Whole proteome screening and identification of potential epitopes of SARS-CoV-2 for vaccine design-an immunoinformatic, molecular docking and molecular dynamics simulation accelera…

Article *in* Journal of biomolecular Structure & Dynamics · February 2021

DOI: 10.1080/07391102.2021.1886171

CITATIONS
3

READS
386

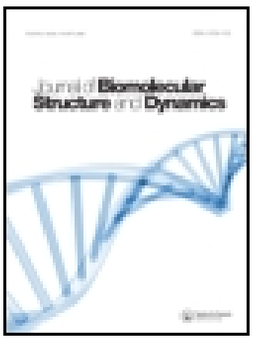
10 authors, including:

Md. Muzahid Ahmed Ezaj
University of Chittagong
9 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Md. Junaid
Advanced Bioinformatics, Computational Biology and Data Science Laboratory, B…
72 PUBLICATIONS **229** CITATIONS
SEE PROFILE

Yeasmin Akter
West Virginia University
23 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Afsana Nahrin
North South University
16 PUBLICATIONS **6** CITATIONS
SEE PROFILE



Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbsd20

Whole proteome screening and identification of potential epitopes of SARS-CoV-2 for vaccine design-an immunoinformatic, molecular docking and molecular dynamics simulation accelerated robust strategy

Md. Muzahid Ahmed Ezaj , Md. Junaid , Yeasmin Akter , Afsana Nahrin , Aysha Siddika , Syeda Samira Afrose , S. M. Abdul Nayeem , Md. Sajedul Haque , Mohammad Ali Moni & S. M. Zahid Hosen

To cite this article: Md. Muzahid Ahmed Ezaj , Md. Junaid , Yeasmin Akter , Afsana Nahrin , Aysha Siddika , Syeda Samira Afrose , S. M. Abdul Nayeem , Md. Sajedul Haque , Mohammad Ali Moni & S. M. Zahid Hosen (2021): Whole proteome screening and identification of potential epitopes of SARS-CoV-2 for vaccine design-an immunoinformatic, molecular docking and molecular dynamics simulation accelerated robust strategy, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2021.1886171

To link to this article: https://doi.org/10.1080/07391102.2021.1886171

JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
https://doi.org/10.1080/07391102.2021.1886171

Whole proteome screening and identification of potential epitopes of SARS-CoV-2 for vaccine design-an immunoinformatic, molecular docking and molecular dynamics simulation accelerated robust strategy

Md. Muzahid Ahmed Ezaj[a,b*], Md. Junaid[b,c*], Yeasmin Akter[b,d*], Afsana Nahrin[e], Aysha Siddika[b,f], Syeda Samira Afrose[b,f], S. M. Abdul Nayeem[b,f], Md. Sajedul Haque[f], Mohammad Ali Moni[g] and S. M. Zahid Hosen[c,h]

[a]Department of Genetic Engineering and Biotechnology, University of Chittagong, Chattogram, Bangladesh; [b]Reverse Vaccinology Research Division, Advanced Bioinformatics, Computational Biology and Data Science Laboratory, Chattogram, Bangladesh; [c]Molecular Modeling Drug-design and Discovery Laboratory, Pharmacology Research Division, BCSIR Laboratories Chattogram, Bangladesh Council of Scientific and Industrial Research, Chattogram, Bangladesh; [d]Department of Biotechnology & Genetic Engineering, Noakhali Science & Technology University, Noakhali, Bangladesh; [e]Department of Pharmacy, University of Science and Technology Chittagong, Chattogram, Bangladesh; [f]Department of Chemistry, University of Chittagong, Chattogram, Bangladesh; [g]WHO Collaborating Centre on eHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, UNSW Sydney, Sydney, NSW, Australia; [h]Pancreatic Research Group, South Western Sydney Clinical School, and Ingham Institute for Applied Medical Research, Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia

Communicated by Ramaswamy H. Sarma

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the most cryptic pandemic outbreak of the 21st century, has gripped more than 1.8 million people to death and infected almost eighty six million. As it is a new variant of SARS, there is no approved drug or vaccine available against this virus. This study aims to predict some promising cytotoxic T lymphocyte epitopes in the SARS-CoV-2 proteome utilizing immunoinformatic approaches. Firstly, we identified 21 epitopes from 7 different proteins of SARS-CoV-2 inducing immune response and checked for allergenicity and conservancy. Based on these factors, we selected the top three epitopes, namely KAYNVTQAF, ATSRTLSYY, and LTALRLCAY showing functional interactions with the maximum number of MHC alleles and no allergenicity. Secondly, the 3D model of selected epitopes and HLA-A*29:02 were built and Molecular Docking simulation was performed. Most interestingly, the best two epitopes predicted by docking are part of two different structural proteins of SARS-CoV-2, namely Membrane Glycoprotein (ATSRTLSYY) and Nucleocapsid Phosphoprotein (KAYNVTQAF), which are generally target of choice for vaccine designing. Upon Molecular Docking, interactions between selected epitopes and HLA-A*29:02 were further validated by 50 ns Molecular Dynamics (MD) simulation. Analysis of RMSD, Rg, SASA, number of hydrogen bonds, RMSF, MM-PBSA, PCA, and DCCM from MD suggested that ATSRTLSYY is the most stable and promising epitope than KAYNVTQAF epitope. Moreover, we also identified B-cell epitopes for each of the antigenic proteins of SARS CoV-2. Findings of our work will be a good resource for wet lab experiments and will lessen the timeline for vaccine construction.

