1   **Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2**

2   **etiologic agent of global pandemic COVID-19: an *in silico* approach**

3

4   M. Shaminur Rahman[1*], M. Nazmul Hoque[1,2*], M. Rafiul Islam[1*], Salma Akter[1,3], A. S. M.

5   Rubayet-Ul-Alam[1,4], Mohammad Anwar Siddique[1], Otun Saha[1], Md. Mizanur Rahaman[1**],

6   Munawar Sultana[1], M. Anwar Hossain[1,5**]

7   [1]Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh

8   [2]Department of Gynecology, Obstetrics and Reproductive Health, Bangabandhu Sheikh Mujibur

9   Rahman Agricultural University, Gazipur-1706, Bangladesh

10   [3]Department of Microbiology, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

11   [4]Department of Microbiology, Jashore University of Science and Technology, Jashore 7408,

12   Bangladesh

13   [5]Present address: Vice-Chancellor, Jashore University of Science and Technology, Jashore 7408,

14   Bangladesh

15   *Equal contribution

16   **Corresponding to:
17
18   M. Anwar Hossain, PhD
19   Professor
20   Department of Microbiology
21   University of Dhaka, Dhaka, Bangladesh
22   E-mail: hossaina@du.ac.bd
23   Or,
24   Dr. Md. Mizanur Rahaman
25   Assistant Professor
26   Department of Microbiology
27   University of Dhaka, Dhaka, Bangladesh
28   E-mail: razu002@du.ac.bd
29
30
31
32

1

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the ongoing pandemic of coronavirus disease 2019 (COVID-19), a public health emergency of international concern declared by the World Health Organization (WHO). An immuno-informatics approach along with comparative genomic was applied to design a multi-epitope-based peptide vaccine against SARS-CoV-2 combining the antigenic epitopes of the S, M and E proteins. The tertiary structure was predicted, refined and validated using advanced bioinformatics tools. The candidate vaccine showed an average of $\geq$ 90.0% world population coverage for different ethnic groups. Molecular docking of the chimeric vaccine peptide with the immune receptors (TLR3 and TLR4) predicted efficient binding. Immune simulation predicted significant primary immune response with increased IgM and secondary immune response with high levels of both IgG1 and IgG2. It also increased the proliferation of T-helper cells and cytotoxic T-cells along with the increased INF-γ and IL-2 cytokines. The codon optimization and mRNA secondary structure prediction revealed the chimera is suitable for high-level expression and cloning. Overall, the constructed recombinant chimeric vaccine candidate demonstrated significant potential and can be considered for clinical validation to fight against this global threat, COVID-19.

Keywords: SARS-CoV-2, Muti-epitope, Chimeric Peptide Vaccine, B-cell Epitope, T-cell Epitope.

57

Introduction

59   Emergence of novel coronavirus, known as severe acute respiratory syndrome

60   coronavirus 2 (SARS-CoV-2), which was first reported in Wuhan, Hubei Province, China in

61   December 2019 is responsible for the ongoing global pandemic of coronavirus disease 2019

62   (COVID-19). The ongoing outbreak of COVID-19 has already made its alarmingly quick

63   transmission across the globe in 199 countries and regions with a total of 723,077 active

64   infection cases and 33,983 deaths until March 29, 2020[1]. World Health Organization (WHO) has

65   declared a global health emergency, and scientists are racing over the clock to develop effective

66   interventions for controlling and preventing COVID-19 by analyzing novel SARS-CoV-2

67   genomics, functional structures, and its host-pathogen interactions. Efforts to contain the virus

68   are ongoing; however, given the many uncertainties regarding pathogen transmissibility and

69   virulence, the effectiveness of these efforts is unknown.

70   This SARS-CoV-2 is the third coronovairus (CoV) that can infect human after the two

71   previously reported coronavirus- severe acute respiratory syndrome (SARS-CoV)[2,3] and Middle

72   East respiratory syndrome (MERS-CoV)[4-6]. Alike SARS-CoV and MERS-CoV, the recent

73   SARS-CoV-2 is a positive-sense single-stranded RNA virus (+ssRNA) belonging to Genus

74   *Betacoronavirus* within the family Coronaviridae[7]. The genome of SARS-CoV-2 is

75   approximately 30 kilobases (between 26,000 and 32,000 bases) in length[8], and encodes for

76   multiple structural and non-structural proteins[7]. The four major structural proteins of CoVs are

77   the spike (S) glycoprotein, small envelope protein (E), membrane protein (M), and nucleocapsid

78   protein (N)[7]. Among these proteins, the S glycoprotein is composed of two main domains, the

79   receptor-binding domain (RBD) and N-terminal domain (NTD),[9,10] which play essential role for

3

80   the attachment and entry of viral particles into the host cells[11,12] and characterized by high

81   antigenicity and surface exposure[2,11,13].

82   To combat the devastating effects of SARS-CoV-2, Scientists applied different vaccine

83   strategies globally. In addition, a curative approach using different drug candidates have been

84   tried so far to combat the current pandemic COVID-19. However, COVID-19 is marching too

85   fast and reached almost every territory in earth. Hence, apart from developing novel drugs

86   against this viral infection, a very potent vaccine candidate is very much desirable. Being an

87   RNA virus, the virus had undergone mutations several times already in the last 4 months[14]. As a

88   result, instead of single epitope, multi-epitope based vaccines can play a great role in fighting

89   against this novel CoV.

90   On coronaviruses, the component primarily targeted by monoclonal antibodies (mAbs) is

91   the trimeric S glycoprotein of the virion consisting of three homologous chains (A, B and C),

92   which mediate receptor recognition and membrane fusion during viral entry into sensitive cells

93   through the host angiotensin-converting enzyme 2 (ACE2) receptor[5,9,11,13]. However, with the

94   rapid evolution and high genomic variation of RNA viruses, mutations on the RBD may enable

95   the new strains to escape neutralization by currently known RBD-targeting antibodies.

96   Therefore, other functional regions of the S glycoprotein can be important for developing

97   effective prophylactic and therapeutic interventions against SARS-CoV-2 infection. Several

98   mAbs targeting non-RBD regions, especially the NTD has recently been reported[13,15]. The NTD

99   is located on the side of the spike trimer and has not been observed to undergo any dynamic

100   conformational changes[11], and therefore, might play role in viral attachment, inducing

101   neutralizing antibody responses and stimulates a protective cellular immunity against SARS-

102   CoV[2,6,11]. Besides, administration of RBD and NTD proteins also induced highly potent

103   neutralizing antibodies and long-term protective immunity in animal models[9]. Furthermore, the E

104   and M proteins also have important functions in the viral assembly of a coronavirus and can

105   augment the immune response against SARS-CoV[11,16,17]. Monoclonal antibodies with potent

106   neutralizing activity have become promising candidates for both prophylactic and therapeutic

107   interventions against SARS-CoV-2 infections[13]. Therefore, the generation of antibodies targeting

108   the RBD and/or NTD of the S glycoprotein, M and E proteins of SARS-CoV-2 would be an

109   important preventive and treatment strategy that can be tested further in suitable models before

110   clinical trials[18]. The nucleocapsid (N) protein of SARS-coronavirus (SARS-CoV), buried inside

111   phospholipid bilayer, is the major protein in the helical nucleocapsid of virion. This protein is

112   reported to be more conserved than other structural proteins, an important B cell immunogen, as

113   well as, can elicit long☐lasting and persistent cellular immune responses. Nevertheless, we did

114   not consider this protein in the chimeric vaccine formation because of its initial unavailability

115   outside of the host cell during infection[19].

116   Most previous studies of the CoVs have focused on the use of single recombinant

117   antigens, however, the immune responses generated have been inadequate to warrant their use in

118   the development of an effectual protective tool[20,21]. Alternatively, we propose the development

119   of a multi-epitope vaccine candidate which could lead to the generation of a more potent

120   protective immune response. Multi-epitope vaccine candidates have already been designed for

121   several diseases, including FMD (in our laboratory; unpublished data), MERS and SARS, and

122   their efficacies have been further reported in MERS-CoV and SARS-CoV[2,6,22]. Hence, we can

123   assume that chimeric vaccine targeting multiple epitopes on the RBD and NTD segments of the

124   S protein, M and E proteins would be a potentially effective vaccine candidate in combatting

125   SARS-CoV-2, and therefore, could be used for vaccine design. Combining subunit vaccines with

126 established or new adjuvants may represent a faster and safer strategy to move through early

127 clinical development and analyze the efficacy. This study was therefore aimed to *in-silico* design

128 and construction of a multi-epitope-based chimeric vaccine candidate against the globally

129 pandemic highly pathogenic SARS-CoV-2.

130

Results

131

Retrieving protein sequences

132

133   We retrieved the spike (S), envelop (E) and membrane (M) protein sequences of the

134 SARS-CoV-2, and the S protein sequences of the SARS-CoV and MERS-CoV from the whole

135 genome reference sequences of the respective three viruses from the NCBI database. To map the

136 structural variability of the S protein of SARS-CoV-2, SARS-CoV and MERS-CoV, we applied

137 homology search on the respective S protein sequences, and finally selected three protein data

138 bank (PDB) templates such as 6vsb, 6acd and 5w9h, respectively for predicting the three

139 dimensional (3D) conformation of the S proteins. Finally, the 3D conformation of the S proteins

140 of the respective three viruses were validated using the Ramachandran plot analysis

141 (Supplementary Fig. 1).

142

Physicochemical properties and trimeric structure of spike (S) protein

144   The physiochemical properties of the spike (S) glycoproteins of the SARS-CoV-2,

145 SARS-CoV and MERS-CoV computed using the Protparam on the ExPASy server demonstrates

146 that it varies significantly among the three viruses. The theoretical isoelectric points (PI) of the

147 subject proteins were 6.24, 5.56 and 5.70, respectively. Protparam computed instability-index

148   (II) of the S protein of SARS-CoV-2, SARS-CoV and MERS-CoV were 33.01, 32.42 and 36.60

149   indicating that these proteins are quite stable[4,23].

150   The S protein of the SARS-CoV-2, SARS-CoV and MERS-CoV is a trimeric

151   conformation consisting of three homologous chains naming as chain A, B and C[5,10] (Fig. 1). The

152   individual structural domain of coronavirus S proteins reflects high degree of structural

153   divergences in the receptor-binding domain (RBD) and N-terminal domain (NTD) of the chain A

154   and chain C compared to that of chain B. Multiple sequence alignment revealed that the S

155   protein of SARS-CoV-2 shares 77.38% and 31.93% sequence identity with the S proteins of the

156   SARS-CoV and MERS-CoV, respectively, and these findings are in line with many earlier

157   studies[3-7] (Supplementary Fig. 2). Due to structural divergence in RBD and NTD of S proteins in

158   SARS-CoV-2, SARS-CoV and MERS-CoV, these regions are the main focus of structural

159   studies including epitope-based chimeric vaccine candidate designing, selection, and

160   development.

161

162   **Structure-based B-cell epitopes prediction**

163   Linear epitopes prediction based on solvent-accessibility and flexibility revealed that

164   chain A of S protein possessed 15 epitopes, of which three epitopes having different residues

165   positions (395-514, 58-194, 1067-1146) were highly antigenic (score > 0.8). However, the B and

166   C chains had 18 and 19 epitopes, respectively, and of them three epitopes in chain B (residues

167   position 1067-1146, 89-194, 58-87) and two epitopes in chain C (residues position 56-194, 1067-

168   1146) seem to be highly antigenic (score > 0.8) (Table 1). The amino acid residues 395-514 and

169   56-194 of the detected epitopes belonged to RBD and NTD regions of the S protein,

170   respectively. These regions were considered as the potential epitope candidates in the IEDB

7

171   analysis resource Elipro analysis. However, the epitopes with amino-acids position of 1067-1146

172   were not selected as the potential epitope candidate because of their presence in viral

173   transmembrane domains (Supplementary Fig. 3). The tertiary structures of the RBD and NTD

174   illustrate their surface-exposed position on the S protein (Fig. 2).

175

176   **Sequence-based B-cell epitopes**

177   Antigenicity analysis of full-length S glycoprotein, M (membrane) and E (envelope)

178   proteins using VaxiJen v2.0 showed antigenicity scores of 0.465, 0.510 and 0.603, respectively

179   exhibiting them as potential antigens. Therefore, they were considered to be promising epitopes

180   for vaccine candidate selection against globally pandemic SARS-CoV-2 infections. A total of 22

181   B-cell epitopes were predicted in S protein of the SARS-CoV-2 using IEDB analysis resource

182   and Bepipred linear epitope prediction 2.0 tools. Of the detected epitopes, eight and six epitopes

183   were respectively found in RBD (aa position 328-528) and NTD (aa position 56-194) regions

184   while the E and M proteins had 2 and 6 epitopes, respectively (Fig. 3, Supplementary Table 1).

185   However, only 5 epitopes were exposed on the surface of the virion, and had a high antigenicity

186   score (> 0.4), indicating their potentials in initiating immune responses (Table 2). To find out the

187   most probable peptide-based epitopes with better confidence level, selected peptides were further

188   tested using VaxiJen 2.0 tool, and five peptides having antigenicity score of ≥ 0.5 were

189   annotated. Among the 5 annotated epitopes, RBD region had two highly antigenic epitopes (75-

190   IRGDEVRQIAPGQTGKIADYNYKLPD-100, antigenicity score = 0.932 and 166-

191   QSYGFQPTNGVGYQ-189, antigenicity score = 0.668). Similarly, the NTD region harbored

192   two highly antigenic epitopes (117-SQPFLMDLEGKQGNFKNLRE-136, antigenicity score =

193   0.749 and 17-GTNGTKRFDN-26, antigenicity score = 0.667). In addition, the envelop (E)

8

194   protein possessed only one highly antigenic epitope (57-YVYSRVKNLNSSRVP-71,

195   antigenicity score = 0.449) which might be potentially functional in host cell binding (Table 2).

196   Furthermore, the Kolaskar and Tongaonkar antigenicity profiling found five highly antigenic

197   epitopes in RBD region with an average (antigenicity) score of 1.042 (minimum = 0.907,

198   maximum = 1.214), and seven highly antigenic epitopes in NTD with an average (antigenicity)

199   score of 1.023 (minimum = 0.866, maximum = 1.213) (Supplementary Fig. 4, Supplementary

200   Table 2). The average Kolaskar scores for envelop protein B-cell epitope (EBE) and membrane

201   protein B-cell epitope (MBE) were 0.980 and 1.032, respectively (Supplementary Table 2).

202   However, through ABCPred analysis, we identified 18 and 11 B-cell epitopes in RBD and NTD

203   regions with average antigenicity score of 0.775 and 0.773 in the associated domains,

204   respectively (Supplementary Table 3).

205

206   **Recognition of T-cell epitopes**

207   The IEDB analysis resource tool was employed to identify T-cell epitopes in RBD and

208   NTD of the S protein of SARS-CoV-2. The prediction method included the major

209   histocompatibility complex class I and Π (MHC-I, MHC-Π), proteasomal cleavage and TAP

210   transport. The RBD and NTD regions were analyzed using IEDB MHC- I binding prediction tool

211   to predict T lymphocytes epitopes that have binding affinity with MHC-I alleles based on

212   Artificial Neural Network (ANN) with half-maximal inhibitory concentration (IC_{50}) $\leq$ 50 nM. A

213   lower IC_{50} value indicates higher binding affinity of the epitopes with the MHC class I and II

214   molecules. While most of the studies[24,25] reported that a binding affinity (IC_{50}) threshold of 250

215   nM identifies peptide binders recognized by T-cells, and this threshold can be used to select

216   peptides, we kept binding affinity within 50 nM to get better confidence level in predicting

9

217   epitopes for MHC-I and MHC-II alleles. The IEDB MHC-I prediction tool retrieved 77 T-cell

218   epitopes in RBD that interacted with 21 possible MHC-I alleles whereas the NTD domain

219   possessed 35 T-cell epitopes with 17 possible MHC-I alleles (Supplementary Data 1). Similarly,

220   the IEDB MHC-II prediction tool generated 13-mer124 peptides from the RBD, and 10-mer 73

221   peptides in the NTD segments of the S protein that showed interaction with many different

222   and/or common MHC-II alleles with an IC_{50} value ranging from 1.4 to 49.9 nM (Supplementary

223   Data 1).   Furthermore, for MHC-I and MHC-II processing, the analysis tool of the IEDB

224   generates an overall score for each epitope's intrinsic potential of being a T-cell epitope based on

225   proteasomal processing, TAP transport, and MHC-binding efficiency (Supplementary Data 1).

226   The outcomes of these tools are quite substantial because they utilize vast number of alleles of

227   HLAs (human-leukocyte-antigens) during computation.

228

Molecular docking analysis of T-cell epitopes with HLA alleles

230   From the selected epitopes from the RBD and NTD segments, top five based on IC_{50}

231   score were used in molecular docking analysis using the GalaxyWeb server with their respective

232   HLA allele binders, in which they revealed significantly favorable molecular interaction for

233   binding affinity. Docking complexes thus formed have significantly negative binding energy,

234   and most of the aa residues of the epitopes were involved in molecular interactions with their

235   respective HLA alleles (Supplementary data 1). The epitope-HLA docking complexes were

236   further refined with GalaxyRefineComplex, and their binding affinity was analyzed through

237   PRODIGY web-server. All of the selected epitopes showed significantly negative binding

238   affinity (ΔG always remained $\leq$ -8.2 kcal mol^{-1}, average = -9.94 kcal mol^{-1}, Fig. 4,

239   Supplementary data 1).

10

240

IFN-γ inducing epitope prediction

The findings of IFNepitope program suggests that, both the target RBD and NTD regions of S protein, and B-cell linear epitope (MBE) had great probability to release of IFN-γ with a positive score. A total of 56 potential positive IFN-γ inducing epitopes (15-mer) were predicted for the RBD domain with an average epitope prediction score of 0.255 and the maximum SVM score of 0.625. On the other hand, a total of 33 potential positive epitopes were predicted for the NTD domain with an average epitope prediction score of 0.312 and the maximum SVM score of 0.811. Moreover, the M protein also possessed several IFN-γ inducing epitopes having an average epitope prediction score of 0.980 (Supplementary Table 4).

