 B-cell and T-cell Epitope Prediction

The B-cell and T-cell epitopes of the selected surface glycoprotein sequence was predicted via IEDB (The Immune Epitope Database). The IEDB database holds large amount of experimental data on epitopes and antibodies. It allows robust analysis on many epitopes in the context of some tools like: conservation across antigens, population coverage, and clusters with similar sequences (Vita et al., 2019). In order to obtain MHC class-I restricted CD8+ cytotoxic T-lymphocyte (CTL) epitopes of the selected surface glycoprotein sequence, NetMHCpan EL 4.0 prediction method was applied for HLA-A*11-01 allele. For MHC class-II restricted CD4+ helper T-lymphocyte (HTL) epitopes were obtained for HLA DRB1*04-01 allele using Sturniolo prediction method. Top ten MHC class-I and MHC class-II epitopes were randomly selected on the basis of their percentile scores and antigenicity scores (AS). Five random B-cell lymphocyte epitopes (BCL) were selected based on of their higher length using Bipipered linear epitope prediction method(Ullah et al., 2020).

2.6 Antigenicity and Allergenicity of the predicted epitopes

VaxiJen v2.0 was utilized to predict protein antigenicity. During antigenicity analysis, threshold was kept at 0.4. The allergenicity of the selected epitopes was predicted via AllerTOP v2(https://www.ddg-pharmfac.net/AllerTOP/).

2.7 Transmembrane Helix and Toxicity Prediction of the predicated epitopes

The transmembrane helix of the selected epitopes was predicted using the TMHMM v2.0 server (http://www.cbs.dtu.dk/services/TMHMM/). The server predicts whether the epitope would be transmembrane, remain inside or outside of the membrane. The toxicity prediction of the selected epitopes was carried out via ToxinPred server https://webs.iiitd.edu.in/raghava/toxinpred/protein.php.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.8 Prediction of Conservancy for the Selected Epitopes

The conservancy analysis of the epitopes was performed via the epitope conservancy analysis tool of IEDB server (Vita et al., 2019). During analysis, the sequence identity threshold was kept at '>=50'.

2.9 The MHC Alleles; Cluster analysis

Cluster analysis was carried out by MHCcluster 2.0 (Thomsen, Lundegaard, Buus, Lund, & Nielsen, 2013). During cluster analysis, the number of peptides to be included was kept at 50,000, the number of bootstrap calculations were kept at 100. For cluster analysis, the NetMHCpan-2.8 prediction method was used.

2.10. Generation of the 3D Structures of the Selected Epitopes

The PEP-FOLD3 online tool was used to predict 3D structures of the selected best epitopes (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3/). This server is a tool for generating de novo peptide 3-D structure (Lamiable et al., 2016; Shen, Maupetit, Derreumaux, & Tufféry, 2014; Thévenet et al., 2012).

2.11. Molecular Docking and Molecular Dynamics Simulation

Pre-docking was carried out by UCSF Chimera (Pettersen et al., 2004). The peptide-protein docking of the selected epitopes was carried out by online docking tool PatchDock(https://bioinfo3d.cs.tau.ac.il/PatchDock/php.php) results of PatchDock were refined and re-scored by FireDock server (http://bioinfo3d.cs.tau.ac.il/FireDock/php.php). Later on, docking was performed by HPEPDOCK server(Zhou, Jin, Li, & Huang, 2018). Docking pose analysis was done by using Ligplot(Wallace, Laskowski, & Thornton, 1995). The molecular simulation was executed with GROMACS 2018.1 package with the force field as Gromos43a1. The protein salvation was performed with SPC water model in a cubic box (10.8 × 10.8 × 10.8 nm^3). The solvated system of protein was processed for energy minimization using the steepest algorithm up to a maximum 25,000 steps or until the maximum force (Fmax) is not greater than 1000 kJ/ mol nm which is the default threshold. The NVT and NPT ensembles for 50,000 steps (100 ps) at 300 K and 1 atm. Here, the system was firstly equilibrated using NVT ensemble followed by NPT ensemble. Then after the final Molecular Dynamic Simulation was performed for dock complex of [GVYFASTEK epitope docked against the HLA-A*11-01 allele (PDB ID: 5WJL)]. Finally, the simulations were evaluated RMSD and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RMSF were calculated for complete episode of simulations. The all steps are kept similar except the final the molecular dynamics simulation was carried out for 50 ns long.

3. Results

3.1 Selection and Retrieval of Viral Protein Sequences

The *SARS-CoV-2*, strain was identified. Twelve viral protein sequences of *SARS-CoV2* against (Host: Human, Country: India) were retrieved from ViPR database and selected for the possible vaccine candidate identification **(Table-1)**. These proteins were: Orf10 protein (**QIA98591.1)**, Orf8 protein **(QIA98589.1)**, Orf7a protein **(QIA98588.1)**, Orf6 protein **(QIA98587.1)**, Orf3a protein **(QIA98584.1)**, Membrane glycoprotein **(QIA98586.1)**, Envelope protein **(QIA98585.1)**, Surface glycoprotein **(QIA98583.1)**, Surface glycoprotein **(QHS34546.1)**, Nucleocapsid protein **(QII87776.1)**, Nucleocapsid protein **(QII87775.1)** and Nucleocapsidphosphoprotein **(QIA98590.1)**. The FASTA sequence of proteins mentioned in **(Supplementary File: S1)**

Table 1: SARS-CoV-2 (Host: Human, Country: India) viral protein sequence identification and retrieval via ViPR database (Viral pathogen database and analysis resource).