Population coverage analysis

The distribution of HLA allele varies between different geographical regions and ethnic groups around the globe. Therefore, population coverage during the development of an efficient vaccine must be taken into account. In this study, epitopes having IC_{50} values ≤ 50 nM and their respective alleles (CTL and HTL) were considered for population coverage analysis both individually (MHC-I and MHC- II) and in combination (Supplementary Data 2). Our selected CTL and HTL epitopes were found to cover 94.9% and 73.11% of the world population, respectively. Importantly, CTL and HTL epitopes showed 98.63% population coverage worldwide when used in combination. The highest population coverage was found to be 99.99% in the Latin American country, Peru (Fig. 5, Supplementary Data 2). In China, where the viral strain (SARS-CoV-2) first appeared and had more devastating outbreaks, the population coverage for CTL and HTL epitopes was 92.67% and 53.44%, respectively with a combined

11

263   coverage of 96.59%. SARS-CoV-2 is currently causing serious pandemics in different continents

264   of the globe including Italy, England, Spain, Iran, South Korea and United States of America

265   where the combined population coverage was found to be 98.8%, 99.44%, 95.35%, 98.48%,

266   99.19% and 99.35%, respectively (Fig. 5a, Supplementary Data 2). In addition to geographical

267   distribution, the ethnic groups also found to be an important determinant for good coverage of

268   the CTL and HTL epitopes (Fig. 5b). Of the studied 147 ethnic groups, the Peru Amerindian had

269   highest population coverage for CTL (99.98%) while the HTL epitopes had highest population

270   coverage for Austria Caucasoid (88.44%) (Fig. 5b, Supplementary Data 2). Furthermore, 53.06%

271   of the ethnic groups had a combined population coverage of more than 90.0% for both CTL and

272   HTL epitopes.

273

274   **Design and construction of chimeric vaccine candidate**

275   The selected sequences for designing of chimeric construct were PADRE (13 aa), MBE

276   (20 aa), NTD (139 aa), RBD (200 aa), EBE (15 aa), and Invasin (16 aa), and the construct was

277   named as CoV-RMEN. The schematic diagram of protein domain structures, their junctions and

278   linker sites in CoV-RMEN are shown in Fig. 6a. These fragments were fused together using a

279   repeat of hydrophobic (glycine; G) and acidic aa (glutamic acid; E) linkers. The EGGE, GG, GG,

280   GG and EGGE sequences were introduced between different domains, and epitopes for more

281   flexibility and efficient separation with balancing of acidic and basic amino acids.

282

283   **Prediction of the antigenicity and allergenicity of the CoV-RMEN**

284   The antigenicity of the final chimeric protein sequence was predicted by the VaxiJen 2.0

285   server to be 0.450 with a virus model at a threshold of 0.4 and 0.875 with ANTIGENpro. The

12

286   results indicate that the generated CoV-RMEN sequences are both antigenic in nature. The

287   vaccine sequence without the adjuvant was also predicted to be non-allergenic on both the

288   AllerTOP v.2 and AllergenFP servers[21].

289

290   **Physiochemical properties and solubility prediction of the CoV-RMEN**

291   The molecular weight (MW) of the finally constructed vaccine candidate was predicted to

292   be 46.8 kDa with a theoretical isoelectric point value (pI) of 8.71. The protein is predicted to be

293   slightly basic in nature based on the pI. The half-life was assessed to be 4.4 hours in mammalian

294   reticulocytes *in vitro*, and >20 hours in yeast and >10 hours in *E. coli in vivo*. The protein was

295   predicted to be less soluble upon expression with a solubility score of 0.330 corroborating the

296   findings of many previous studies[21,26]. An instability index (II) of 29.74 was predicted,

297   classifying the protein as stable (II of > 40 indicates instability). The estimated aliphatic index

298   was predicted to be 66.59, indicating thermostability. The predicted Grand average of

299   hydropathicity (GRAVY) was –0.300. The negative value indicates the protein is hydrophilic in

300   nature and can interact with water molecules[21].

301

302   **Secondary and tertiary structures prediction of the CoV-RMEN**

303   The CoV-RMEN peptide was predicted to contain 43.2% alpha helix, 67.4% beta sheet,

304   and 12% turns (Fig. 6b, Supplementary Fig. 5) using CFSSP:Chou and Fasman secondary

305   structure prediction server. In addition, with regards to solvent accessibility of aa residues, 34%

306   were predicted to be exposed, 30% medium exposed, and 34% were predicted to be buried. Only

307   2 aa residues (0.0%) were predicted to be located in disordered domains by the RaptorX Property

308   server (Supplementary Fig. 6). The Phyre2 server predicted the tertiary structure model of the

13

309 designed chimeric protein in 5 templates (c5x5bB, c2mm4A, c6vsbB, c5x29B and c6vybB)

310 based on heuristics to maximize confidence, percent identity and alignment coverage. The final

311 model of the CoV-RMEN peptide modelled at 82% with more than 90% confidence (Fig. 6c).

312 Moreover, 65 residues were modelled by *ab initio*.

313

314

315

316 **Tertiary structure refinement and validation of the CoV-RMEN**

317   After energy minimization of the CoV-RMEN using GROMOS96 program[27], refinement

318 of the vaccine model on the ModRefiner server followed by further refinement on the

319 GalaxyRefine server yielded five models. The selected model has parameters scores of GDT-HA

320 (0.9538), RMSD (0.414), and MolProbity (2.035). The Ramachandran plot analysis of the finally

321 modelled protein revealed that 94.7% of residues in the protein are in favored regions (Fig. 6d),

322 consistent with the 94.0% score predicted by the GalaxyRefine analysis. Additionally, 4.8% of

323 the residues were predicted to be in allowed regions, and only 0.5% in disallowed region (Fig.

324 6d). The chosen model after refinement had an overall quality factor of 74.45% with ERRAT

325 (Supplementary Fig. 7) while ProSA-web gave a Z-score of -6.17 (Fig. 6e) for the CoV-RMEN

326 protein model.

327

328 **Immune simulation**

329   The immune-stimulatory ability of the predicted vaccine CoV-RMEN was conducted

330 through the C-ImmSimm server. The analysis predicts the generation of adaptive immunity in

331 target host species (human) using position-specific scoring matrix (PSSM), and machine learning

14

332   techniques for the prediction of epitopes and immune interactions[28]. The cumulative results of

333   immune responses after three times antigen exposure with four weeks interval each time revealed

334   that the primary immune response against the antigenic fragments was elevated indicated by

335   gradual increase of IgM level after each antigen exposure (Fig. 7a). Besides, the secondary

336   immune response, crucial for immune stability, have been shown as increased with adequate

337   generation of both IgG1 and IgG2. Also, the elevated level of all circulating immunoglobulins

338   indicates the accuracy of relevant clonal proliferation of B-cell and T-cell population. The level

339   of cytokines after antigen exposure increased concomitantly reflected by escalation of IFN-γ and

340   IL-2 which are most significant cytokines for anti-viral immune response and clonal selection

341   (Fig. 7b). The abundance of different types of B-cells and T-cells, like antigen processing B-

342   cells, resting memory B- and T- cells, B-cells with IgM and IgG remains significantly higher

343   indicating development of immune memory and consequently increased clearance of antigen

344   after exposure (Fig. 7c,d). Additionally, T-helper cells and cytotoxic T-cells were found with a

345   drastic up-regulation of Th1 concentration enhancing the B-cell proliferation and immune

346   memory development (Fig. 7e,f). The high level of immunoglobulin IgG1 + IgG2, active B-cell

347   and T-helper cell population reflected the development of strong immune response reinforcing

348   the indelible and peerless antigenicity of the CoV-RMEN vaccine candidate.

349

350   **Molecular docking of CoV-RMEN with immune receptors (TLR3 and TLR4)**

351   The ClusPro server was used to determine the protein binding and hydrophobic

352   interaction sites on the protein surface. The immune responses of TLR3 and TLR4 against

353   vaccine construct (CoV-RMEN) were estimated by analyzing the overall conformational stability

354   of vaccine protein-TLRs docked complexes. The active interface aa residues of refined

15

355   complexes of CoV-RMEN and TLRs were predicted (Fig. 8, Table 3). The relative binding free

356   energies (ΔG) of the protein-TLRs complexes were significantly negative (Table 3) which

357   suggest that the interaction of the chimeric protein might favor stimulation of the TLR receptors.

358   Interface contacts (IC) per property (ICs charged-charged: 16, ICs charged-polar: 22, ICs

359   charged-apolar: 26, polar-polar: 6, ICs polar-apolar: 25 and apolar-apolar: 29) were for the

360   vaccine protein-TLR3 complex. Also, vaccine protein-TLR4 complex showed similar (ICs) per

361   property (ICs charged-charged: 5, ICs charged-polar: 11, ICs charged-apolar: 30, polar-polar: 4,

362   ICs polar-apolar: 31 and apolar-apolar: 39). These data validate the selected docked complexes

363   that may result high binding among TLRs and the chimeric protein (Fig. 8).

364

365   **Codon optimization of the CoV-RMEN**

366   The high yield expression of a recombinant protein is always remained as a challenge,

367   whereas heterologous prokaryotic expression vector can potentiate a robust, low cost and user-

368   friendly technology for large scale vaccine production. For this reason, the recombinant protein

369   encoding sequence need to be optimized for escalation of protein expression and downstream

370   processing for a final product development. In order to optimize codon usage of the vaccine

371   construct CoV-RMEN in *E. coli* (strain K12) for maximal protein expression, Rare Codon

372   Analysis Tool (GenScript) was used. The length of the optimized codon sequence was 1,251

373   nucleotides. The Codon Adaptation Index (CAI) of the optimized nucleotide sequence was 0.87,

374   and the average GC content of the adapted sequence was 50.26% showing the possibility of good

375   expression of the vaccine candidate in the *E. coli* host. The optimal percentage range of GC

376   content is between 30% and 70% (Fig. 9 a,b,c).

377

16

378 **Prediction of mRNA secondary structure of the CoV-RMEN**

379   The evaluation of minimum free energy for 25 structures of chimeric mRNA, the

380 optimized sequences carried out by the 'Mfold'server. The results showed that ΔG of the best

381 predicted structure for the optimized construct was ΔG = -386.50 kcal/mol. The first nucleotides

382 at 5' did not have a long stable hairpin or pseudoknot. Therefore, the binding of ribosomes to the

383 translation initiation site, and the following translation process can be readily accomplished in

384 the target host. These outcomes were in the agreement with data obtained from the

385 'RNAfold'web server (Fig. 9 d,e) where the free energy was -391.37 kcal/mol.

386

387 **Expression of CoV-RMEN with SUMO-fusion**

388   After codon optimization and mRNA secondary structure analysis, the sequence of the

389 recombinant plasmid was designed by inserting the adapted codon sequences into pETite vector

390 (Lucigen, USA) using SnapGene software (Fig. 10). The utilization of pETite vector containing

391 SUMO (Small Ubiquitin-like Modifier) tag and 6x-His tag will facilitate both the solubilization

392 and affinity purification of the recombinant protein[29]. After successful cloning of the gene, the

393 recombinant plasmid can be propagated efficiently using *E. coli* cells, and subsequent protein

394 expression can be performed in *E. coli* K-12 strain using IPTG (Isopropyl β-d-1-

395 thiogalactopyranoside) induction and cultivation at 28 °C as also reported earlier[29].

396
397

398 **Discussion**

399   The causative agent of the ongoing COVID-19 pandemic caused by SARS-CoV-2, the

400 seventh coronavirus with high zoonotic importance and transmission rate, has proved to be

401 transmitted through direct and indirect contact along with airborne transmission which contribute

402   very potentially for community transmission[30-32]. Since the emergence of SARS-CoV-2 in 2019

403   in Wuhan, China[31,32], it has exceeded both SARS-CoV and MERS-CoV in its rate of

404   transmission among humans[30], and therefore, effective measures counteracting its infection have

405   become a major research focus. Scientific efforts have been going on extensively to learn about

406   the virus, its biological properties, epidemiology, and pathophysiology[7]. At this early stage,

407   different vaccine development efforts have been going on globally along with the novel drug

408   development strategies[33].

409   Development of an effective live attenuated vaccine for viral pathogens has been the

410   most practiced strategy for prevention. But due to its extremely high contagious nature and

411   ability for community transmission, this novel SARS-CoV-2 attenuated vaccine research

412   essentially requires high biosafety level for researchers and work stations, which may eventually

413   make it difficult to avail the vaccine for mass population of all the countries. Although live

414   attenuated vaccine is often very immunogenic, however, the risk of reversion to a virulent virus

415   has limited its usage as SARS-CoV and MERS-CoV vaccine[22]. Unlike other types of vaccines,

416   chimeric peptide vaccine production does not involve virus replication, therefore reduce the cost

417   of production. Hence, a low cost strategy should be adopted for developing a highly demanded

418   vaccine for the mankind. Heterologous expression of any vaccine candidate protein has very

419   promising scopes for developing such low cost vaccine, providing that all essential properties for

420   antigenicity, immunogenicity and functional configuration are being conserved to mimic the

421   structural and functional property of the actual antigen[34]. Construction of a vaccine candidate

422   with multiple potential epitopes can obviously potentiate the multi-valency of the antigen to

423   develop immune response against a number of epitopes of any pathogen. Also, rational

424   engineering of epitopes for increased potency and magnitude, ability to enhance immune

18

425   response in conserved epitopes, increased safety and absence of unnecessary viral materials and

426   cost effectiveness all these cumulatively include potential benefit to multi-epitope recombinant

427   protein based vaccine[20]. This study was designed to assist with the initial phase of multi-epitope

428   vaccine candidate selection. Thereby, safe and effective vaccine development by providing

429   recommendations of epitopes that may potentially be considered for incorporation in vaccine

430   design for SARS-CoV-2.

431   Pathophysiology and virulence mechanisms of coronaviruses (CoVs), and therefore

432   similarly of SARS-CoV-2 have links to the function of the structural proteins such as spike (S)

433   glycoprotein, envelop (E), and membrane (M) proteins[34,35]. Across the CoV families, the

434   antibodies against the S glycoprotein, M and E proteins of SARS-CoV-2 would provide

435   immunity to the infection[2,12,22]. Therefore, this work focused on the *in silico* design and

436   development of a potential multi-epitope vaccine for SARS-CoV-2 selecting epitopes and

437   domains from the S, E and M proteins.

438   Among the structural elements of CoVs, the spike S glycoproteins are composed of two

439   major domains, the RBD and NTD. Furthermore, the homotrimeric S proteins of the SARS-

440   CoV-2, SARS-CoV[9,10] and MERS-CoV[5] comprises three chains: the chain A, B and C on the

441   viral surface, guiding the link to host receptors. Unlike the full-length S protein of the CoVs, the

442   RBD and NTD segments possessed the critical neutralizing domains but lacking the non-

443   neutralizing immunodominant region[6,10,11,22]. Therefore, considering the safety and effectiveness

444   perspectives, the RBD and NTD are more promising candidates in the development of SARS-

445   CoV-2 vaccines over the full-length S protein. The presence of E and M proteins on the envelope

446   can augment the immune response against SARS-CoV[2,36] and thus, considered for suitable

19

447   candidate for vaccine development[7,12,22]. Therefore, antibodies against the S, M and E proteins

448   of SARS-CoV-2 would provide protective immunity to the infection[6,7,11,12,22].

449   Immuno-informatics analyses of the selected RBD and NTD regions indicated that these

450   contain ample amount of high-affinity B-cell, MHC Class I, MHC Class II and interferon-γ

451   (IFN-γ) epitopes with high antigenicity scores. Moreover, membrane B-cell epitope (MBE) and

452   envelop B-cell epitope (EBE) were selected from the M and E proteins, respectively for the final

453   vaccine construct to enhance the overall stability, immunogenicity and antigenicity of the

454   chimeric vaccine candidate, the CoV-RMEN[21]. The HLA alleles maintain the response to T-cell

455   epitopes, and in different ethnic communities and geographical regions, these alleles are highly

456   polymorphic. T-cell epitopes from RBD and NTD regions having high interaction with more

457   HLA alleles were found to cover more than 98% of the world population with different ethnic

458   groups[6,37,38]. Moreover, the molecular docking between T-cell epitopes and their respective HLA

459   alleles revealed their highly significant binding affinity reflecting the immune activation of B-

460   and T-cells[23].

461   Vaccine design is improved through the use of specialized spacer sequences[39]. To

462   designing the CoV-RMEN (vaccine candidate) GG and EGGE linkers were incorporated

463   between the predicted epitopes to produce sequences with minimized junctional

464   immunogenicity, thereby, allowing the rational design construction of a potent multi-epitope

465   vaccine[21,38]. The molecular weight of our vaccine candidate, the CoV-RMEN is 46.8 kDa with a

466   predicted theoretical pI of 8.71, indicating that the protein is basic in nature. Also, the predicted

467   instability index indicates that the protein will be stable upon expression, thus further

468   strengthening its potential for use. The aliphatic index showed that the protein contains aliphatic

469   side chains, indicating potential hydrophobicity. All these parameters indicate that the

20

470   recombinant protein is thermally stable, hence would be best suited for use in different endemic

471   areas worldwide[6,21].

472   The knowledge of secondary and tertiary structures of the target protein is essential in

473   vaccine design[39,40]. Secondary structure analysis of the CoV-RMEN indicated that the protein

474   consisted of 43.2% alpha helix, 67.4% beta sheet, and 12% turns with only 2 residues disordered.

475   Natively unfolded protein regions and alpha-helical and beta sheet peptides have been reported

476   to be important forms of structural antigens[21,40]. These two structural forms (secondary and

477   tertiary), when tested as the synthetic peptides, have the ability to fold into their native

478   structure,hence, could be recognized by naturally induced antibodies in response to infection[2].

479   The tertiary (3D) structure of the vaccine candidate improved markedly after the refinement, and

480   showed desirable properties based on Ramachandran plot predictions. The Ramachandran plot

481   shows that most of the residues are found in the favoured and allowed regions (94.7%) with very

482   few residues in the outlier region; this indicates that the quality of the overall model is

483   satisfactory[21,23]. The lack of allergenic properties of the CoV-RMEN further strengthens its

484   potential as a vaccine candidate. Strikingly, the multi-epitope peptide we designed and

485   constructed (CoV-RMEN), showed higher antigenicity scores on Vaxijen v2.0 (0.450 with a

486   virus model at a threshold of 0.4) and ANTIGENpro (0.875) gave an antigenic score almost

487   twice that of Ov-103 on ANTIGENpro (0.93 for the chimeric peptide and 0.59 for Ov-103)

488   supporting that this multiple-epitopes carrying vaccine candidate would be poorly immunogenic

489   and require coupling to adjuvants. However, the designed protein must show similar antigenicity

490   scores with and without the addition of an adjuvant sequence[21].