S.No.	*Gene Symbol*	*Protein Name*	*GenBank Accession*	*GenBank Protein Accession*
1	orf10	Orf10 protein	MT050493	QIA98591.1
2	orf8	Orf8 protein	MT050493	QIA98589.1
3	orf7a	Orf7a protein	MT050493	QIA98588.1
4	orf6	Orf6 protein	MT050493	QIA98587.1
5	orf3a	Orf3a protein	MT050493	QIA98584.1
6	M	Membrane glycoprotein	MT050493	QIA98586.1
7	E	Envelope protein	MT050493	QIA98585.1
8	S	Surface glycoprotein	MT050493	QIA98583.1
9	S	Surface glycoprotein	MT012098	QHS34546.1
10	N	Nucleocapsid protein	MT163715	QII87776.1
11	N	Nucleocapsid protein	MT163714	QII87775.1
12	N	Nucleocapsid phosphoprotein	MT050493	QIA98590.1

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2 Physicochemical Property Analysis and Protein Antigenicity

The physicochemical property analysis like number of amino acids, the molecular weights, theoretical pI, extinction coefficients (in M-1 cm-1), Est. half-life (in mammalian cell), instability indexes, aliphatic indexes and grand average of hydropathicity (GRAVY) of the twelve proteins were predicted **(Table-2)**. For antigenicity prediction threshold value kept at 0.4, all proteins were found to be antigenic **(Table-3)**. The physicochemical study revealed that the surface glycoprotein **(QIA98583.1)** had the comparatively maximum extinction co-efficient of 148960M-1 cm-1 and lowest GRAVY value of -0.077. In addition, surface glycoprotein was stable and antigenic. We selected this surface glycoprotein for further analysis.

Table 2: Physiochemical property analysis of SARS-CoV-2 against (Host: Human, Country: India) viral proteins.

	Properties							
Gene Symbol	**No. of amino acids**	**Molecular weight**	**Theoretical pI**	**Ext. coefficient (in M-1 cm-1)**	**Est. half-life (in mammalian cell)**	**Instability index**	**Aliphatic index**	**GRAVY (gra average of hydropathicit**
orf10	38	4449.23	7.93	4470	30 hours	16.06 (stable)	107.63	0.637
orf8	121	13804.93	5.42	16305	30 hours	46.24 (unstable)	94.13	0.181
orf7a	121	13744.17	8.23	7825	30 hours	48.66 (unstable)	100.74	0.318
orf6	61	7272.54	4.60	8480	30 hours	31.16 (stable)	130.98	0.233
orf3a	275	31122.94	5.55	58705	30 hours	32.96 (stable)	103.42	0.275
M	222	25146.62	9.51	52160	30 hours	39.14 (stable)	120.86	0.446
E	75	8365.04	8.57	6085	30 hours	38.68 (stable)	144.00	1.128
S	1273	141206.52	6.24	148960	30 hours	33.01 (stable)	84.82	-0.077
S	1272	140972.27	6.16	147470	30 hours	32.78 (stable)	85.05	-0.071
N	88	9827.08	10.23	8480	4.4 hours	36.54 (stable)	61.14	-1.067
N	133	14363.88	11.37	8480	1 hours	58.97 (unstable)	44.21	-1.170
N	419	45625.70	10.07	43890	30 hours	55.09 (unstable)	52.53	-0.971

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: Antigenicity predication of SARS-CoV-2 viral proteins (Threshold value: 0.4)

S. No.	*Protein Name*	*Antigenicity Score*	*Antigenicity; Threshold=0.4*
1	Orf10 protein	0.7185	Antigenic
2	Orf8 protein	0.6063	Antigenic
3	Orf7a protein	0.6441	Antigenic
4	Orf6 protein	0.6131	Antigenic
5	Orf3a protein	0.4945	Antigenic
6	Membrane glycoprotein	0.5102	Antigenic
7	Envelope protein	0.6025	Antigenic
8	Surface glycoprotein	0.4654	Antigenic
9	Surface glycoprotein	0.4687	Antigenic
10	Nucleocapsid protein	0.5767	Antigenic
11	Nucleocapsid protein	0.6235	Antigenic
12	Nucleocapsid phosphoprotein	0.5059	Antigenic

3.3 T-cell and B-cell Epitope Prediction

The T-cell epitopes of MHC class-I was determined by NetMHCpan EL 4.0 prediction method of the IEDB server keeping the sequence length 9. The server generated epitopes further analyzed on the basis of the antigenicity scores (AS) and percentile scores, top ten potential epitopes were selected randomly for antigenicity, allergenicity, toxicity and conservancy tests. The server ranks the predicted epitopes based on the ascending order of percentile scores **(Table-4a)**. The T-cell epitopes of MHC class-II (HLA DRB1*04-01 allele) of the protein was also determined by IEDB server **(Table-4b)**. , where the Sturniolo prediction methods was used. For protein, ten out of the top ranked epitopes were selected randomly for further analysis. Additionally, the B-cell epitopes of the protein was selected using Bipipered linear epitope prediction method of the IEDB server and the selection of epitopes were based on their higher lengths **(Fig-1).**

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4a: MHC Class-I epitopes of Surface glycoprotein (QIA98583.1): Table represents topology, antigenicity, allergenicity, toxicity and conservancy analysis of protein.