491   It has been reported that immunity to viral infections is dependent on both B- and T-

492   cells[20,21]. Toll-Like Receptors (TLRs) recognize Pathogen Associated Molecular Patterns

21

493   (PAMPs) from viral components or replication intermediates, resulting in signaling cascades that

494   initiate an antiviral state in cells as a result of infection[41]. However, it is likely that TLRs might

495   play an important role in the innate immune response to SARS-CoV-2 infection[42]. Though, we

496   did not use any adjuvants, it is reported that TLRs (TLR3 and TLR4) can effectively bind with

497   spike protein of the CoV[43]. The molecular docking analysis of the CoV-RMEN through

498   HADDOCK showed that the chimeric protein can establish stable protein-protein interactions

499   with TLRs (TLR-3,TLR-4)[41] indicating efficient activation of these surface molecules which is

500   very crucial for immune activation of dendritic cells for subsequent antigen processing and

501   presentation to CD4+ and CD8+ T-cells via MHC-II and MHC-1, respectively corroborating the

502   findings of different earlier studies[11,20,21]. TLR☐4 might bind to S protein leading to the

503   activation of the MyD88☐dependent signaling pathway, which ultimately releases

504   proinflammatory cytokines[35]. Immune simulation of the designed CoV-RMEN showed results

505   consistent with typical immune responses, and there was a general increase in the generated

506   immune responses after repeated antigen exposures. The comprehensive immune response

507   developed after repeated antigen exposure reflected a perfect host defence development with the

508   uplifted primary immune response and subsequent secondary immune response resulting in

509   immune memory development[2]. In the present study, the development of memory B-cells and T-

510   cells was evident, with memory in B-cells lasting for several months. Helper T- cells were

511   particularly stimulated. The engrossing findings of the study is the development of Th1 response

512   which enhance the growth and proliferation of B- cells augmenting the adaptive immunity[44]. The

513   antiviral cytokine IFN-γ and cell stimulatory IL-2 level significantly increased which also

514   contribute to the subsequent immune response after vaccination in host[2]. This indicates high

515   levels of helper T-cells and consequently efficient antibody production, supporting a humoral

22

516   response[21,45]. The Simpson index, D for investigation of clonal specificity suggests a possible

517   diverse immune response. This is plausible considering the generated chimeric peptide is

518   composed of sufficient B- and T-cell epitopes.

519   One of the first steps in validating a candidate vaccine is to screen for immunoreactivity

520   through serological analysis. This requires the expression of the recombinant protein in a suitable

521   host. *Escherichia coli* expression systems are the preferred choice for the production of

522   recombinant proteins[33]. In order to achieve high-level expression of our recombinant vaccine

523   protein in *E. coli* (strain K12), codon optimization was carried out. Stable mRNA structure,

524   codon adaptability index (0.87), and the GC content (50.26%) were favourable for high-level

525   expression of the protein in bacteria. The next step projected at the moment, is to express this

526   peptide in a bacterial system and perform the various immunological assays needed to validate

527   the results obtained here through immuno-informatics analyses.

528

529   **Conclusions**

530   COVID-19 is reaching entire globe and emergence of the causative agent SARS-CoV-2

531   kept us nowhere as we are racing for treatment and prevention. This multi-epitope chimera that

532   we named as CoV-RMEN possess domains of RBD and NTD segments of S, M and E proteins

533   all of which showed significant antigenic properties compared to any other viral proteins. This

534   chimera also includes potential CTL, HTL and B-cell epitopes to ensure humoral as well as

535   cellular immune response and the optimal expression and stability of the chimera was validated.

536   With multiple limitations and high cost requirements for the attenuated vaccine preparation for

537   contagious agents like SARS-CoV-2, this chimeric peptide vaccine candidate gives us the hope

538   to ensure it's availability and relatively cheap option to reach entire world. This CoV-RMEN can

539 be very effective measure against COVID-19 to reach globally. Hence, this could be cloned,

540 expressed and tried for *in vivo* validations and animal trials at the laboratory level.

541

542

543 **Methods**

544 **Sequence retrieval and structure generation**

545 A total of 250 partial and complete genome sequences of SARS-CoV-2 were retrieved

546 from NCBI (Supplementary Table 5). We aligned these sequences through MAFFT online

547 server (https://mafft.cbrc.jp/alignment/server/) using default parameters, and Wu-Kabat protein

548 variability was analyzed in Protein variability server (http://imed.med.ucm.es/PVS/) with

549 respect   to   NCBI   (National   Center   for   Biotechnology   Information,

550 https://www.ncbi.nlm.nih.gov/protein) reference genome (Accession no : NC_045512.2).

551 Despite having minor heterogeneity in spike glycoprotein (S), membrane (M) and envelop (E)

552 proteins, the NCBI reference genome (Accession no: NC_045512.2) of SARS-CoV-2 from

553 Wuhan, China was finally selected for domains and epitopes selection, secondary and tertiary

554 (3D) structure prediction, antigenicity and allergenicity assessment, refinement and validation,

555 molecular docking, cDNA and mRNA analysis for cloning and expression[20,21]. Moreover, two

556 reference genome sequences of SARS-CoV (NC_004718.3) and MERS-CoV (NC_019843.3)

557 were also retrieved from the NCBI database to reveal the structural heterogeneity of S protein.

558 Multiple sequence alignment of the S proteins of the viruses was performed by CLUSTALW[45].

559 Since the S glycoprotein of coronaviruses is important for viral attachment to host cell, antibody

560 against this protein acts as inhibitor[30], therefore, S proteins of SARS-CoV, MERS-CoV and

561 SARS-CoV-2 were structurally compared using SWISS homology modeling[46] based on the

24

562   protein databank (PDB) templates 6acd, 5w9h and 6vsb, respectively. The models were then

563   optimized by energy minimization using the GROMOS96 program[27] implemented within the

564   Swiss-PdbViewer, version 4.1.0[47]. The Ramachandran plots of the derived models were

565   evaluated using a PROCHECK (version 3.5)-server to check the stereochemical properties of the

566   modeled structures[48]. Finally, the homology models were structurally aligned using PyMOL[49],

567   and observed for heterogeneity in the conformations. The physiochemical properties of the S

568   glycoprotein of the SARS-CoV-2, SARS-CoV and MERS-CoV was computed using the

569   Protparam ExPasy-server[50].

570

571

572   **Linear B-cell epitopes prediction**

573   We employed both structure and sequence-based methods for B-cell epitopes prediction.

574   Conformational B-cell epitopes on the S protein were predicted by Ellipro (Antibody Epitope

575   Prediction tool; http://tools.iedb.org/ellipro/) available in IEDB analysis resource[51] with the

576   minimum score value set at 0.4 while the maximum distance selected as 6 Å. The ElliPro allows

577   the prediction and visualization of B-cell epitopes in a given protein sequence or structure. The

578   ElliPro method is based on the location of the residue in the protein's three-dimensional (3D)

579   structure. ElliPro implements three algorithms to approximate the protein shape as an ellipsoid,

580   calculate the residue protrusion index (PI), and cluster neighboring residues based on their

581   protrusion index (PI) value. The residues lying outside of the ellipsoid covering 90% of the inner

582   core residues of the protein score highest PI of 0.9[23]. Antigenicity of full-length S (spike

583   glycoprotein), M (membrane protein) and E (envelope protein) proteins was predicted using

584   VaxiJen v2.0 (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html)[52]. The linear B-cell

25

585 epitopes of RBD and NTD regions of S protein, full length E and M proteins were predicted by

586 "BepiPred Linear Epitope Prediction" (propensity scale method such as hidden Markov model)[53]

587 and ABCPred under default parameters[54]. To find out the most probable peptide-based epitopes

588 with better confidence level, selected peptides were further tested using VaxiJen antigenicity

589 scores[51]. The Kolaskar and Tongaonkar antigenicity profiling from IEDB analysis resource was

590 also used for RBD and NTD segments, E and M proteins[55].

591

592 **Screening for T-cell epitopes**

593 Immune Epitope Database (IEDB) tool peptide binding to MHC class I molecules and

594 Proteasomal cleavage/TAP transport/MHC class I combined predictor

595 (http://tools.iedb.org/main/tcell/) were used to screen CTL epitopes, proteasomal cleavage and

596 transporter associated with antigen processing (TAP) with all parameters set to default.

597 Threshold for strong binding peptides (IC_{50}) was set at 50 nM to determine the binding and

598 interaction potentials of helper T-cell epitope peptide and major histocompatibility complex

599 (MHC) class I allele[20,23]. The total score generated by the tool is a combined score of

600 proteasome, major histocompatibility complex (MHC), TAP (N-terminal interaction), processing

601 analysis scores. The HTL epitopes were screened using the IEDB tool "MHC-II Binding

602 Predictions" (http://tools.iedb.org/mhcii/). The tool generated the median percentile rank for each

603 predicted epitope through combination of three methods (Combinatorial Library, SMM-align,

604 and Sturniolo) by comparing the score of peptide against the scores of other random five million

605 15-mer peptides from the SwissProt database[28]. Top five HLA epitopes for each RBD and NTD

606 segments were docked against the respective HLA (MHC-□ and MHC-□) allele binders by

607 interaction similarity-based protein-peptide docking system GalaxyPepDock of the GalaxyWeb,

608   docked HLA-epitope complexes were refined in GalaxyRefineComplex and binding affinity

609   (ΔG) was determined PROtein binDIng enerGY prediction (PRODIGY) tool[56].

610

Population coverage by CTL and HTL epitopes

612   Due to the MHC restriction of T-cell response, the peptides with more different HLA

613   binding specificities mean more population coverage in defined geographical regions and ethnic

614   groups where-and-to-whom the peptide-based vaccine might be employed. The predicted T-cell

615   epitopes were shortlisted based on the aligned Artificial Neural Network (ANN) with half-

616   maximal inhibitory concentration (annIC$_{50}$) values up to 50 nM. The IEDB "Population

617   Coverage" tool (http://tools.iedb.org/population/)[57] was used to determine the world human

618   population coverage by the shortlisted CTL and HTL epitopes. We used multi-epitopes involving

619   both CTL and HTL epitopes to have the higher probability of larger human population coverage

620   worldwide. We used OmicsCircos to visualize the association between world population and

621   different ethnic groups[58].

622

IFN-γ-inducing epitope prediction

624   Potential IFN-γ epitopes of all the selected antigenic sites of RBD, NTD, envelop protein

625   B-cell epitope (EBE), and membrane protein B-cell epitope (MBE) were predicted by

626   "IFNepitope" server (http://crdd.osdd.net/raghava/ ifnepitope/scan.php). To identify the set of

627   epitopes associated with MHC alleles that would maximize the population coverage, we adopted

628   the "Motif and SVM hybrid" (MERCI: Motif-EmeRging and with Classes-Identification, and

629   SVM) approach. The tool generates overlapping sequences from the query antigens for IFN-γ-

630   inducing epitope prediction. The estimated population coverage represents the percentage of

27

631 individuals within the population that are likely to elicit an immune response to at least one T

632 cell epitope from the set, and finally the coverage was observed by adding epitopes associated

633 with any of the remaining MHC alleles[2]. The prediction is based on a dataset of IFN-γ-inducing

634 and IFN-γ-noninducing MHC allele binders[59].

635

Design and construction of multi-epitope vaccine candidate

637 The candidate vaccine design and construction method follows previously developed

638 FMD peptide vaccine development protocol in Professor M Anwar Hossain's laboratory

639 (unpublished data, and under patent application in Bangladesh Application No.

640 P/BD/2018/000280, date: October 1, 2018; India Patent Number: 201924036919, Date:

641 September 13, 2019). The multi-epitope protein was constructed by positioning the selected

642 RBD, NTD, MBE and EBE amino acid sequences linked with short, rigid and flexible linkers

643 GG. To develop highly immunogenic recombinant proteins, two universal T-cell epitopes were

644 used, namely, a pan-human leukocyte antigen DR-binding peptide (PADRE)[60], and an invasin

645 immunostimulatory sequence taken from *Yersinia* (Invasin)[61] were used to the N and C terminal

646 of the vaccine construct respectively, linked with EGGE. Here, acidic aa, Glutamate (E) was

647 added to balance the ratio of acidic and basic amino acids. We denoted the vaccine candidate as

648 CoV-RMEN. VaxiJen 2.0[52] and ANTIGENpro (http://scratch.proteomics.ics.uci.edu/) web-

649 servers were used to predict the antigenicity of the CoV-RMEN while the AllerTOP 2.0

650 (http://www.ddg-pharmfac.net/AllerTOP) and AllergenFP (http://ddg-pharmfac.net/AllergenFP/)

651 web-servers were used to predict vaccine allergenicity.

652

Physiochemical properties and solubility prediction of the CoV-RMEN

28

654   Various physiochemical parameters of the CoV-RMEN were assessed using the online

655   web-server ProtParam[50]. These parameters included aa residue composition, theoretical pI,

656   instability index, *in vitro* and *in vivo* half-life, aliphatic index, molecular weight, and grand

657   average of hydropathicity (GRAVY). The solubility of the multi-epitope vaccine peptide was

658   evaluated using the Protein-Sol server (https://protein-sol.manchester.ac.uk/).

659

660   **Secondary and tertiary structure prediction**

661   Chou   and   Fasman   secondary   structure   prediction   server   (CFSSP:

662   https://www.biogem.org/tool/chou-fasman/)   and   RaptorX   Property

663   (http://raptorx.uchicago.edu/StructurePropertyPred/predict/)[62], web-servers   were   used   for

664   secondary structure predictions. The RaptorX web-server (http://raptorx.uchicago.edu/)[62] was

665   used to predict the three-dimensional (3D) structure and binding residues of the chosen protein.

666   Homology modelling of the CoV-RMEN was carried out using the protein homology/analogy

667   recognition engine (Phyre2) (http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index) web-

668   server.

669

670   **Refinement and validation of the tertiary structure of the CoV-RMEN**

671   The 3D model obtained for the CoV-RMEN was refined in a three-step process, initially

672   energy minimization using the GROMOS96 program[27] implemented within the Swiss-

673   PdbViewer (version 4.1.0)[47]. After energy minimization, the model was refined using

674   ModRefiner (https://zhanglab.ccmb.med.umich.edu/ModRefiner/) and then GalaxyRefine server

675   (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE). The construction and refinement

676   of the CoV-RMEN was further assessed through ModRefiner server for atomic-level energy

29

677 minimization. This results in improvements in both global and local structures, with more

678 accurate side chain positions, better hydrogen-bonding networks, and fewer atomic overlaps[63].

679 The GalaxyRefine server was further used to improving the best local structural quality of the

680 CoV-RMEN according to the CASP10 assessment, and ProSA-web

681 (https://prosa.services.came.sbg.ac.at/prosa.php) was used to calculate overall quality score for a

682 specific input structure, and this is displayed in the context of all known protein structures. The

683 ERRAT server (http://services.mbi.ucla.edu/ERRAT/) was also used to analyze non-bonded

684 atom-atom interactions compared to reliable high-resolution crystallography structures. A

685 Ramachandran plot was obtained through the RAMPAGE server

686 (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php). The server uses the PROCHECK

687 principle to validate a protein structure by using a Ramachandran plot and separates plots for

688 Glycine and Proline residues[64].

689

690 **Immune simulation**

691 To further characterize the immunogenicity and immune response profile of the CoV-

692 RMEN, *in silico* immune simulations were conducted using the C-ImmSim server

693 (http://150.146.2.1/C-IMMSIM/index.php)[28]. All simulation parameters were set at default with

694 time steps set at 1, 84, and 170 (each time step is 8 hours and time step 1 is injection at time = 0).

695 Therefore, three injections were given at four weeks apart. The Simpson index, D (a measure of

696 diversity) was interpreted from the plot.

697

698 **Molecular docking of the CoV-RMEN with TLRs**

30

699   The generation of an appropriate immune response is dependent on the interaction

700   between an antigenic molecule and a specific immune receptor like Toll Like Receptors

701   (TLRs)[33]. Molecular docking of the CoV-RMEN with the TLR3 (PDB ID: 1ZIW) and TLR4

702   (PDB ID: 4G8A) receptor was performed using the High Ambiguity Driven DOCKing

703   (HADDOCK, version 2.4)[65] web-server in order to evaluate the interaction between ligand and

704   receptor and consequently the development of an immune response. In order to predict such

705   interaction, data-driven docking of designed chimeric protein and TLR3 and TLR4 complexes

706   were performed[33]. In this regard, CPORT (https://milou.science.uu.nl/services/CPORT/) was

707   used to predict active interface residues of the CoV-RMEN and TLRs. The HADDOCK server

708   was then employed to perform docking simulations of the CoV-RMEN-TLRs complexes due to

709   perform short MD simulations of docked complexes[56]. Finally, the binding affinities of the top-

710   ranking docking pose of chimeric protein-TLRs complexes was predicted using the PRODIGY

711   (PROtein binDIng enerGY prediction) (https://nestor.science.uu.nl/prodigy/) web-server[56].

712

713   **Analysis of cDNA and mRNA for cloning and expression**

714   Reverse translation and codon optimization were performed using the GenScript Rare

715   Codon Analysis Tool (https://www.genscript.com/tools/rare-codon-analysis) in order to express

716   the CoV-RMEN in a selected expression vector. Codon optimization was performed in order to

717   express the final vaccine construct in the *E. coli* (strain K12). The analysis output includes the

718   codon adaptation index (CAI) and percentage GC content, which can be used to assess protein

719   expression levels. CAI provides information on codon usage biases; the ideal CAI score is 1.0

720   but >0.8 is considered a good score[66]. The GC content of a sequence should range between 30–

721   70%. GC content values outside this range suggest unfavorable effects on translational and

31

722   transcriptional efficiencies[22]. Stability of the mRNA was measured using two different tools

723   namely RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) and the mfold

724   (http://unafold.rna.albany.edu/?q=mfold) web-servers. To clone the optimized gene sequence of

725   the final vaccine construct in *E. coli* into pETite vector (Lucigen, USA) through enzyme-free

726   method. The primers for amplification of the DNA fragment must contain 18-nucletide short

727   DNA sequence (5'-CGCGAACA-GATTGGAGGT-3' in upstream of forward primer and 5'-

728   GTGGCGGCCGCTCTATTA-3' in upstream of reverse primer) to facilitate the enzyme-free

729   cloning of the gene at the insertion site. Finally, the sequence of the recombinant plasmid was

730   designed by inserting the adapted codon sequences into pETite vector using SnapGene software

731   (from Insightful Science; available at snapgene.com).

732

733   **Conflict of interest statement**

734   The authors declare no competing interests.

735

736   **Author contributions**

737   MSR, MNH, and MRI done the overall study and also draft the manuscript. SA drafted

738   some parts of results and discussion. MNH finally compiled the manuscript. OS, and MAS

739   helped in final discussion and reference indexing. ASMRUA, MMR, MS and MAH contributed

740   intellectually to the interpretation and presentation of the results. MS and MAH developed the

741   concept of peptide vaccine development. Finally, all authors have approved the manuscript for

742   submission.

743

744   **Acknowledgements**

32

749

750   **Supplementary Material**

751   Supplementary information supporting the findings of the study are available in this

752   article as Supplementary Data files, or from the corresponding author on request.

753

754

755

756   **References**

757   1. WHO, 2020. Novel Coronavirus (2019-nCoV) situation reports - World Health

758   Organization (WHO)

759   2. Almofti, Y. A., Abd-elrahman, K. A., Gassmallah, S. A. E. & Salih, M. A.   Multi

760   Epitopes Vaccine Prediction against Severe Acute Respiratory Syndrome (SARS)

761   Coronavirus Using Immunoinformatics Approaches. *American J. Microbiol. Res.* **6(3)**,

762   94-114 (2018).

763   3. Wu, F. et al. "A new coronavirus associated with human respiratory disease in China."

764   *Nat*. 1-5 (2020)

765   4. Badawi, M. M. et al. *In Silico* Prediction of a Novel Universal Multi-epitope Peptide

766   Vaccine in the Whole Spike Glycoprotein of MERS CoV. *Am. J. Microbiol Res.* **4(4)**,

767   101-21 (2016).

768   5. Pallesen, J. et al. Immunogenicity and structures of a rationally designed prefusion
769   MERS-CoV spike antigen. *Proc. Nat. Aca. Sci*. **114(35)**, E7348-E7357 (2017).

770   6. ul Qamar, M. T., Saleem, S., Ashfaq, U. A., Bari, A., Anwar, F. & Alqahtani, S.
771   Epitope☐based peptide vaccine design and target site depiction against Middle East
772   Respiratory Syndrome Coronavirus: an immune-informatics study. *J. Translational Med*.
773   **17(1)**, 362 (2019).

774   7. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential
775   vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV
776   immunological studies. *Viruses*. **12(3)**, 254 (2020).

777   8. Abdelmageed, M. I. et al. Design of multi epitope-based peptide vaccine against E
778   protein of human 2019-nCoV: An immunoinformatics approach. *BioRxiv*. (2020).

779   9. Song, W., Gui, M., Wang, X. & Xiang, Y. Cryo-EM structure of the SARS coronavirus
780   spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog.* **14(8)**,
781   e1007236 (2018).

782   10. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion
783   conformation. *Science* 2020.

784   11. Shang, W. et al. The outbreak of SARS-CoV-2 pneumonia calls for viral vaccines. *npj*
785   *Vaccines* **5**, 18 (.2020).

786   12. Gralinski, L. E. & Menachery, V. D. Return of the coronavirus: 2019-nCoV. *Viruses* **12**,
787   135 (2020).

788   13. Zhou, H. et al. Structural definition of a neutralization epitope on the N-terminal domain
789   of MERS-CoV spike glycoprotein. *Nat. Commun.* **10**, 3068 (2019).

14. Rabi, F. A., Al Zoubi, M. S., Kasasbeh, G. A., Salameh, D. M. & Al-Nasser, A. D. SARS-CoV-2 and Coronavirus Disease 2019: What We Know So Far. *Pathogens* **9(3)**, 231 (2020).

15. Wang, N. et al. Structural Definition of a Neutralization-sensitive Epitope on the MERS-CoV S1-NTD. *Cell Rep.* **28(13)**, 3395-405 (2019).

16. Shi, S. Q. et al. The expression of membrane protein augments the specific responses induced by SARS-CoV nucleocapsid DNA immunization. *Mol. Immunol.* **43**, 1791–1798 (2006).

17. Schoeman, D. & Fielding, B. C. Coronavirus envelope protein: current knowledge. *Virol J.* **16**, 69 (2019).

18. Chan, J. F. et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* **9**, 221–236 (2020).

19. Li, G. et al. Coronavirus infections and immune responses. *J. Med. Virology* **92(4)**, 424-432 (2020).

20. Shi, J. et al. Epitope-Based Vaccine Target Screening against Highly Pathogenic MERS-CoV: An *In Silico* Approach Applied to Emerging Infectious Diseases. *PLoS One* **10(12)**, e0144475 (2015).

21. Shey, R. A. et al. *In-silico* design of a multi-epitope vaccine candidate against onchocerciasis and related filarial diseases. *Sci Rep.* **9(1)**, 1-18 (2019).

22. Yong, et al. Recent Advances in the Vaccine Development Against Middle East Respiratory Syndrome-Coronavirus. *Front. Microbiol.* **10**, 1781 (2019).

812   23. Srivastava, S. et al. Design of novel multi-epitope vaccines against severe acute

813   respiratory syndrome validated through multistage molecular interaction and dynamics. *J.*

814   *Biomol. Struc. Dyn.* **37(16)**, 4345-4360 (2019).

815   24. Sakib, M. S., Islam, M., Hasan, A. K. M. & Nabi, A. H. M. Prediction of epitope-based

816   peptides for the utility of vaccine development from fusion and glycoprotein of nipah

817   virus using *in silico* approach. *Adv. Bioinform.* (2014).

818   25. Adhikari, U. K., Tayebi, M. & Rahman, M. M. Immunoinformatics approach for

819   epitope-based peptide vaccine design and active site prediction against polyprotein of

820   emerging oropouche virus. *J. Immunol. Res.* (2018).

821   26. Hebditch, M., M. A. Carballo-Amador, S. Charonis, R. Curtis & J. Warwicker. Protein–

822   Sol: a web tool for predicting protein solubility from sequence. *Bioinform.* **33(19)**, 3098-

823   3100 (2017).

824   27. Van Gunsteren, W., S. et al. The GROMOS96 manual and user guide. (1996).

825   28. Rapin, N.,Lund, O, Bernaschi, M & Castiglione, M. Computational immunology meets

826   bioinformatics: the use of prediction tools for molecular binding in the simulation of the

827   immune system. *PLoS One* **5(4)** (2010).

828   29. Biswal, J. K., Bisht, P., Mohapatra, J. K., Ranjan, R., Sanyal, A. & Pattnaik, B.

829   "Application of a recombinant capsid polyprotein (P1) expressed in a prokaryotic system

830   to detect antibodies against foot-and-mouth disease virus serotype O." *J. Virol. Methods*

831   **215,** 45-51 (2015).

832   30. Tai, W. et al. Characterization of the receptor-binding domain (RBD) of 2019 novel

833   coronavirus: implication for development of RBD protein as a viral attachment inhibitor

834   and vaccine. *Cell. Mol. Immunol.* (2020).

835 31. Zhou, P. et al. Discovery of a novel coronavirus associated with the recent pneumonia
836   outbreak in humans and its potential bat origin. *bioRxiv*. (2020).

837 32. Hui, D. S. et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to
838   global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Inft.*
839   *Dis.* **91**, 264-266 (2020).

840 33. Pei, H. et al. Expression of SARS-coronavirus nucleocapsid protein in *Escherichia coli*
841   and *Lactococcus lactis* for serodiagnosis and mucosal vaccination. *Applied Microbiol.*
842   *Biotech.* **68(2)**, 220-227 (2005).

843 34. Pang, J. et al. Potential rapid diagnostics, vaccine and therapeutics for 2019 novel
844   Coronavirus (2019-ncoV): a systematic review. *J. Clinical Med.* **9(3)**, 623 (2020).

845 35. Frieman, M. et al. "SARS coronavirus and innate immunity." *Virus Research. 133(1),*
846   101-112 (2018).

847 36. Millet, J. K. & Whittaker, G. R. "Host cell proteases: Critical determinants of coronavirus
848   tropism and pathogenesis." *Virus Res*. **202,** 120-134 (2015).

849 37. Jaimes, J. A., Andre, N. M., Millet, J. K. & Whittaker, G. R. Structural modeling of
850   2019-novel coronavirus (nCoV) spike protein reveals a proteolytically-sensitive
851   activation loop as a distinguishing feature compared to SARS-CoV and related SARS-
852   like coronaviruses. *arXiv*. (2020).

853 38. Huang, H. et al. CD4(+) Th1 cells promote CD8(+) Tc1 cell survival, memory response,
854   tumor localization and therapy by targeted delivery of interleukin 2 via acquired pMHC I
855   complexes. *Immunology* **120**, 148–159 (2007).

856   39. Mohammad, N. et al. *In silico* design of a chimeric protein containing antigenic

857   fragments of Helicobacter pylori; a bioinformatic approach. *The Open Microbiol. J.* 10,

858   97 (2016).

859   40. Meza, B., Ascencio, F., Sierra-Beltrán, A. P., Torres, J. & Angulo, C. A novel design of a

860   multi-antigenic, multistage and multi-epitope vaccine against *Helicobacter pylori*: an *in*

861   *silico* approach. *Infect. Genet. Evol.* **49**, 309-317 (2017).

862   41. Totura, A. L. et al. Toll-like receptor 3 signaling via TRIF contributes to a protective

863   innate immune response to severe acute respiratory syndrome coronavirus

864   infection. *MBio 6*(3), e00638-15 (2015).

865   42. Mosaddeghi, P. et al. Therapeutic approaches for COVID-19 based on the dynamics of

866   interferon-mediated immune responses. (2020).

867   43. Zander, R. A. et al. Th1-like plasmodium-specific memory CD4+ T cells support

868   humoral immunity. *Cell Rep.* **21(7)**, 1839-1852 (2017).

869   44. Carvalho, L. H. et al. IL-4-secreting CD4+ T cells are crucial to the development of

870   CD8+ T-cell responses against malaria liver stages. *Nat. Med.* **8**, 166–70 (2002).

871   45. Hoque, M. N. et al. Metagenomic deep sequencing reveals association of microbiome

872   signature with functional biases in bovine mastitis. *Sci. Rep.* **9**, 13536 (2019).

873   46. Waterhouse, A. et al.. "SWISS-MODEL: homology modelling of protein structures and

874   complexes." *Nucleic Acids Res.* **46(W1)**, W296-W303 (2018).

875   47. Guex, N. & Peitsch, M. C. "SWISS□MODEL and the Swiss□Pdb Viewer: an

876   environment for comparative protein modeling." *Electrophoresis* **18(15)**, 2714-2723

877   (1997).

48. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. "PROCHECK: a program to check the stereochemical quality of protein structures." *J. Applied Crystal* **26(2)**, 283-291 (1993).

49. Faure, G. et al. iPBAvizu: a PyMOL plugin for an efficient 3D protein structure superimposition approach. *Source Code Biol. Med.* 14, 5 (2019).

50. Gasteiger, E. et al. Protein identification and analysis tools on the ExPASy server. *The Proteomics Protocols Handbook*, Springer**:** 571-607 (2005).

51. Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable B cell epitope predictions: Impacts of method development and improved benchmarking. *PLoS Comput. Biol.* 8(12), e1002829 (2012).

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

53. Larsen, J. E. P., Lund, O. & Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2(1)**, 2 (2006).

54. Saha, S. & Raghava G. P. S. "Prediction of continuous B☐cell epitopes in an antigen using recurrent neural network." *Proteins: Structure, Function, Bioinformatics* **65(1)**, 40-48 (2006).

55. Kolaskar, A. & Tongaonkar, P. C. "A semi☐empirical method for prediction of antigenic determinants on protein antigens." *FEBS Letters* **276(1-2)**, 172-174 (1990).

56. Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M. & Vangone, A. "PRODIGY: a web server for predicting the binding affinity of protein–protein complexes." *Bioinformatics* **32(23)**, 3676-3678 (2016).

900   57. Bui, H. H., Sidney, J., Dinh, K., Southwood, S., Newman, M. J & Sette, A. Predicting

901   population coverage of T-cell epitope-based diagnostics and vaccines. *BMC*

902   *Bioinformatics* **7(1)**,153 (2006).

903   58. Hoque, M. N. et al. Resistome diversity in bovine clinical mastitis microbiome, a

904   signature concurrence. *bioRxiv*, 829283 (2019).

905   59. Dhanda, S. K., Vir, P. & Raghava, G. P. "Designing of interferon-gamma inducing MHC

906   class-II binders." *Biology Direct* **8(1)**, 30 (2013).

907   60. Agadjanyan, M. G. et al. "Prototype Alzheimer's disease vaccine using the

908   immunodominant B cell epitope from β-amyloid and promiscuous T cell epitope pan

909   HLA DR-binding peptide." *The J. Immunology* **174(3)**, 1580-1586 (2005).

910   61. Li, H. et al. Co-expression of the C-terminal domain of *Yersinia enterocolitica* invasin

911   enhances the efficacy of classical swine-fever-vectored vaccine based on human

912   adenovirus. *J. Biosci.* **40**, 79–90 (2015).

913   62. Källberg, M., Margaryan, G., Wang, S., Ma, J. & Xu, J. RaptorX server: a resource for

914   template-based protein structure modeling. In *Protein Structure Prediction* **(pp. 17-27)**,

915   Humana Press, New York, NY.2014.

916   63. Xu, D. & Zhang, Y. "Improving the physical realism and structural accuracy of protein

917   models by a two-step atomic-level energy minimization." *Biophysical J.* **101(10)**, 2525-

918   2534 (2011).

919   64. Lovell, S. C. et al. "Structure validation by Cα geometry: □, ψ and Cβ deviation."

920   *Proteins: Structure, Function, Bioinformatics* **50(3)**, 437-450 (2003).

921   65. De Vries, S. J., Van Dijk, M. & Bonvin, A. M. The HADDOCK web server for data-

922   driven biomolecular docking. *Nat. Protoc*. **5(5)**, 883 (2010).

40

923   66. Morla, S., Makhija, A. & Kumar, S. "Synonymous codon usage pattern in glycoprotein

924   gene of rabies virus." *Gene* **584(1)**, 1-6 (2016).

925

926

Spike Protein S

N-terminal

NTD

RBD

C-terminal

Chain A
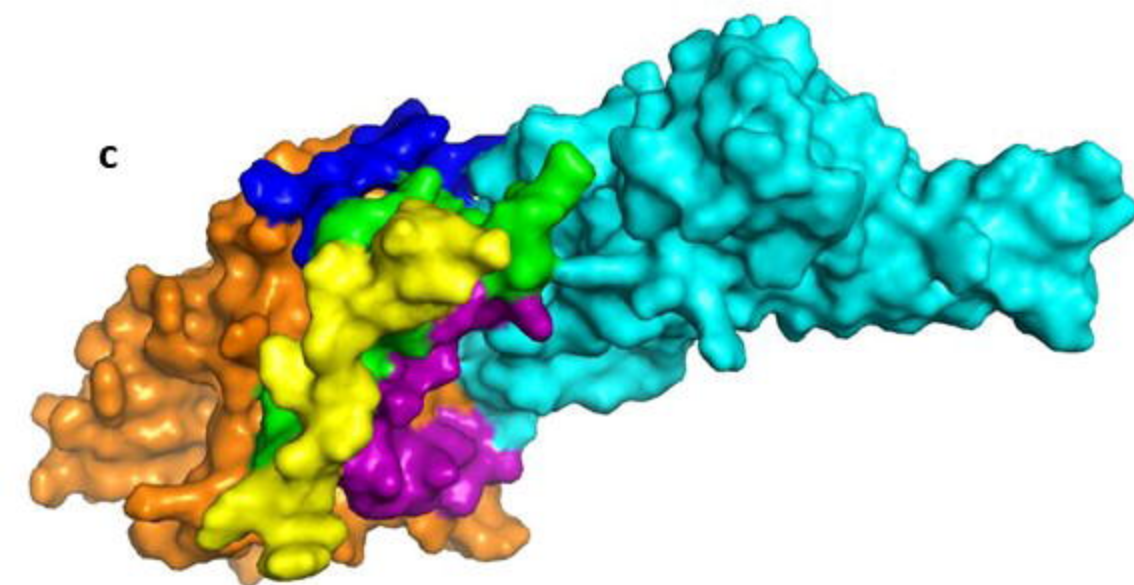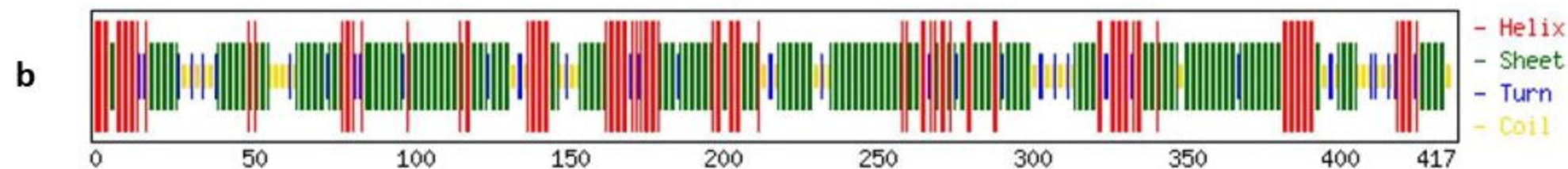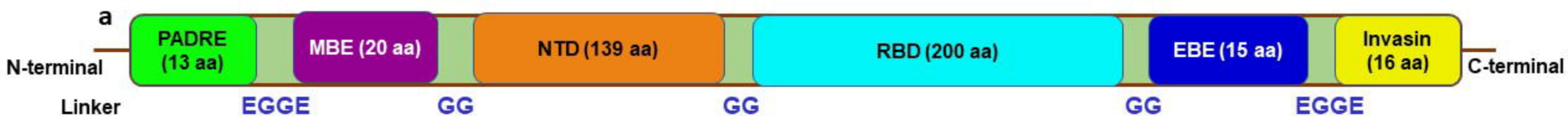
Chain B