Surface glycoprotein (QIA98583.1)									
Epitope	**Start**	**End**	**Topology**	**Antigenicity**	**Score**	**Allergenicity**	**Toxicity**	**Minimum Identity**	**Conservanc**
GVYFASTEK	19	27	Inside	Antigen	0.7112	Non-allergen	Non-toxic	11.11%	100%
VTYVPAQEK	15	23	Inside	Antigen	0.8132	Allergen	Non-toxic	22.22%	100%
ASANLAATK	40	48	Inside	Antigen	0.7041	Allergen	Non-toxic	22.22%	100%
TLADAGFIK	57	65	Inside	Antigen	0.5781	Non-allergen	Non-toxic	22.22%	100%
TLKSFTVEK	22	30	Inside	Non-antigen	0.0809	Allergen	Non-toxic	11.11%	100%
NSASFSTFK	20	28	Inside	Non-antigen	0.1232	Allergen	Non-toxic	11.11%	100%
TEILPVSMTK	24	33	Inside	Antigen	1.4160	Allergen	Non-toxic	10.00%	100%
SSTASALGK	29	37	Outside	Antigen	0.6215	Allergen	Non-toxic	22.22%	100%
GTHWFVTQR	49	57	Inside	Non-antigen	0.0723	Allergen	Non-toxic	11.11%	100%
EILPVSMTK	25	33	Inside	Antigen	1.6842	Allergen	Non-toxic	11.11%	100%

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4b: MHC Class-II epitopes of Surface glycoprotein (QIA98583.1): Table represents topology, antigenicity, allergenicity, toxicity and conservancy analysis of protein.

Surface glycoprotein (QIA98583.1)									
Epitope	**Start**	**End**	**Topology**	**Antigenicity**	**Score**	**Allergenicity**	**Toxicity**	**Minimum Identity**	**Conservanc**
SNFRVQPTESI	36	46	Inside	Antigen	0.9897	Allergen	Non-toxic	11.11%	100%
NFRVQPTESIV	37	47	Inside	Antigen	1.0669	Non-Allergen	Non-toxic	22.22%	100%
FRVQPTESIVR	38	48	Inside	Non-antigen	0.3493	Allergen	Non-toxic	9.09%	100%
VYYHKNNKSWM	3	13	Inside	Non-antigen	0.3726	Allergen	Non-toxic	18.18%	100%
LGVYYHKNNKS	1	11	Inside	Antigen	0.8696	Allergen	Non-toxic	9.09%	100%
GVYYHKNNKSW	2	12	Inside	Antigen	0.6685	Allergen	Non-toxic	9.09%	100%
LLIVNNATNVV	47	57	Inside	Antigen	0.4166	Non-Allergen	Non-toxic	9.09%	100%
LIVNNATNVVI	48	58	Inside	Non-antigen	0.2045	Non-Allergen	Non-toxic	9.09%	100%
IVNNATNVVIK	49	59	Inside	Non-antigen	0.2274	Allergen	Non-toxic	9.09%	100%
VFVSNGTHWFV	44	54	Outside	Non-antigen	0.0957	Allergen	Non-toxic	18.18%	100%

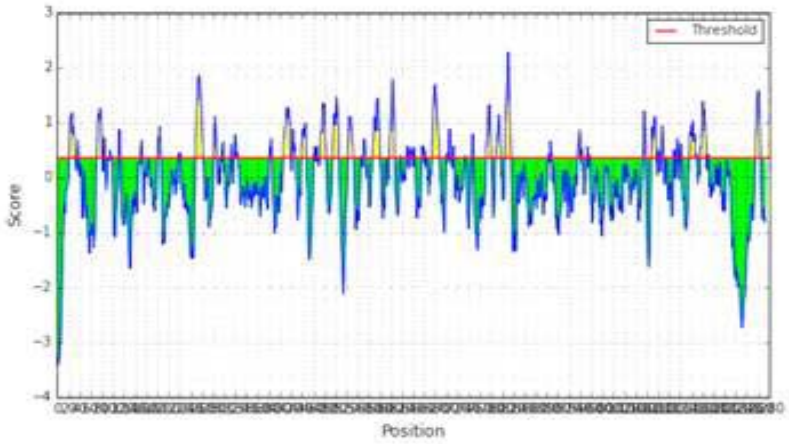

Fig 1: B cell Epitopes prediction: The graph of epitope prediction for surface glycoproteins QIA98583.1

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.4 Topology Identification of the Epitopes

The topology of the selected epitopes was determined by TMHMM v2.0 server (http://www.cbs.dtu.dk/services/TMHMM/). The **Table 4a** and **Table 4b** represent the potential T-cell epitopes of selected surface glycoprotein**. Table 5** shows the potential B-cell epitopes with their respective topologies.

Table 5: B cell epitopes of Surface glycoprotein (QIA98583.1): Table represents topology, antigenicity and allergenicity analysis of protein.

Surface glycoprotein (QIA98583.1)

Epitope	**Topology**	**Antigenicity**	**Allergenicity**	**Epitope**	**Topology**	**Antigenicity**	**Allergenicity**
RTQLPPAYTNS	Inside	Antigen	Allergen	QIAPGQTGKIAD	Inside	Antigen	Non-Allergen
SGTNGTKRFDN	Inside	Antigen	Allergen	YGFQPTNGVGYQ	Outside	Antigen	Allergen
LTPGDSSSGWTAG	Outside	Antigen	Non-Allergen	RDIADTTDAVRDPQ	Inside	Antigen	Allergen
VRQIAPGQTGKIAD	Inside	Antigen	Non-Allergen	QTQTNSPRRARSV	Inside	Non-antigen	Non-Allergen
YQAGSTPCNGV	Inside	Non-antigen	Non-Allergen	ILPDPSKPSKRS	Outside	Antigen	Non-Allergen

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.4. Antigenicity, allergenicity, toxicity and conservancy analysis of the epitopes

The selected T-cell epitopes found to be highly antigenic as well as non-allergenic, non-toxic, and had conservancy of greater than 90%. Among the ten selected MHC class-I epitopes and ten selected MHC class-II epitopes, total four epitopes were selected based on the mentioned criteria: GVYFASTEK, TLADAGFIK, NFRVQPTESI and LLIVNNATNV.

3.5. Cluster Analysis of the MHC Alleles

The cluster analysis of the MHC class-I alleles that will possibly interact with the predicted epitopes were carried out by online tool MHCcluster 2.0 (http://www.cbs.dtu.dk/services/MHCcluster/). The tool generates the clusters of the alleles in phylogenetic manner. The Results illustrate the outcome of the experiment where the red zone indicates strong interaction and the yellow zone corresponds to weaker interaction **(Fig-2)**.

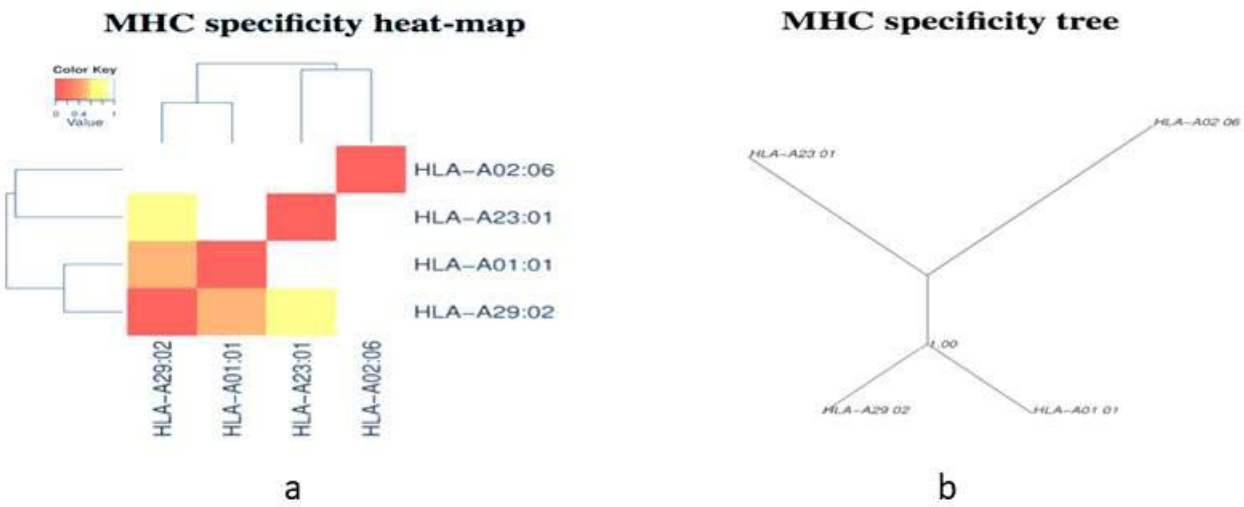

Fig 2: MHC Class Cluster Analysis (a) Heat map (b) Specificity tree

Note: Here, red zone indicates strong interaction and the yellow zone corresponds to weaker interaction.

3.6. The 3D Structure prediction of the Epitopes

All the T-cell epitopes were subjected to 3D structure is predicted by the PEP-FOLD3 server. The predicted 3D structures were used for peptide-protein docking **(Fig-3)**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

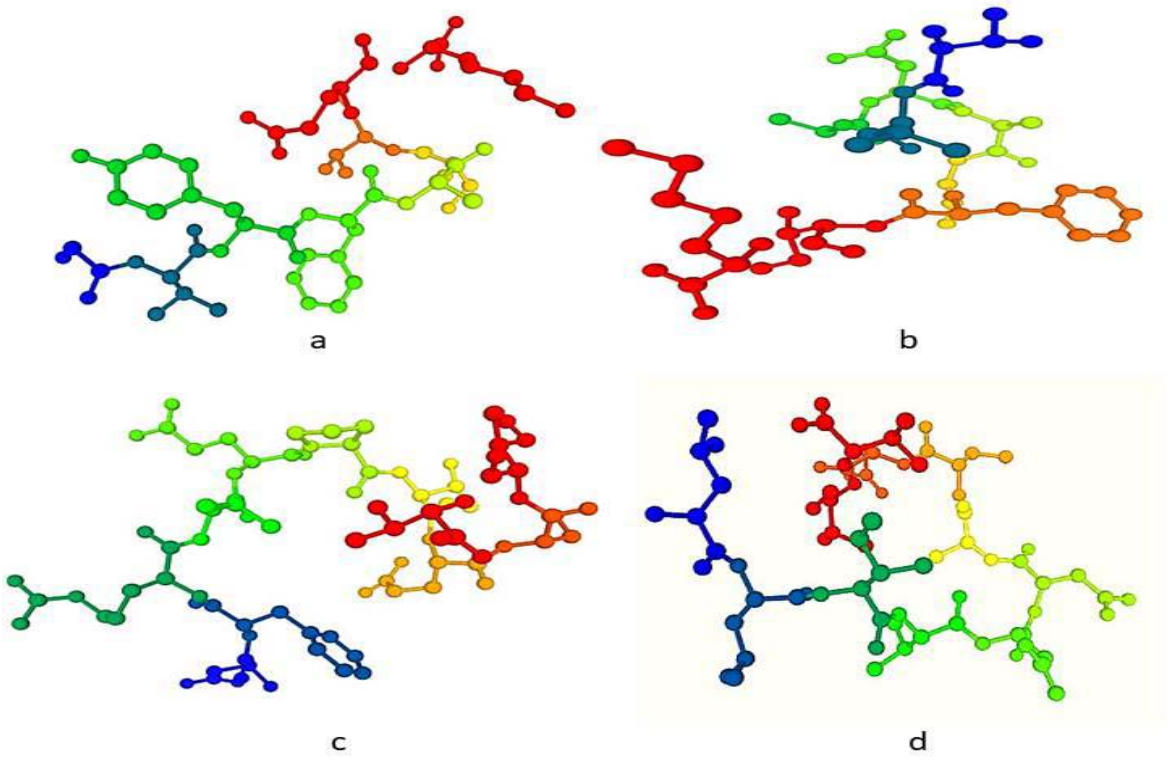

Fig 3: 3D structure generation of T-cell epitopes by the PEP-FOLD3 server. Epitope representation: (a) GVYFASTEK (b) TLADAGFIK (c) NFRVQPTESI and (d) LLIVNNATNV.