Chain C

SARS-CoV-2   SARS-CoV   MERS-CoV
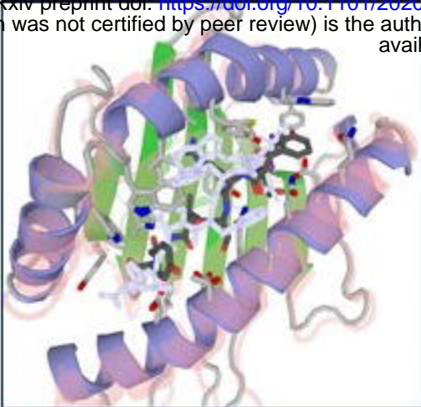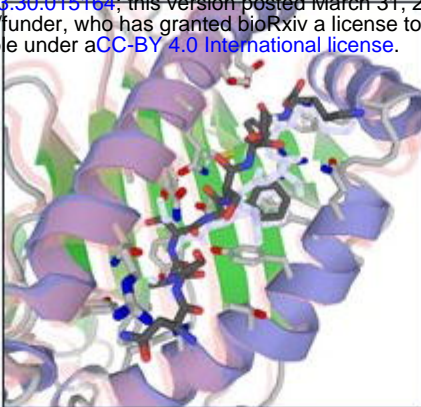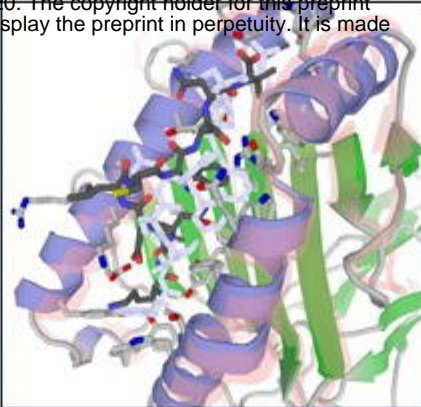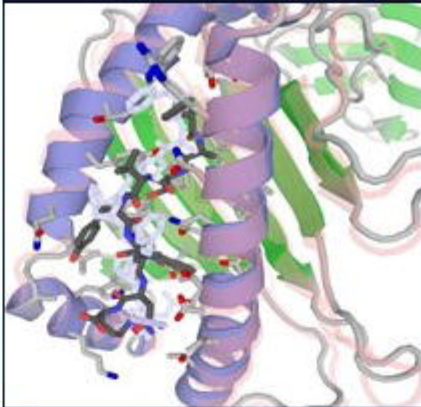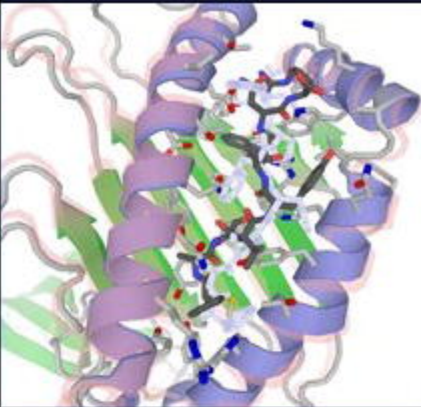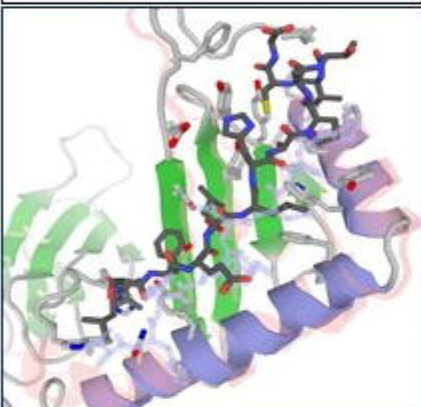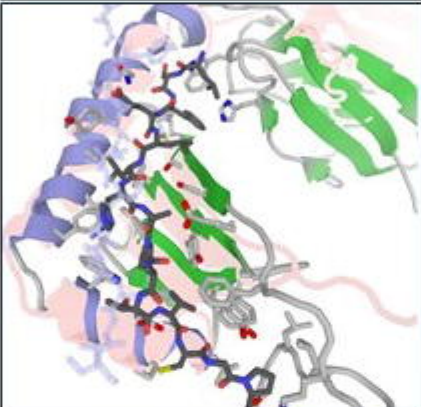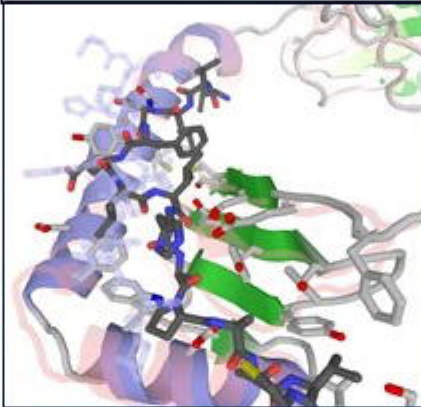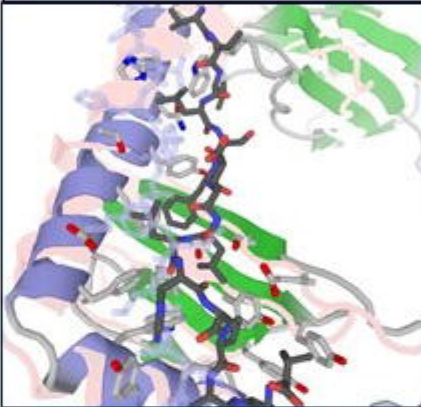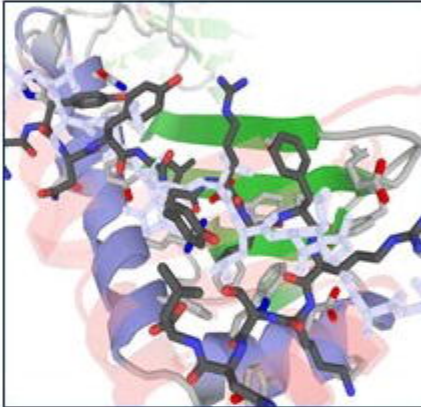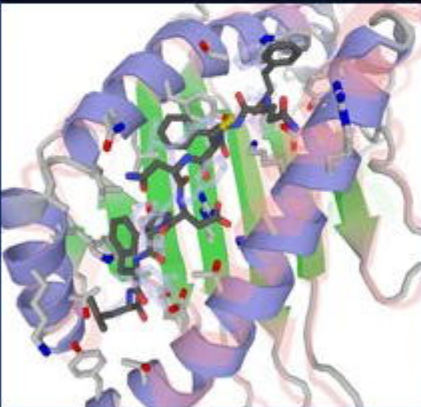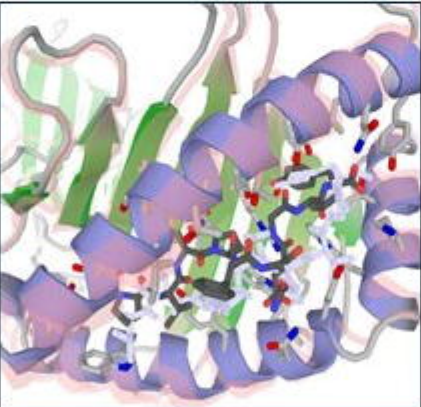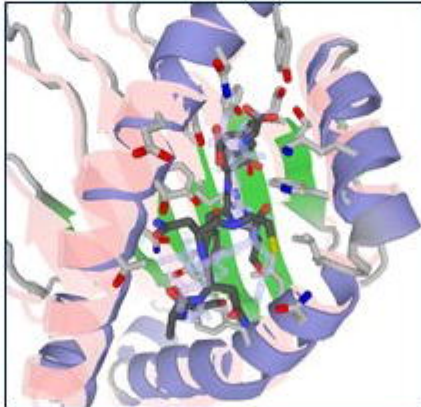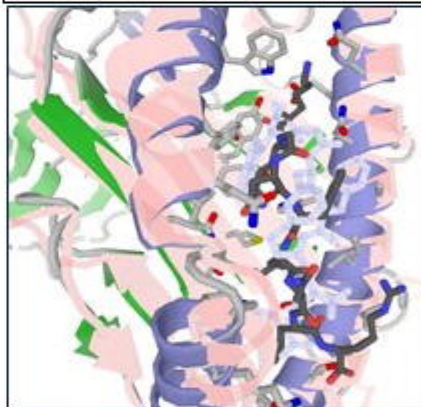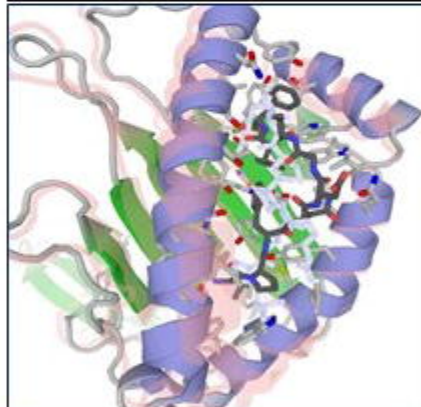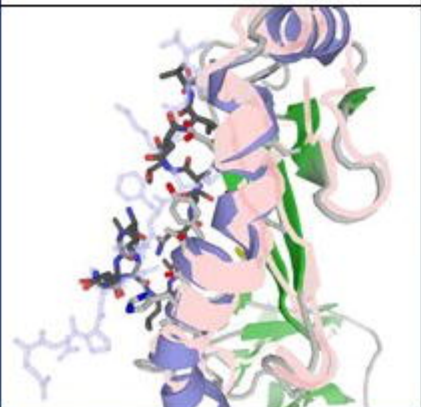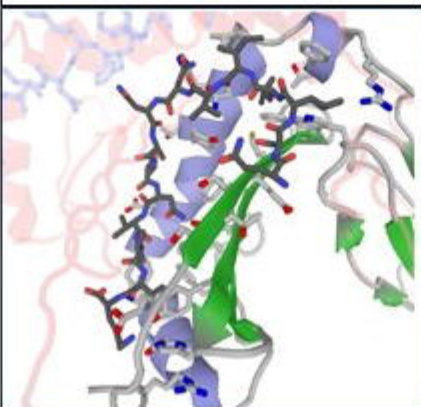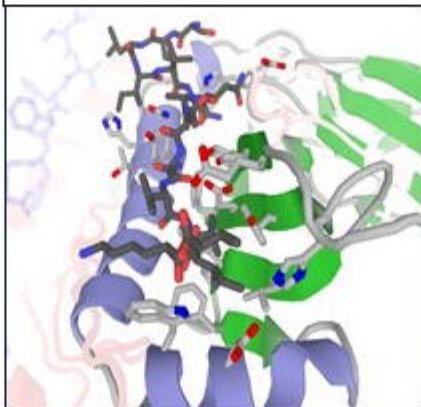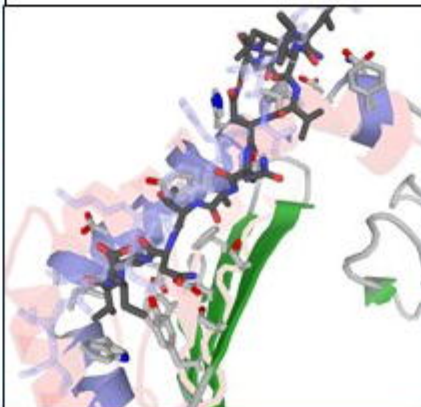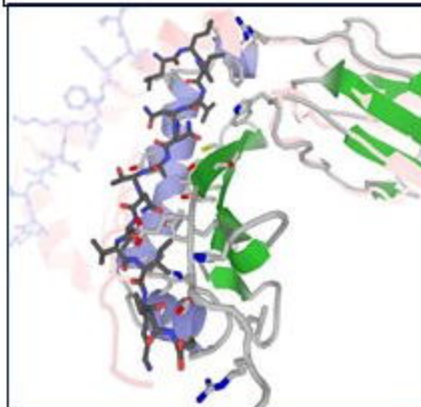
	1	2	3	4	5
MHC-I of RBD	ATRFASVYA-A*30:01 ΔG/-10.1(kcal mol-1)	NSASFSTFK -A*68:01 ΔG/-8.4(kcal mol-1)	KLNDLCFTNV -A*02:03 ΔG/-10.8(kcal mol-1)	FASVYAWNR -A*68:01 ΔG/-11.1(kcal mol-1)	RLFRKSNLK -A*33:01 ΔG/-11.9(kcal mol-1)
MHC-II of RBD	VLSFELLHAPATVCG - DRB1*01:01 ΔG/-8.6(kcal mol-1)	LSFELLHAPATVCGP - DRB1*01:01 ΔG/-11.8(kcal mol-1)	VVLSFELLHAPATVC - DRB1*01:01 ΔG/-11.4(kcal mol-1)	VVVLSFELLHAPATV - DRB1*01:01 ΔG/-9.8(kcal mol-1)	GNYNYLYRLFRKSNL - DRB1*11:01 ΔG/-9.8(kcal mol-1)
MHC-I of NTD	FQFCNDPFL -A*02:06 ΔG/-9.2(kcal mol-1)	LPFFSNVTW -B*15:01 ΔG/-11.0(kcal mol-1)	SANNCTFEY -B*35:01 ΔG/-8.4(kcal mol-1)	KQGNFKNLR -A*31:01 ΔG/-10.7(kcal mol-1)	LPFNDGVYF -B*35:01 ΔG/-10.3(kcal mol-1)
MHC-II of NTD	QSLLIVNNATNVVIK - DRB1*13:02 ΔG/-8.2(kcal mol-1)	SLLIVNNATNVVIKV - DRB1*13:02 ΔG/-9.6(kcal mol-1)	TQSLLIVNNATNVVI - DRB1*13:02 ΔG/-10.4(kcal mol-1)	KTQSLLIVNNATNVV - DRB1*13:02 ΔG/-8.4(kcal mol-1)	LLIVNNATNVVIKVC - DRB1*13:02 ΔG/-8.8(kcal mol-1)

a

b

a

PADRE (13 aa) | MBE (20 aa) | NTD (139 aa) | RBD (200 aa) | EBE (15 aa) | Invasin (16 aa)