3.7 Peptide-Protein Docking &Vaccine candidate's prioritization

The Molecular Docking study was performed to observe, whether all the epitopes had the capacity to bind with the MHC class-I and MHC class-II molecule. The selected epitopes were docked with the HLA-A*11-01 allele (PDB ID: 5WJL) and HLA DRB1*04-01 (PDB ID: 5JLZ). The docking was performed using PatchDock online docking tool and then the results were refined by FireDock online server. Results were also analyzed by HPEPDOCK server **(Supplementary Figure: S2)**. Among the Four epitopes of selected glycoprotein QIA98583.1, GVYFASTEK (MHC class I epitope) showed the best result with the lowest global energy of -52.82. Further, docking pose was analyzed via Ligplot (**Fig 4a)** and docking site can be clearly visualized in **(Fig 4b).** We also identified highly antigenic and non-allergenic B-cell vaccine candidates LTPGDSSSGWTAG and VRQIAPGQTGKIAD from the selected Surface glycoprotein (QIA98583.1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

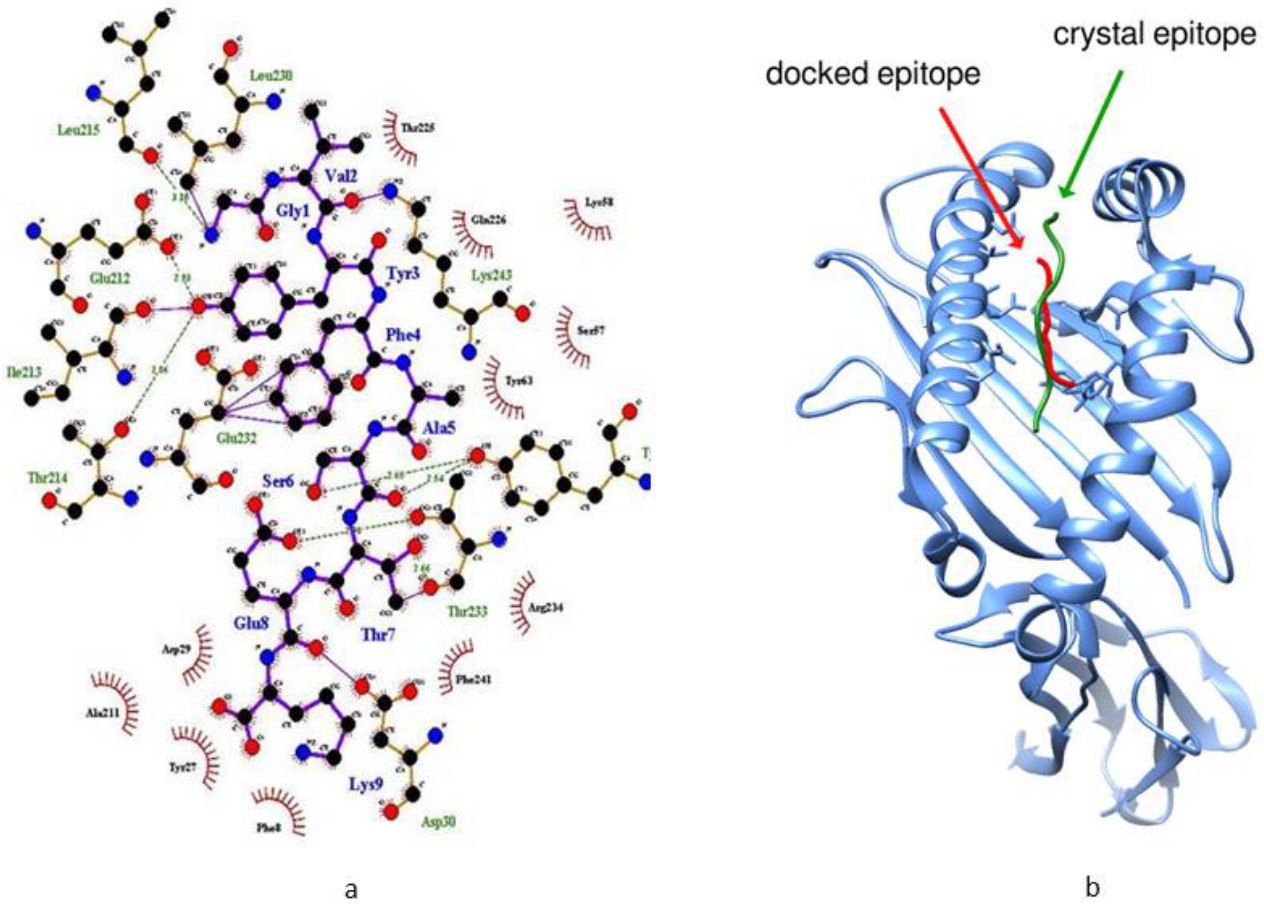

Fig 4a: Docking pose analysis via LigPlot [GVYFASTEK epitope docking against the HLA-A*11-01 allele (PDB ID: 5WJL)] Note: Molecular Docking result showing protein -ligand interaction; where Oxygen (O), Nitrogen (N) and Carbon (C) atoms are represented in red, blue and black circle.

Fig 4b: Molecular docking analysis showing docking site of ligand (GVYFASTEK epitope) in our study is similar to ligand used in crystal structure of HLA-A*11-01 allele (PDB ID: 5WJL).

Molecular Dynamics Simulation

Molecular Dynamics Simulation study of dock complex of GVYFASTEK epitope docked against the HLA-A*11-01 allele (PDB ID: 5WJL) was successfully executed for 50ns. The complex became stable throughout the simulation with RMSD fluctuation between 0.3-1.0 nm from the original position **(Fig.5a).** In most cases, residues lying in the core protein regions have low RMSF while exposed loops have high RMSF values **(Fig.5b).** As observed, the peaks

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the graph show a value between 0.1 and 0.6 nm. Both these results indicate that the protein complexes were stable throughout MD simulations and thus proteins possess the ability to stability.

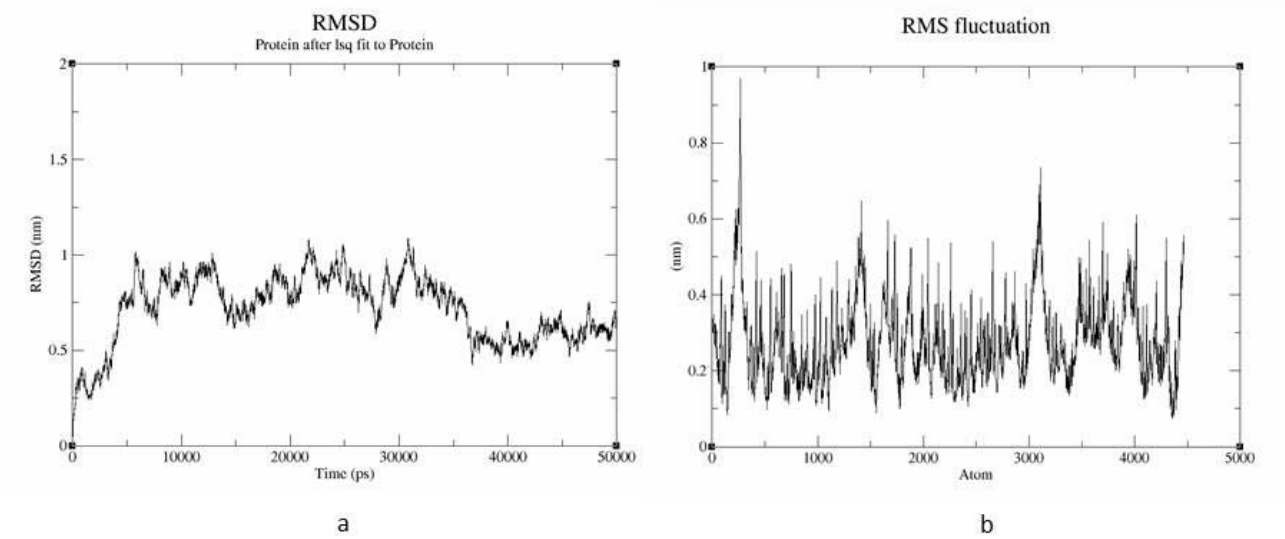

Fig 5: Molecular dynamics simulation- (a) RMSD graph (b) RMSF graph of dock complex [GVYFASTEK epitope docked against the HLA-A*11-01 allele (PDB ID: 5WJL)].

4. Discussion

Vaccine is enormously imperative and expansively formed therapeutic products. Millions of infants and people are getting vaccinated every year. However, the development and research processes of vaccines are expensive and occasionally, it requires countless months to prepare and advance an appropriate vaccine candidate towards killing of a pathogen. In today's scenario, innumerable tools and approaches of immunoinformatics, Computer-aided drug design (CADD), bioinformatics and converse/reverse vaccinology stance extensively castoff towards the progression of vaccine preparations, which intern reduce prolong duration and price investment on vaccine expansion (Islam et al., 2020; María, Arturo, Alicia, Paulina, & Gerardo, 2017; Ullah et al., 2020).

In current study, the physicochemical study revealed that surface glycoprotein **QIA98583.1** exhibited upper most extinction co-efficient of 148960M-1 cm-1 and lowest GRAVY value of -0.077. In addition, this selected surface glycoprotein was highly stable (instability index of less than 40) and antigenic. The antigenicity of the protein was determined by VaxiJen V2.0 server. If a compound required variability index greater than 40, means the product is considered as unbalanced (Guruprasad, Reddy, & Pandit, 1990). The extinction coefficient referred to the quantity

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of light, that is captivated by a complex at a particular wavelength (Ikai, 1980; Pace, Vajdos, Fee, Grimsley, & Gray, 1995). Various physicochemical properties including the amount of amino acids, molecular mass/weight, theoretic pI, extinction co-efficient, uncertainty index, aliphatic index, GRAVY was resolute by ProtParam (https://web.expasy.org/protparam/) server.

In the immune system, two major functioning cells are B and T lymphocytic cells, which are responsible for several defended roles in the body. Once identified by an antigen presenting cell or APC (such as dendritic cell, macrophage etc.), the antigen is accessible by the MHC class-II molecule existing on the exterior of these usual APCs, to the helper T cell. Subsequently, the helper T cell comprehends CD4+ fragment on its exterior, it is correspondingly termed as CD4+ T cell. Later stimulated by APC, the helper-T cells subsequently stimulate the B cell and reasons to yield antibody constructing plasma B cell alongside memory B cell. The plasma B cell harvests many antibodies and the memory B cell purposes as the immunological, elongated tenure memory. Nevertheless, macrophage and CD8+ cytotoxic T cell are also triggered by the helper-T cell that abolishes the target antigen (Arpin et al., 1995; Cano & Lopera, 2013; Goerdt & Orfanos, 1999; Pavli, Hume, Van De Pol, & Doe, 1993; Tanchot & Rocha, 2003).