N-terminal   C-terminal

Linker   EGGE   GG   GG   GG   EGGE

b

- Helix
- Sheet
- Turn
- Coil

0   50   100   150   200   250   300   350   400   417

c

d

PHE 178

VAL 171

ALA 107

ASN 71

HIS 29

Psi (degrees): 180, 135, 90, 45, 0, -45, -90, -135

Phi (degrees): -180, -135, -90, -45, 0, 45, 90, 135, 180

a   b

e

X-ray
NMR

Z-score: 10, 5, 0, -5, -10, -15, -20
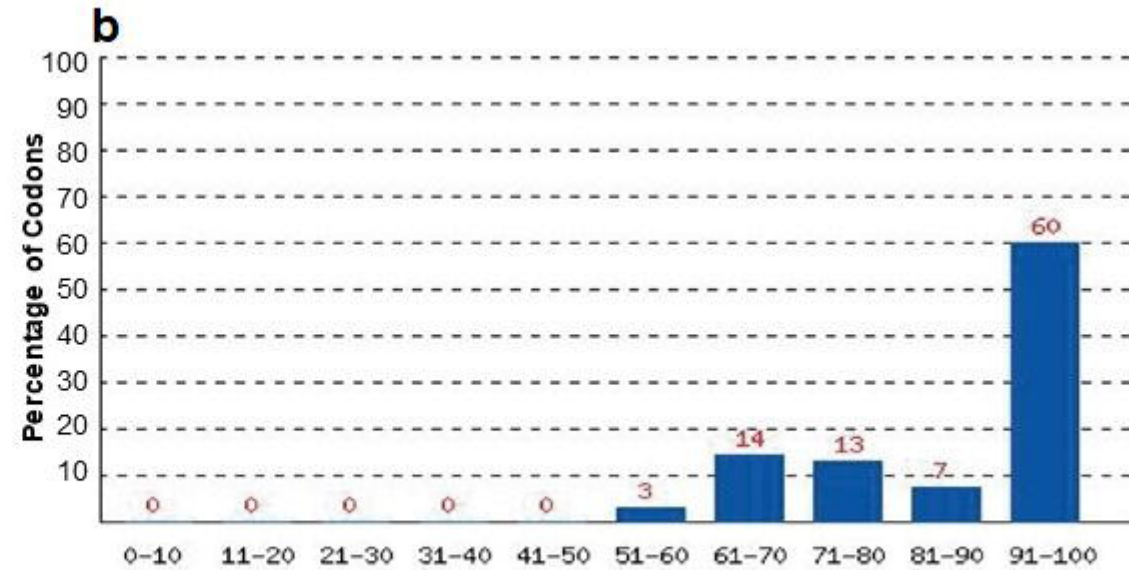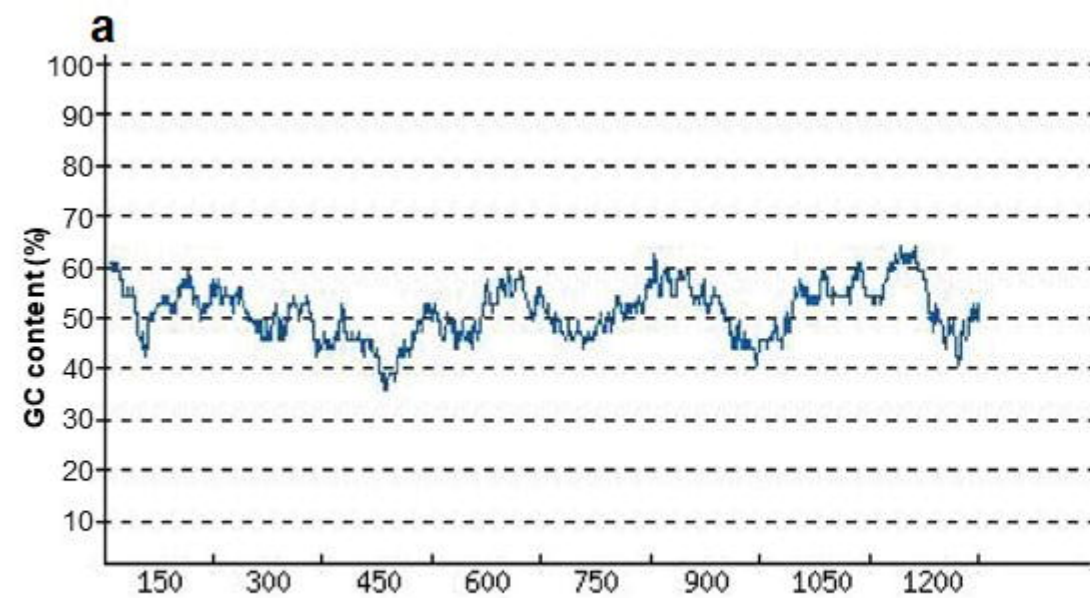
Number of residues: 0, 200, 400, 600, 800, 1000
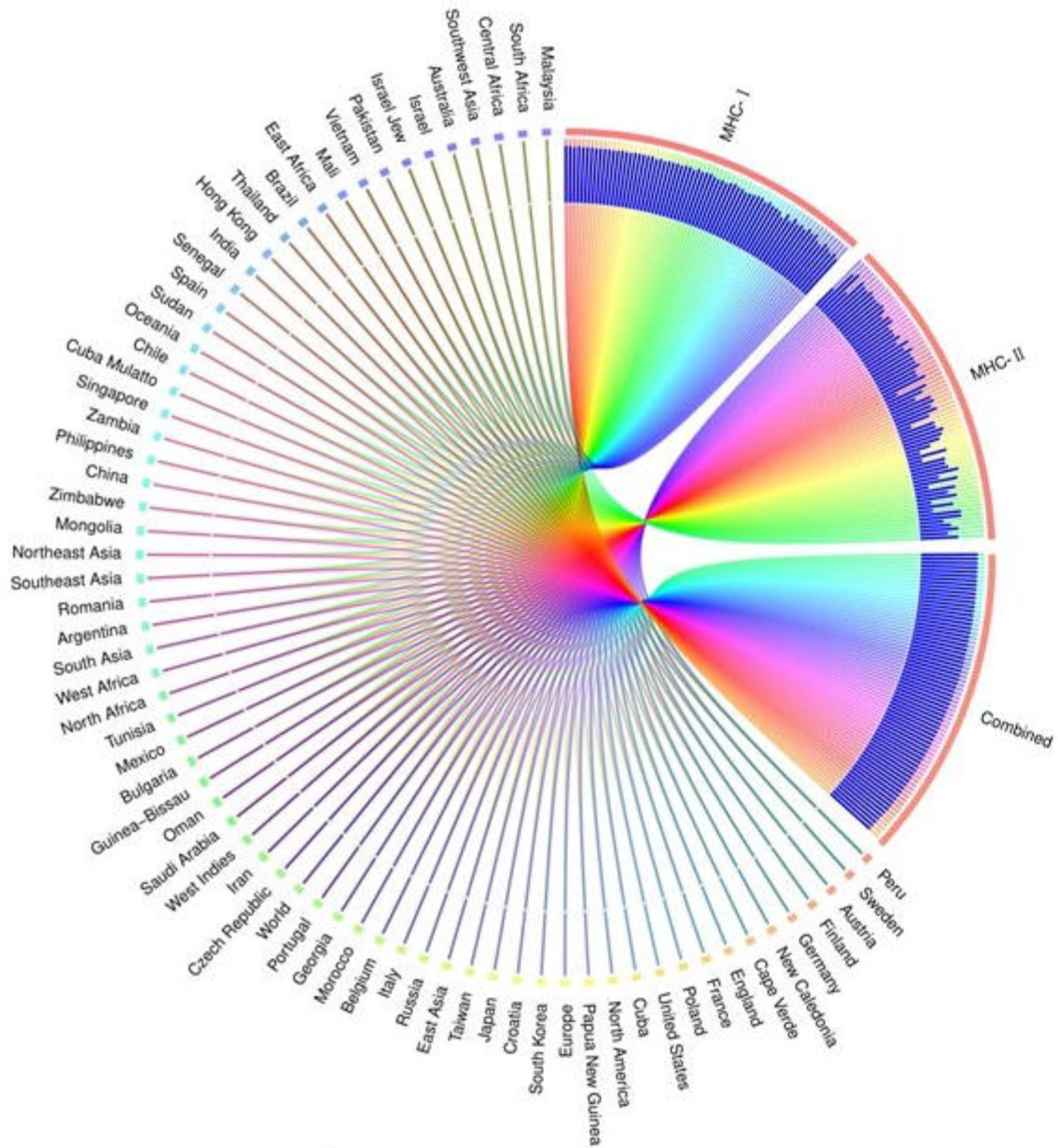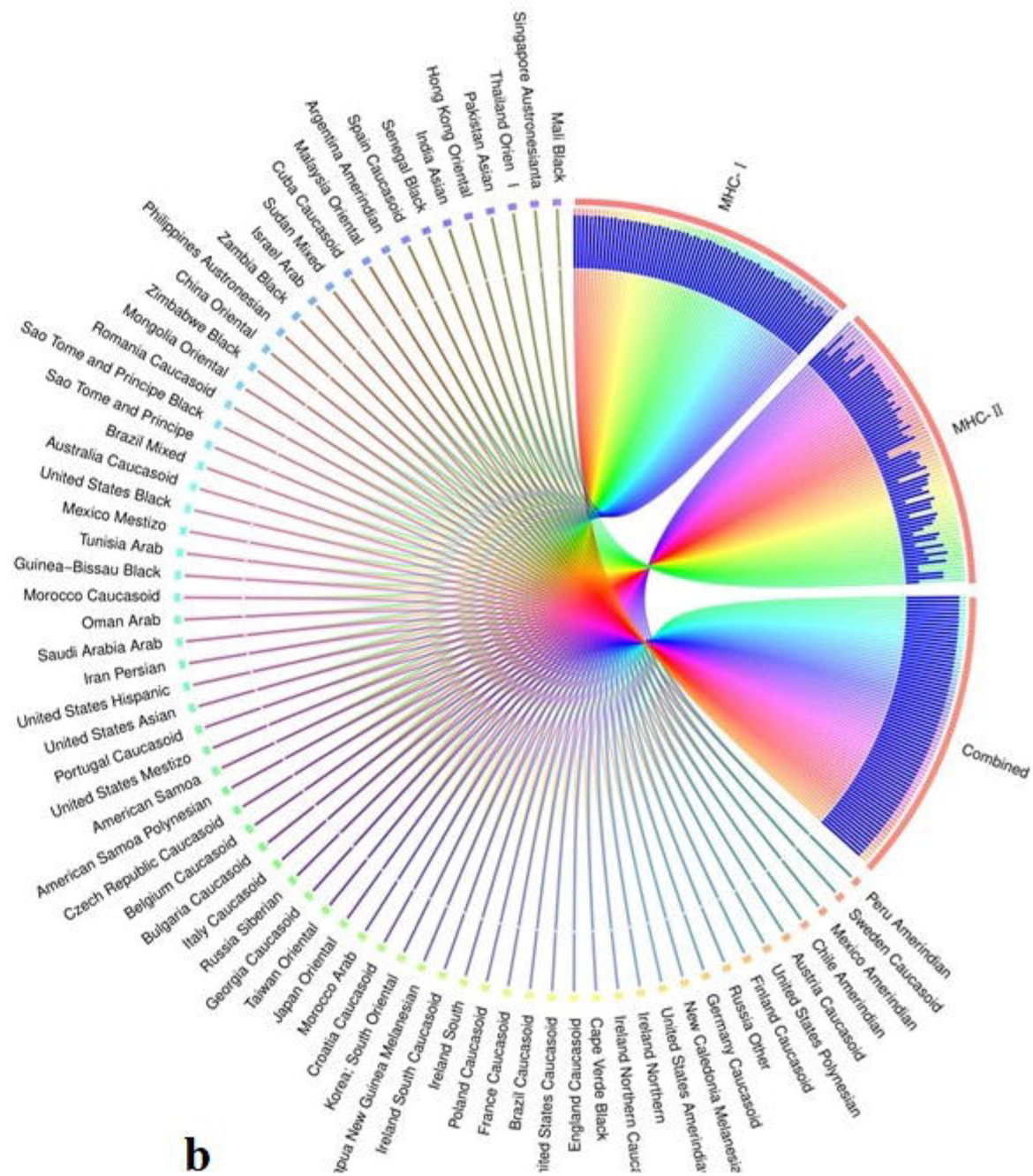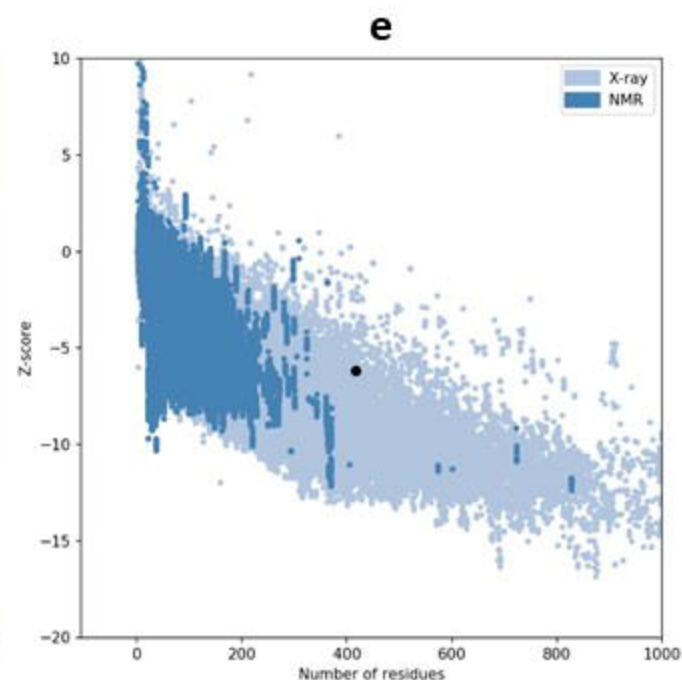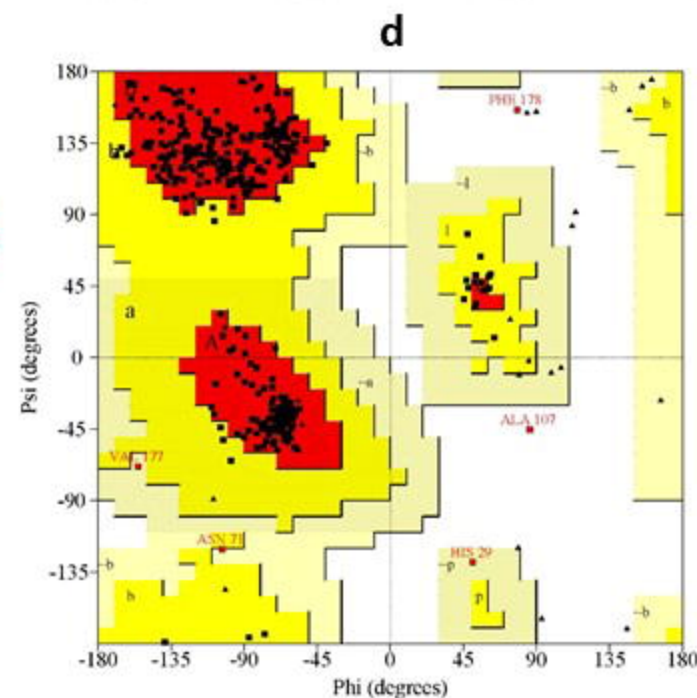
f
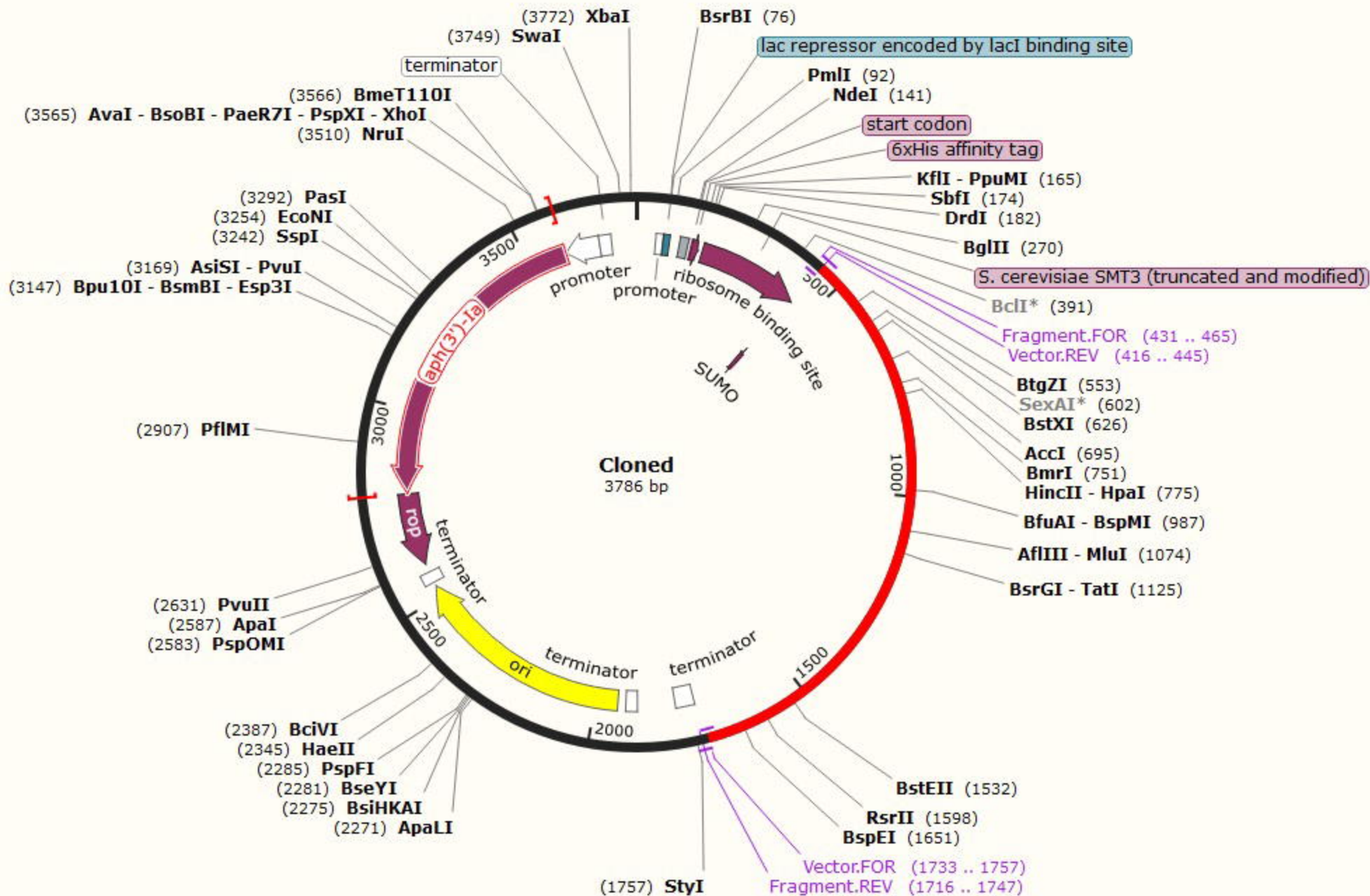
>CoV-RMEN

AKFVAAWTLKAAA EGGE YRIGNYKLNTDHSSSSDNIA GG LPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKV CEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVF GG RFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSA SFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRFKSNLKPFERDISTEIYQAGSTPCNGVEG FNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGP GG YVVYSRVKNLNSSRVP EGGE TAKSKKFPSYTATYQF

a — TLR2 — $\Delta G = -9.0\ \text{kcal mol}^{-1}$

b — TLR3 — $\Delta G = -14.9\ \text{kcal mol}^{-1}$

c — TLR4 — $\Delta G = -16.0\ \text{kcal mol}^{-1}$

d

e

f

Cloned
3786 bp

(3772) **XbaI**
(3749) **SwaI**
terminator
BsrBI (76)
lac repressor encoded by lacI binding site
PmlI (92)
NdeI (141)
start codon
6xHis affinity tag
(3566) **BmeT110I**
(3565) **AvaI - BsoBI - PaeR7I - PspXI - XhoI**
(3510) **NruI**
KflI - PpuMI (165)
SbfI (174)
DrdI (182)
BglII (270)
(3292) **PasI**
(3254) **EcoNI**
(3242) **SspI**
S. cerevisiae SMT3 (truncated and modified)
BclI* (391)
Fragment.FOR (431 .. 465)
Vector.REV (416 .. 445)
(3169) **AsiSI - PvuI**
(3147) **Bpu10I - BsmBI - Esp3I**
BtgZI (553)
SexAI* (602)
BstXI (626)
(2907) **PflMI**
AccI (695)
BmrI (751)
HincII - HpaI (775)
BfuAI - BspMI (987)
AflIII - MluI (1074)
BsrGI - TatI (1125)
(2631) **PvuII**
(2587) **ApaI**
(2583) **PspOMI**
(2387) **BciVI**
(2345) **HaeII**
(2285) **PspFI**
(2281) **BseYI**
(2275) **BsiHKAI**
(2271) **ApaLI**
BstEII (1532)
RsrII (1598)
BspEI (1651)
Vector.FOR (1733 .. 1757)
Fragment.REV (1716 .. 1747)
(1757) **StyI**

promoter
promoter
ribosome binding site
SUMO
aph(3')-Ia
rop
terminator
ori
terminator
terminator
3500
3000
2500
2000
1500
1000
500

Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an *in silico* approach

M. Shaminur Rahman, M. Nazmul Hoque, M. Rafiul Islam, Salma Akter, A.S.M. Rubayet- Ul-Alam, Mohammad Anwar Siddique, Otun Saha, Md. Mizanur Rahaman, Munawar Sultana, M. Anwar Hossain
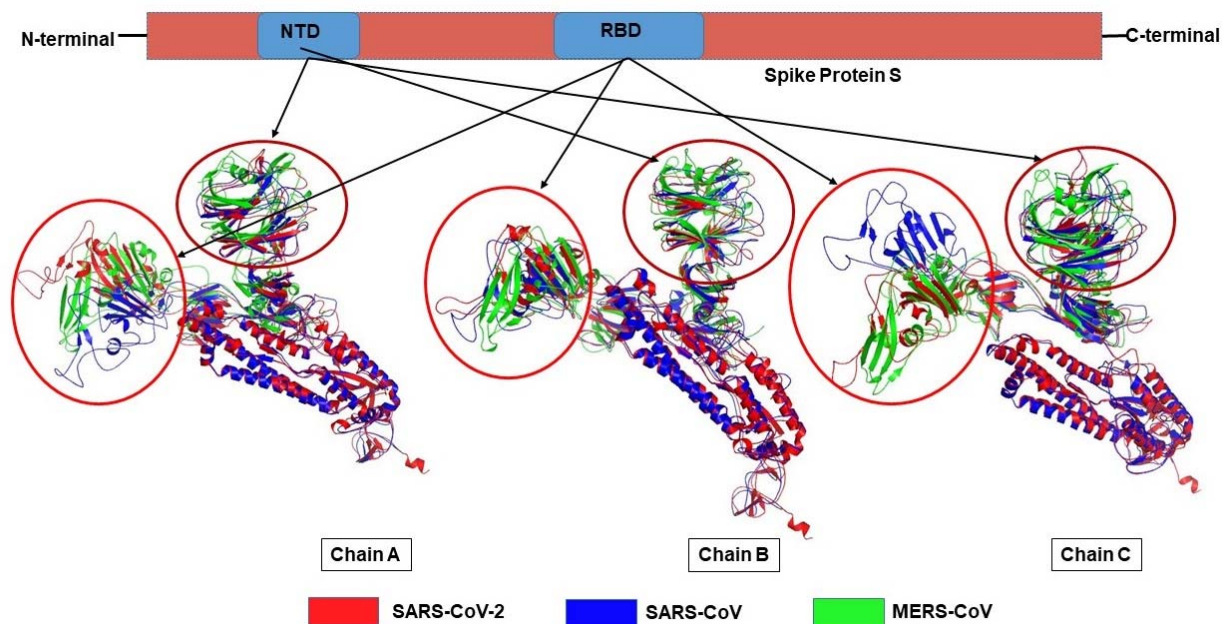
Fig. 1 The trimeric conformation spike (S) proteins. The S proteins of the SARS-CoV-2, SARS-CoV and MERS-CoV is a trimeric conformation consisting of three homologous chains named as chain A, B and C. Respective sequences of these three chains were aligned and visualized using PyMOL which revealed high degree of structural divergences in the N-terminal domains (NTDs) and receptor binding domains (RBDs) of the chain A and C compared to that of chain B.
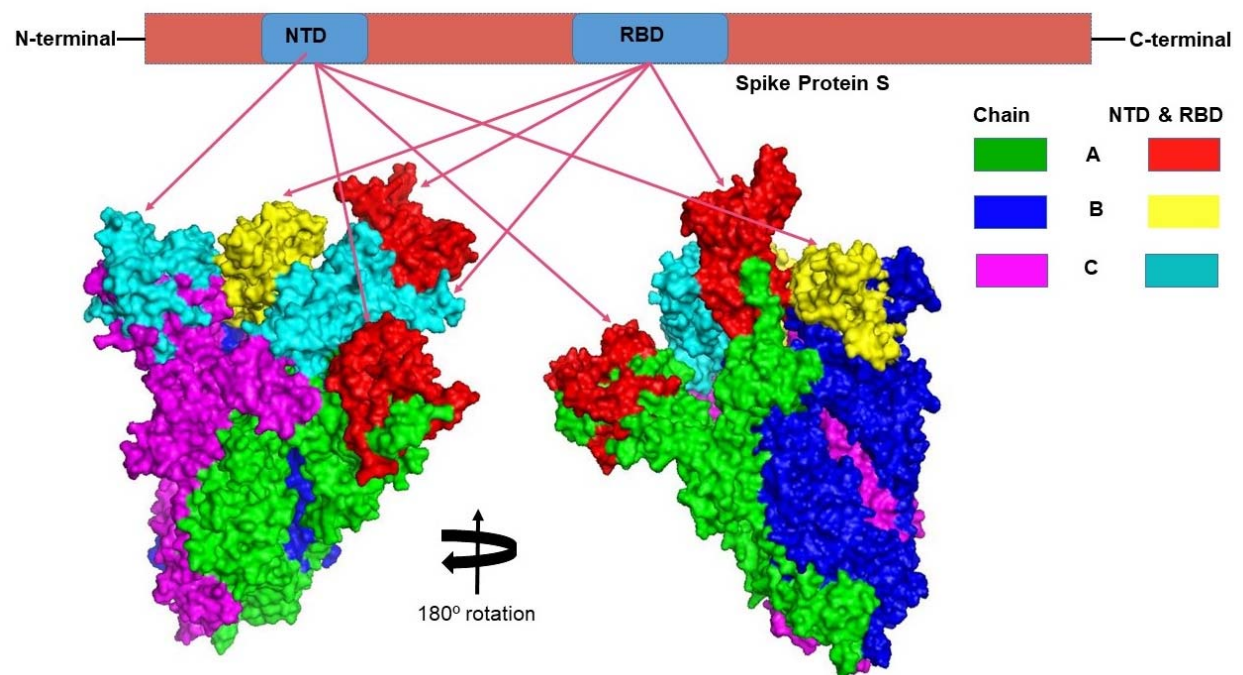
Fig. 2 The three-dimensional (3D) structure of the N-terminal domains (NTDs) and receptor binding domains (RBDs) of the spike (S) proteins of SARS-CoV-2 (surface view). The red, cyan, and yellow colored regions represent the potential antigenic domains predicted by the IEDB analysis resource Elipro analysis.
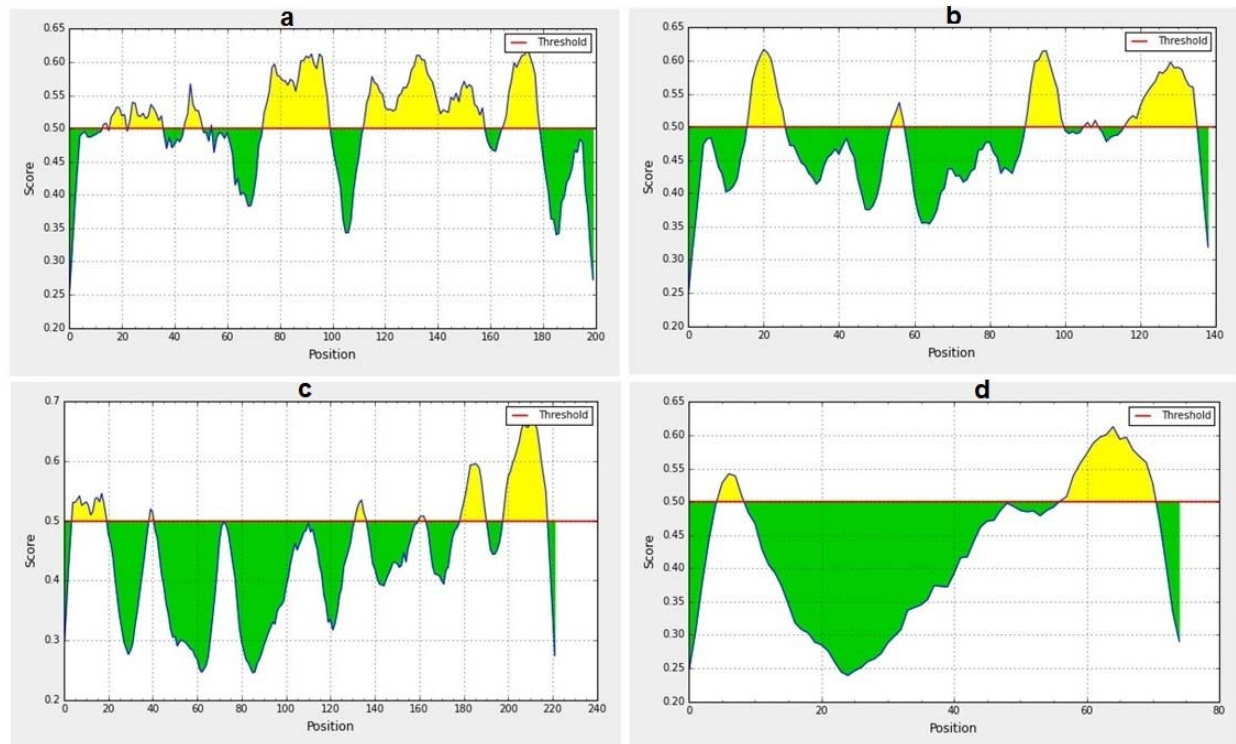
Fig. 3 Predicted B-cell epitopes using BepiPred-2.0 epitope predictor in IEDB-analysis resource web-based repository. Yellow areas above threshold (red line) are proposed to be a part of B cell epitopes in (a) RBD and (b) NTD regions of S protein, (c) envelop (E) and (d) membrane (M) proteins of SARS-CoV-2. While green areas are not.
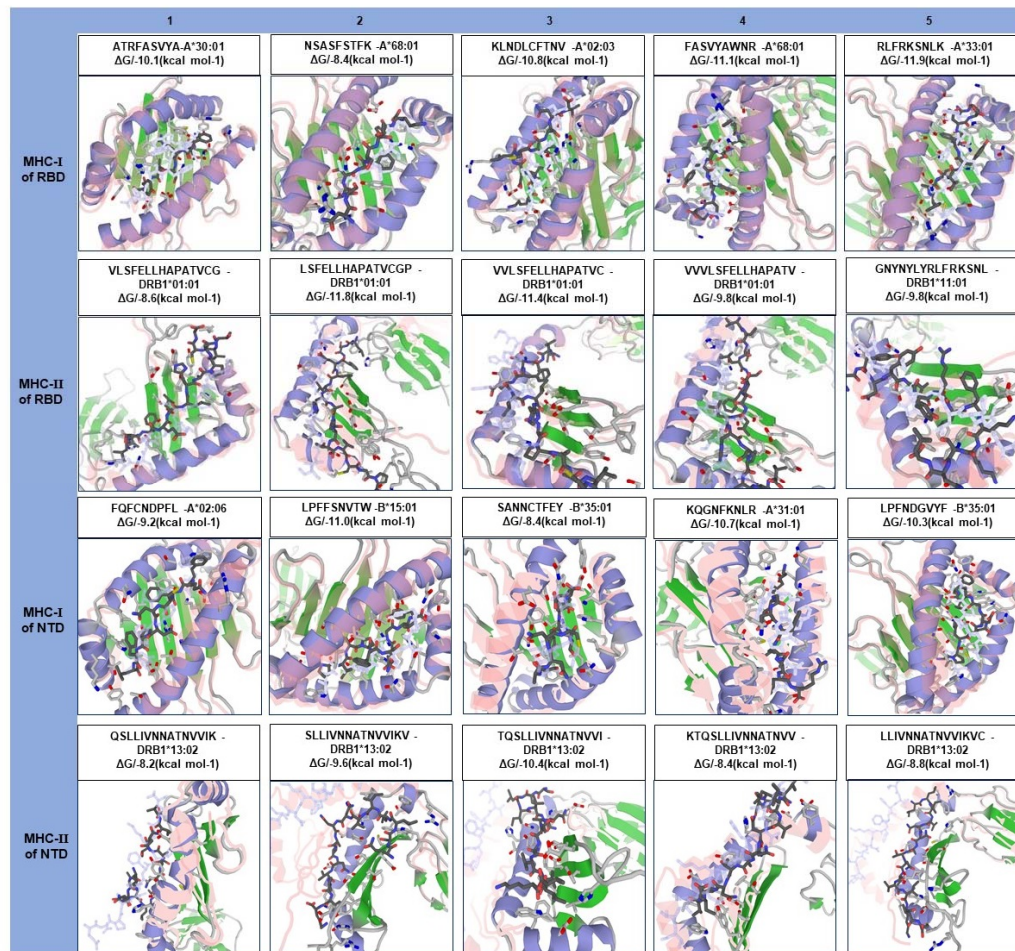
Fig. 4 Molecular docking of top five MHC-Ⅰ and MHC-Ⅱ epitopes of RBD and NTD domains with respect to HLA allele binders. The protein-peptide docking was performed in GalaxyWEB-GalaxyPepDock-server followed by the refinement using GalaxyRefineComplex and free energy (ΔG) of each complex was determined in PRODIGY server. Ribbon structures represent HLA alleles and stick structures represent the respective epitopes. Light color represents the templates to which the alleles and epitopes structures were built. Further information on molecular docking analysis is also available in Supplementary Data 1.
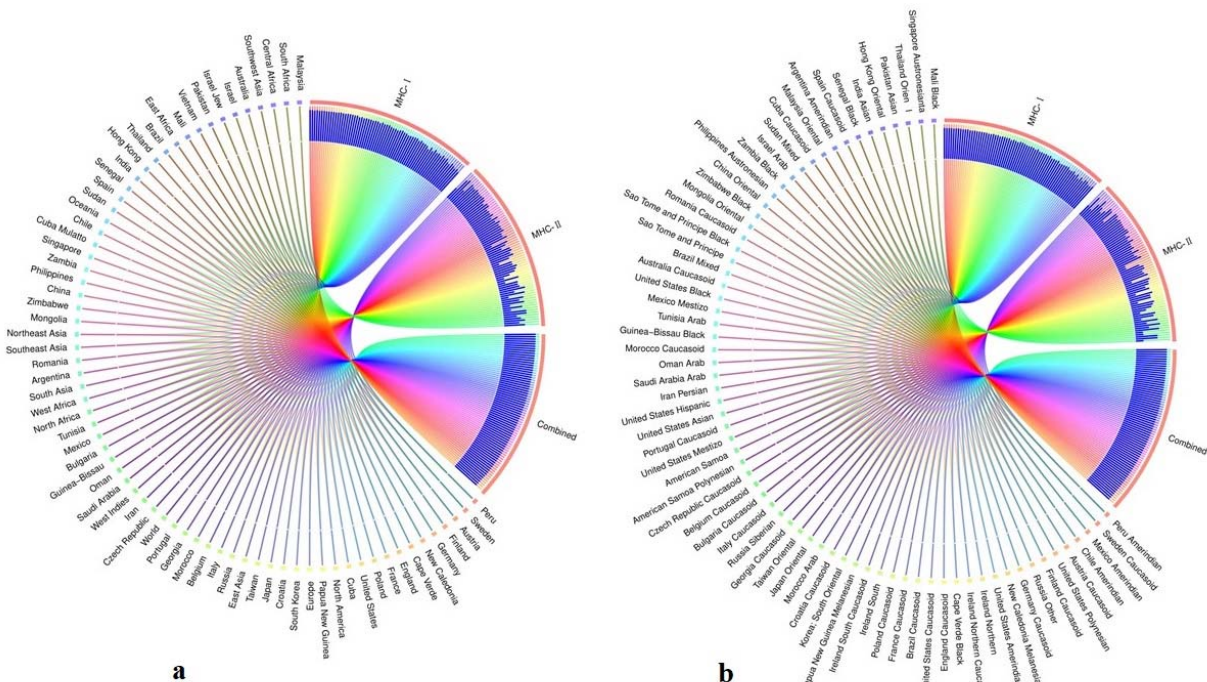
Fig. 5 Population coverage of the selected T-cell epitopes and their respective HLA alleles.

The circular plot illustrates the relative abundance of the top 70 geographic regions and ethnic groups for selected CTL and HTL epitopes, which were used to construct the vaccine and their corresponding MHC HLA alleles were obtained for population coverage analysis both individually (either MHC-□ or MHC-□) and in combination (MHC-□ and MHC-□). (a) Population coverage of top seventy geographical regions out of 123 regions. (b) Population coverage of top seventy ethnic groups selected from 146 ethnic groups. Regions and ethnic groups in the respective MHC-□ and MHC-□ epitopes are represented by different colored ribbons, and the inner blue bars indicate their respective relative coverages. Further information on population coverage analysis is also available in Supplementary Data 1.
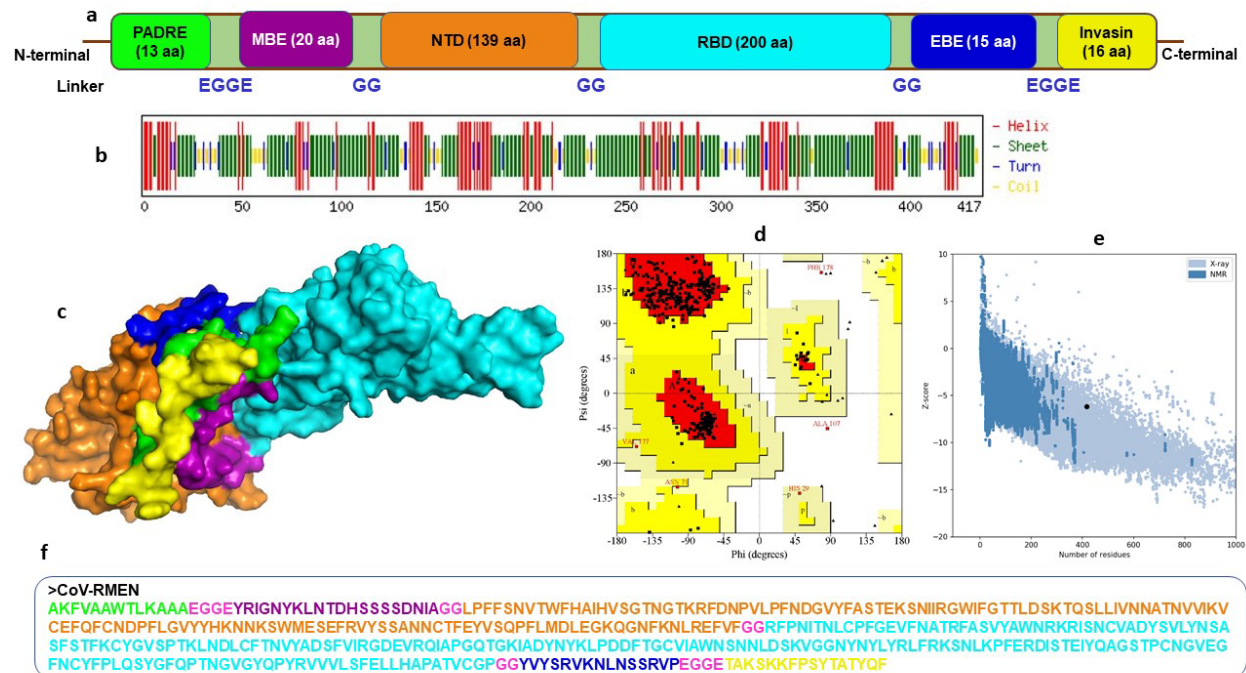
Fig. 6 Design, construction and structural validation of multi-epitope vaccine candidate (CoV-RMEN) for SARS-CoV-2. (a) Structural domains and epitopes rearrangement of CoV-RMEN, (b) secondary structure of CoV-RMEN as analyzed through CFSSP:Chou and Fasman secondary structure prediction server, (c) final tertiary structure of CoV-RMEN (surface view) obtained from homology modelling on Phyre2 in which domains and epitopes are represented in different colors (PADRE-green; membrane B-cell epitope, MBE-magenta; N-terminal domain, NTD-orange; receptor-binding domain, RBD-cyan; envelop B-cell epitope, EBE-blue; invasin-yellow), (d) validation of the refined model with Ramachandran plot analysis showing 94.7%, 4.8% and 0.5% of protein residues in favored, allowed, and disallowed (outlier) regions respectively, (e) ProSA-web, giving a Z-score of –6.17, and (f) the finally predicted primary structure of the CoV-RMEN.
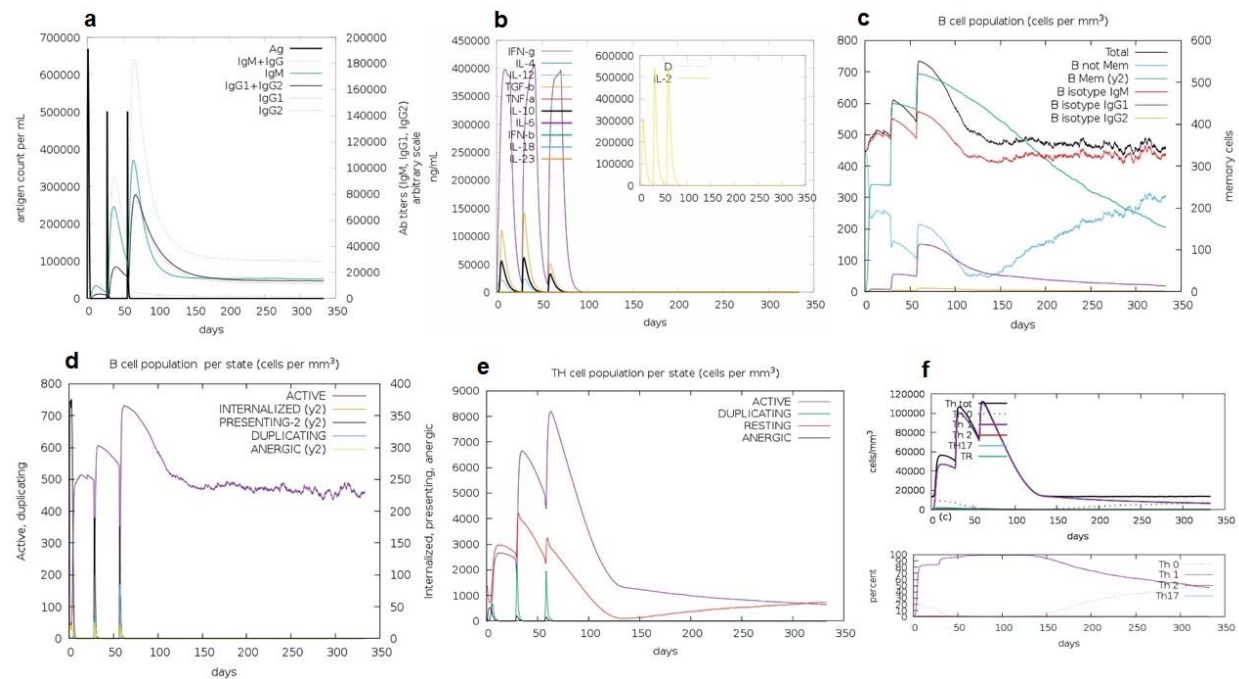
Fig. 7 C-ImmSim presentation of an *in silico* immune simulation with the chimeric peptide.