The possible B and T cell epitopes of the designated *SARS-CoV-2* viral protein was identified by the IEDB (https://www.iedb.org/) server. The IEDB server generates and ranks the epitopes on the basis of their antigenicity scores (AS) and percentile scores. The top ten MHC class-I and class-II epitopes were engaged for investigation. The topology of the precise epitopes was resolute by TMHMM v2.0 server (http://www.cbs.dtu.dk/services/TMHMM/). In all the inflammatory situations such as allergenicity, antigenicity, toxicity and conservancy examination, the T-cell epitopes those were found to be exceedingly antigenic with higher immune response without allergies, non-toxic, and required conservancy of over 90%. Amongst ten certain MHC class-I and ten selected MHC class-II epitopes of the protein, over-all four epitopes were designated based on the revealed measures: GVYFASTEK, TLADAGFIK, NFRVQPTESI and LLIVNNATNVV as well as antigenic and non-allergenic B-cell epitopes were selected for added vaccine candidate investigation. The cluster examination of the conceivable MHC class-I and MHC class-II alleles that might interrelate with the prophesied epitopes were performed by the online tool MHC cluster 2.0 (http://www.cbs.dtu.dk/services/MHCcluster/). The antigenicity, demarcated as the capability of a extraneous ingredient to act as an antigen and stimulate the B and T cell replies, over their epitope correspondingly baptized antigenic determinant portion (Fishman, Wiles, & Wood, 2015). The

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

allergenicity is well-defined as the capability of that ingredient to act as an allergen and persuade latent allergic responses inside the host being (Andreae & Nowak-Węgrzyn, 2017).

Moreover, the cluster scrutiny of the MHC class-I and II alleles were similarly performed to categorize their association with each other and bunch them based on their functionality and foretold requisite specificity(Thomsen et al., 2013). Subsequently in the following steps, the peptide-protein docking was performed amongst the chosen epitopes and the MHC alleles. The MHC class-I epitopes stayed docked to the MHC class-I molecule (PDB ID: 5WJL) and the MHC class-II epitopes were docked with the MHC class-II molecule (PDB ID: 5JLZ) correspondingly. The peptide-protein docking was executed to evaluate the capability of the epitopes to quandary with their particular MHC molecules. Pre-docking was performed by UCSF-chimera and afterward, 3D structure generation of the epitopes was performed. The docking was executed by PatchDock and FireDock servers and analyzed by HPEPDOCK server constructed on global energy. The GVYFASTEK engendered the best scores in the peptide-protein docking respectively. All the vaccine candidates were proved to be potentially antigenic and non-allergenic, for this reason they should not cause any allergenic reaction within the host body. However, more *in vitro* and *in vivo* examinations should be performed to lastly approve the security, usefulness and potentiality of the predicted vaccines candidates.

5. Conclusion

In the face of enormous tragedy of suffering, demise and social adversity caused by COVID-19 pandemic. It is of the extremely importance to develop an effectual and safe vaccine against this highly pandemic disease. Bioinformatics, Reverse vaccinology and related technologies are widely used in vaccine design and development since these technologies reduce the cost and time. In this study, first the potential proteins belong to *SARS-CoV-2* against (host: human, country: India) are identified. Further, the potential B cell and T cell epitopes that can effectively elicit cellular mediated immune response related to these selected proteins were determined through robust processes. These potential T-cell epitope GVYFASTEK and B-cell epitopes LTPGDSSSGWTAG, VRQIAPGQTGKIAD, QIAPGQTGKIAD and ILPDPSKPSKRS play vigorous protagonist in the subunit and multi-epitope vaccine construction soon. In brief, reverse vaccinology vindicated as an authoritative means to recognize novel vaccine candidates and their consequential exact application. This study will lead to the research in an

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

innovative and efficient direction and the outcome of our study will deliver a fast, reliable and significant platform in search of effective and timely cure of this treacherous pandemic disease, COVID-19 caused by *SARS-CoV2*.

Acknowledgement: RKM acknowledges the financial support and award of Ramalingaswami fellowship from Department of Biotechnology, New Delhi, India. Authors highly acknowledge the Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur and Dr. B. Lal Institute of Biotechnology, Jaipur.

Funding:

This research did not receive any specific grant from funding agencies.

Compliance with Ethical Standards:

Conflict of interest

The authors declare there is no conflict of interest.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

6. References

Adeoye, A. O., Oso, B. J., Olaoye, I. F., Tijjani, H., & Adebayo, A. I. (2020). Repurposing of chloroquine and some clinically approved antiviral drugs as effective therapeutics to prevent cellular entry and replication of coronavirus. *Journal of Biomolecular Structure and Dynamics*(just-accepted), 1-14.

Andreae, D., & Nowak-Węgrzyn, A. (2017). The Effect of Infant Allergen/Immunogen Exposure on Long-Term Health. In *Early Nutrition and Long-Term Health* (pp. 131-173): Elsevier.

Arpin, C., Dechanet, J., Van Kooten, C., Merville, P., Grouard, G., Briere, F., . . . Liu, Y.-J. (1995). Generation of memory B cells and plasma cells in vitro. *Science, 268*(5211), 720-722.

Cano, R. L. E., & Lopera, H. D. E. (2013). Introduction to T and B lymphocytes. In *Autoimmunity: From Bench to Bedside [Internet]*: El Rosario University Press.

Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics, 8*(1), 4.

Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.-J., & Jiang, S. (2009). The spike protein of SARS-CoV—a target for vaccine and therapeutic development. *Nature Reviews Microbiology, 7*(3), 226-236.

Enayatkhani, M., Hasaniazad, M., Faezi, S., Guklani, H., Davoodian, P., Ahmadi, N., . . . Ahmadi, K. (2020). Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: an in silico study. *Journal of Biomolecular Structure and Dynamics*(just-accepted), 1-19.

Enjuanes, L., Zuñiga, S., Castaño-Rodriguez, C., Gutierrez-Alvarez, J., Canton, J., & Sola, I. (2016). Molecular basis of coronavirus virulence and vaccine development. In *Advances in virus research* (Vol. 96, pp. 245-286): Elsevier.

Fishman, J. M., Wiles, K., & Wood, K. J. (2015). The acquired immune system response to biomaterials, including both naturally occurring and synthetic biomaterials. In *Host Response to Biomaterials* (pp. 151-187): Elsevier.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Goerdt, S., & Orfanos, C. E. (1999). Other functions, other genes: alternative activation of antigen-presenting cells. *Immunity*, *10*(2), 137-142.

Gupta, E., Gupta, S. R., & Niraj, R. R. K. (2019). Identification of drug and vaccine target in mycobacterium leprae: a reverse vaccinology approach. *International Journal of Peptide Research and Therapeutics*, 1-14.

Guruprasad, K., Reddy, B. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection*, *4*(2), 155-161.

Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., . . . Tural, A. (2020). First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine*.

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *The Journal of Biochemistry*, *88*(6), 1895-1898.

Islam, R., Parves, R., Paul, A. S., Uddin, N., Rahman, M. S., Mamun, A. A., . . . Halim, M. A. (2020). A Molecular Modeling Approach to Identify Effective Antiviral Phytochemicals against the Main Protease of SARS-CoV-2. *Journal of Biomolecular Structure and Dynamics*(just-accepted), 1-20.

Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., & Tufféry, P. (2016). PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research*, *44*(W1), W449-W454.

María, R., Arturo, C., Alicia, J. A., Paulina, M., & Gerardo, A. O. (2017). *The impact of bioinformatics on vaccine design and development*: InTech, Rijeka, Croatia.

Meunier, M., Guyard-Nicodème, M., Hirchaud, E., Parra, A., Chemaly, M., & Dory, D. (2016). Identification of novel vaccine candidates against Campylobacter through reverse vaccinology. *Journal of immunology research*, *2016*.

Mishra, D. (2020). T Cell Epitope-Based Vaccine Design for Pandemic Novel Coronavirus 2019-nCoV.

Mishra, S., & Sinha, S. (2009). Immunoinformatics and modeling perspective of T cell epitope-based cancer immunotherapy: a holistic picture. *Journal of Biomolecular Structure and Dynamics*, *27*(3), 293-305.

Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein science*, *4*(11), 2411-2423.

Pant, S., Singh, M., Ravichandiran, V., Murty, U., & Srivastava, H. K. (2020). Peptide-like and small-molecule inhibitors against Covid-19. *Journal of Biomolecular Structure and Dynamics*(just-accepted), 1-15.

Pavli, P., Hume, D., Van De Pol, E., & Doe, W. (1993). Dendritic cells, the major antigen-presenting cells of the human colonic lamina propria. *Immunology*, *78*(1), 132.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605-1612.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., . . . Gu, Z. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research*, *40*(D1), D593-D598.

Shen, Y., Maupetit, J., Derreumaux, P., & Tufféry, P. (2014). Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *Journal of chemical theory and computation*, *10*(10), 4745-4758.

Sinha, S. K., Shakya, A., Prasad, S. K., Singh, S., Gurav, N. S., Prasad, R. S., & Gurav, S. S. (2020). An in-silico evaluation of different Saikosaponins for their potency against SARS-CoV-2 using NSP15 and fusion spike glycoprotein as targets. *Journal of Biomolecular Structure and Dynamics*(just-accepted), 1-13.

Tanchot, C., & Rocha, B. (2003). CD8 and B cell memory: same strategy, same signals. *nature immunology*, *4*(5), 431-432.

Thévenet, P., Shen, Y., Maupetit, J., Guyon, F., Derreumaux, P., & Tuffery, P. (2012). PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic acids research*, *40*(W1), W288-W293.

Thomsen, M., Lundegaard, C., Buus, S., Lund, O., & Nielsen, M. (2013). MHCcluster, a method for functional clustering of MHC molecules. *Immunogenetics*, *65*(9), 655-665.

Ullah, M. A., Sarkar, B., & Islam, S. S. (2020). Exploiting the Reverse Vaccinology Approach to Design Novel Subunit Vaccine against Ebola Virus. *medRxiv*.

Umesh, Kundu, D., Selvaraj, C., Singh, S. K., & Dubey, V. K. (2020). Identification of new anti-nCoV drug chemical compounds from Indian spices exploiting SARS-CoV-2 main protease as target. *Journal of Biomolecular Structure and Dynamics*(just-accepted), 1-9.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., . . . Peters, B. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic acids research*, *47*(D1), D339-D343.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.13.039198; this version posted October 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Walker, J. M. (2005). *The proteomics protocols handbook*: Springer.

Wallace, A. C., Laskowski, R. A., & Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering, design and selection*, *8*(2), 127-134.

Zhou, P., Jin, B., Li, H., & Huang, S.-Y. (2018). HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. *Nucleic acids research*, *46*(W1), W443-W450.