(a) The immunoglobulins and the immunocomplex response to antigen (CoV-RMEN) inoculations (black vertical lines); specific subclasses are indicated as colored peaks, (b) concentration of cytokines and interleukins, and inset plot shows danger signal together with leukocyte growth factor IL-2, (c) B-cell populations after three injections, (d) evolution of B cell, (e) T-helper cell populations per state after injections, and (f) evolution of T-helper cell classes with the course of vaccination.
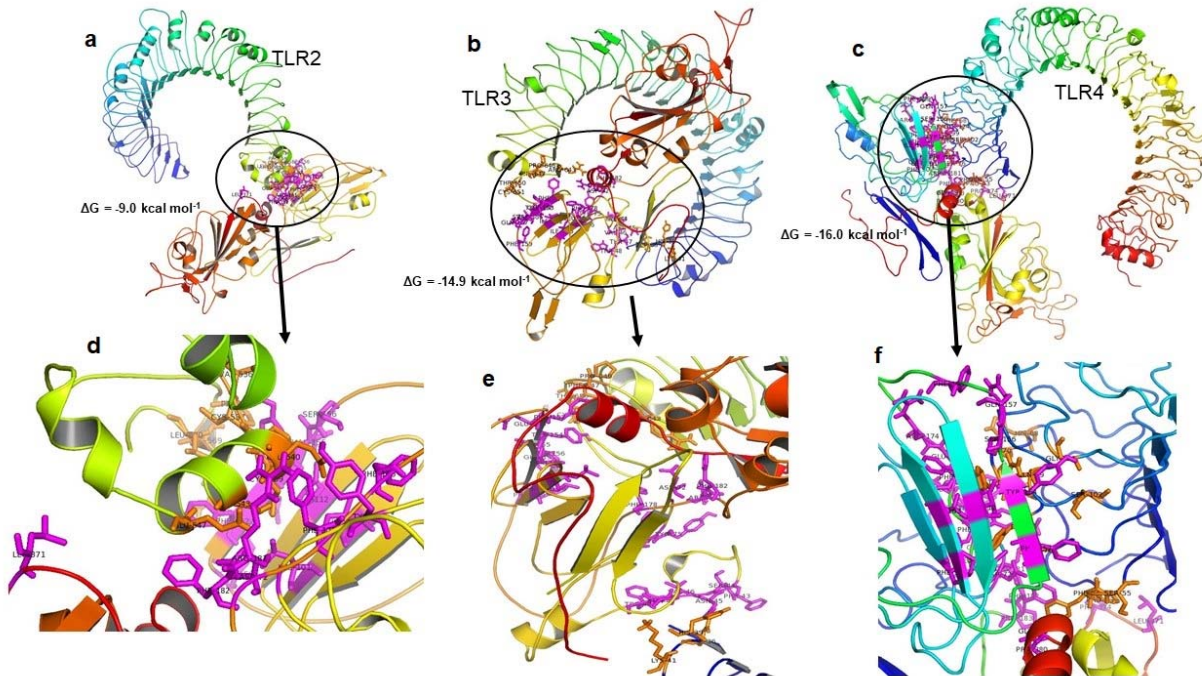
Fig. 8 Molecular docking of CoV-RMEN vaccine with immune receptors (TLR2, TLR3 and TLR4). Docked complexes for (a) CoV-RMEN and TLR2, (b) CoV-RMEN and TLR3, and (c) CoV-RMEN and TLR4. Magnified interfaces of the complexes are figured to (d), (e) and (f) respectively. Active residues of CoV-RMEN colored magenta, and of TLRs colored orange with stick view. ΔG represents the binding affinity of the complexes.
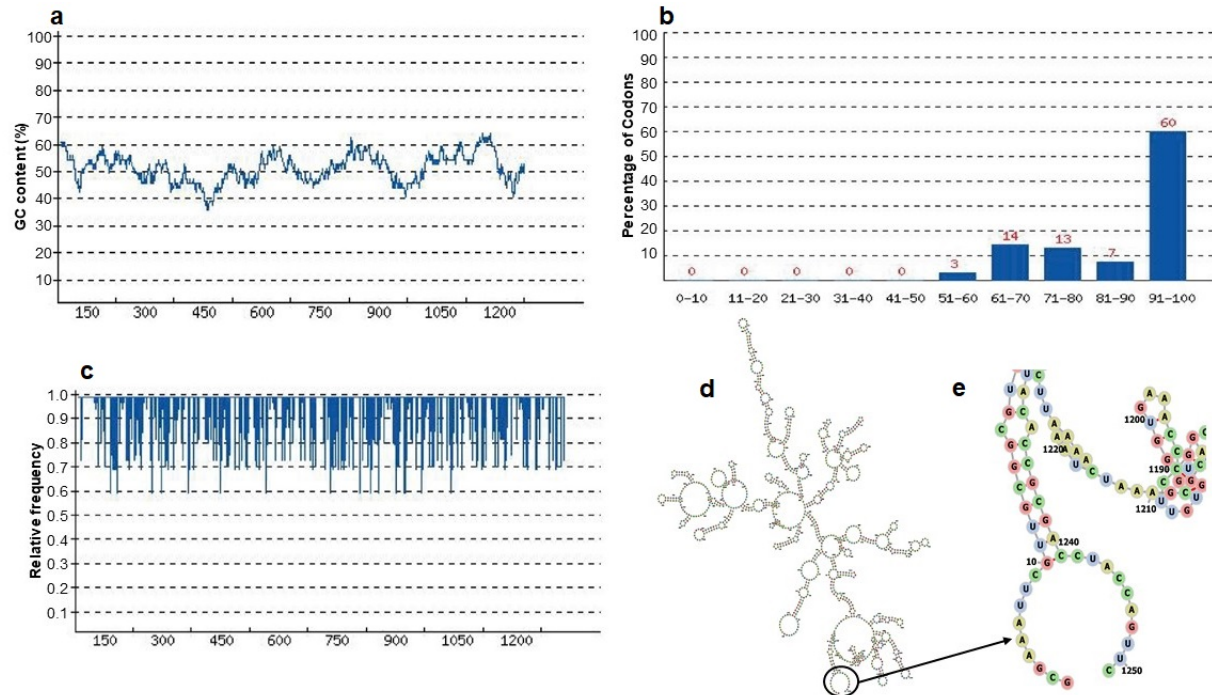
Fig. 9 Codon optimization and mRNA structure of CoV-RMEN gene for expression in *E. coli* (a) GC curve (average GC content: 50.26%) of the optimized CoV-RMEN gene, (b) percentage distribution of codons in computed codon quality groups, (c) relative distribution of codon usage frequency along the gene sequence to be expressed in *E. coli*, and codon adaptation index (CAI) was found to be 0.87 for the desired gene, (d) secondary structure and stability of corresponding mRNA, and (e) resolved view of the start region in the mRNA structure of CoV-RMEN.

Fig. 10 *In silico* fusion cloning of the CoV-RMEN. The final vaccine candidate sequence was inserted into the pETite expression vector where the red part represents the gene coding for the predicted vaccine, and the black circle represents the vector backbone. The six His-tag and SUMU-tag are located at the Carboxy-terminal end.

March 30, 2020

Alan D.T. Barrett, PhD

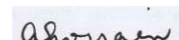Editor-in-Chief

npj Vaccines

Dear Editor,

I am pleased to submit an original research article entitled **"Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an *in silico* approach"** for your kind consideration for publication in the **npj Vaccines**, a renowned Nature publishing journal. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the ongoing pandemic of coronavirus disease 2019 (COVID-19), a public health emergency of international concern declared by the World Health Organization (WHO). An immuno-informatics approach along with comparative genomic was applied to design a multi-epitope-based peptide vaccine against SARS-CoV-2 combining the antigenic epitopes of the S, M and E proteins.

Here, I would like to request you to have APC waivers and discounts (Bangladesh, lower-middle-income country) for this manuscript as per the rules of the journal.

Therefore, me and rest of the co-authors of this manuscript do firmly believe and hope that you and the reviewer panel will consider this manuscript suitable for publication in **npj Vaccines** journal.

Thanking you for kind consideration.

Sincerely,

The Corresponding Author
M. Anwar Hossain, PhD
Vice-Chancellor
Jashore University of Science and Technology, and
Professor, Department of Microbiology
University of Dhaka, Dhaka 1000, Bangladesh
Email: hossaina@du.ac.bd

Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an *in silico* approach

M. Shaminur Rahman, M. Nazmul Hoque, M. Rafiul Islam, Salma Akter, A.S.M. Rubayet- Ul-Alam, Mohammad Anwar Siddique, Otun Saha, Md. Mizanur Rahaman, Munawar Sultana, M. Anwar Hossain

Table 1: Linear epitopes present on spike (S) glycoprotein surface predicted through ElliPro in IEDB-analysis resource based upon solvent-accessibility and flexibility are shown with their antigenicity scores. The highlighted green coloured regions were the potential antigenic domains while the yellow coloured region represents the trans-membrane domain of the S protein.

No.	Chain	Start	End	Peptide	Residues	Score
1		395	514	VYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPF ERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLS	120	0.837
2		58	194	FFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIK VCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVF	137	0.835
3		1067	1146	YVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNT VYDPLQPELD	80	0.83
4	A	201	270	FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYL	70	0.76
5		331	381	NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYG	51	0.706
6		700	720	GAENSVAYSNNSIAIPTNFTI	21	0.668
7		27	35	AYTNSFTRG	9	0.66
8		909	936	IGVTQNVLYENQKLIANQFNSAIGKIQD	28	0.633
9		789	813	YKTPPIKDFGGFNFSQILPDPSKPS	25	0.6
10		623	642	AIHADQLTPTWRVYSTGSNV	20	0.598
11		891	907	GAALQIPFAMQMAYRFN	17	0.591
12		579	583	PQTLE	5	0.551
13		687	692	VASQSI	6	0.55
14		653	659	AEHVNNS	7	0.539
15		679	684	NSPRRA	6	0.521
16		1067	1146	YVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNT VYDPLQPELD	80	0.826
17		89	194	GVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRV YSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVF	106	0.816
18		58	87	FFSNVTWFHAIHVSGTNGTKRFDNPVLPFN	30	0.81
19		203	270	IYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYL	68	0.748
20		465	509	ERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYR	45	0.727
21		436	458	WNSNNLDSKVGGNYNYLYRLFRK	23	0.672
22		700	720	GAENSVAYSNNSIAIPTNFTI	21	0.671
23		27	35	AYTNSFTRG	9	0.666
24	B	909	9036	IGVTQNVLYENQKLIANQFNSAIGKIQD	28	0.641
25		624	643	IHADQLTPTWRVYSTGSNVF	20	0.617
26		328	365	RFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADY	38	0.608
27		891	907	GAALQIPFAMQMAYRFN	17	0.602
28		577	583	RDPQTLE	7	0.598
29		790	817	KTPPIKDFGGFNFSQILPDPSKPSKRSF	28	0.595
30		673	693	SYQTQTNSPRRARSVASQSII	21	0.567
31		526	537	GPKKSTNLVKNK	12	0.553
32		653	661	AEHVNNSYE	9	0.548
33		554	563	ESNKKFLPFQ	10	0.52
34		56	194	LPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNV VIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVF	139	0.84
35		1067	1146	YVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNT VYDPLQPELD	80	0.822
36		201	270	FKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYL	70	0.77
37		27	35	AYTNSFTRG	9	0.676
38		465	509	ERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYR	45	0.675
39		700	720	GAENSVAYSNNSIAIPTNFTI	21	0.658
40		909	936	IGVTQNVLYENQKLIANQFNSAIGKIQD	28	0.633
41	C	437	458	NSNNLDSKVGGNYNYLYRLFRK	22	0.629
42		673	684	SYQTQTNSPRRA	12	0.619
43		790	817	KTPPIKDFGGFNFSQILPDPSKPSKRSF	28	0.602
44		891	907	GAALQIPFAMQMAYRFN	17	0.597
45		578	583	DPQTLE	6	0.589
46		620	631	VPVAIHADQLTP	12	0.579
47		329	362	FPNITNLCPFGEVFNATRFASVYAWNRKRISNCV	34	0.571

48		687	692	VASQSI	6	0.566
49		835	845	KQYGDCLGDIA	11	0.567
50		653	659	AEHVNNS	7	0.559
51		527	536	PKKSTNLVKN	10	0.546
52		635	642	VYSTGSNV	8	0.51

Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an *in silico* approach

M. Shaminur Rahman, M. Nazmul Hoque, M. Rafiul Islam, Salma Akter, A.S.M. Rubayet- Ul-Alam, Mohammad Anwar Siddique, Otun Saha, Md. Mizanur Rahaman, Munawar Sultana, M. Anwar Hossain

Table 2: B-cell epitopes predicted using Bepipred linear epitope prediction 2.0 in IEDB analysis resource web-server along with their start and end positions, average score, and VaxiJen 2.0 determined antigenicity scores.

Domain/proteins	Position	Sequences	Average Score	Antigenicity
RBD	14-15	VF	0.502	-
	17-22	ATRFAS	0.520	-0.151
	24-36	YAWNRKRISNCVA	0.522	0.394
	45-51	ASFSTFK	0.527	0.087
	55	V	0.464	-
	75-100	IRGDEVRQIAPGQTGKIADYNYKLPD	0.575	0.932
	113-158	NLDSKVGGNYNYLYRLFRKSNLKPFER DISTEIYQAGSTPCNGVEG	0.554	0.210
	166-189	QSYGFQPTNGVGYQ	0.535	0.670
NTD	17-26	GTNGTKRFDN	0.573	0.667
	55-58	LDSK	0.511	-
	91-100	HKNNKSWMES	0.573	0.174
	106-107	SS	0.503	-
	109	N	0.499	-
	117-136	SQPFLMDLEGKQGNFKNLRE	0.553	0.749
MBE	199-218	YRIGNYKLNTDHSSSSDNIA	0.614	0.222
EBE	57-71	YVYSRVKNLNSSRVP	0.565	0.449

Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an *in silico* approach

M. Shaminur Rahman, M. Nazmul Hoque, M. Rafiul Islam, Salma Akter, A.S.M. Rubayet- Ul-Alam, Mohammad Anwar Siddique, Otun Saha, Md. Mizanur Rahaman, Munawar Sultana, M. Anwar Hossain

Table 3: Active interface amino acid residues and binding scores among Toll Like Receptors (TLRs) and the constructed vaccine CoV-RMEN.

	Active residues of TLRs	Active residues of CoV-RMEN	HADDOCK score	ΔG (kcal mol^{-1})
TLR-2	V536, C537, S538, C539, E540, S543, E547, P567, R569, L570	D72, Y75, L101, I103, I112, C150, F152, E153, Y154, V155, S156, F176, F178, R181, F182, L371	-30.4	-9.0
TLR-3	D36, H39, K41, R643, F644, P646, F647, T650, C651, E652, S653, I654, W656, F657, V658, N659, W660, I661, N662, E663	F43, S44, N45, V46, T47, W48, D72, Y75, F76, L101, I103, I112, F152, E153, Y154, V155, S156, Q157, F159, F178, R181, F182, L371	-47.2	-14.9
TLR-4	P53, F54, S55, H68, G70, Y72, S73, F75, S76, Q99, S102, G124,	G39, L40, D72, Y75, F76, F90, L101, I103, I112, F152, Y154, S156, Q157, F159, R174, E175, F176, F178, R181, F182, P183, L371, P374, P380, G381	-52.1	-16.0