ARTICLE HISTORY
Received 30 August 2020
Accepted 1 February 2021

KEYWORDS
SARS-CoV-2; T lymphocyte epitope; immunoinformatics; molecular docking; molecular dynamics simulation

1. Introduction

Coronaviruses are enveloped viruses having 27 to 32 kb positive-stranded RNA genome and belong to the genus of the Coronaviridae family. Coronaviruses typically infect animals, generally birds, and mammals (van der Hoek et al., 2004). Among the seven coronavirus strains detected over the last couple of decades, a novel one was reported in Wuhan City, China in December 2019 (Hui et al., 2020; Lu et al., 2020). This virus was named 2019 novel coronavirus (2019-nCoV) and World Health Organization (WHO) declared global health emergency on 30 January 2020 [https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-ofthe-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novelcoronavirus-(2019-ncov)], because of its terrifyingly rapid transmission to 213 other countries around the world [https://www.cdc.gov/coronavirus/2019-ncov/locations-confirmed-cases.html, accessed in 14th May 2020]. Scientists from all over the world are putting collaborative efforts to identify and understand the pathogenicity and mode of action of this severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; originally

CONTACT S. M. Zahid Hosen s.hosen@student.unsw.edu.au Molecular Modeling Drug-design and Discovery Laboratory, Pharmacology Research Division, BCSIR Laboratories Chattogram, Bangladesh Council of Scientific and Industrial Research, Chattogram, Bangladesh; Pancreatic Research Group, South Western Sydney Clinical School, and Ingham Institute for Applied Medical Research, Faculty of Medicine, University of New South Wales, Sydney, NSW, Australia
*These authors equally contributed to this paper.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07391102.2021.1886171.

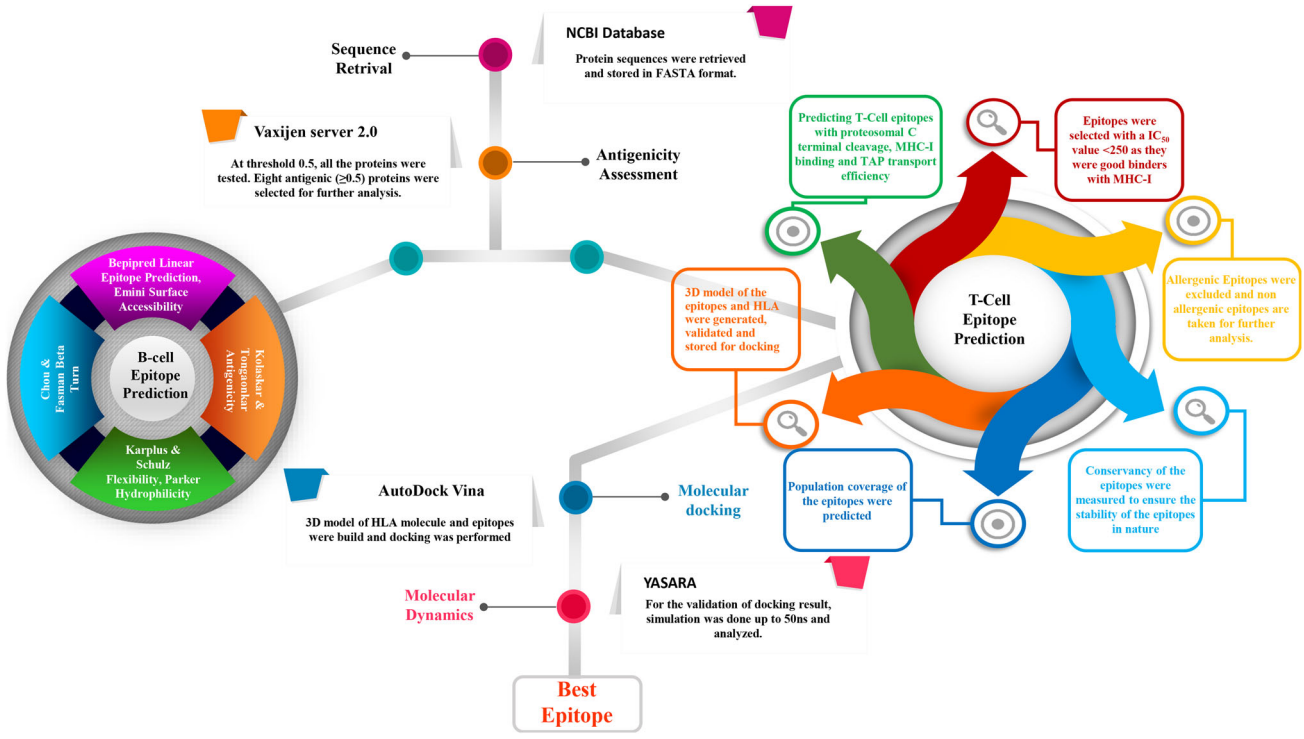


Figure 1. Graphical depiction of the methodologies used in epitope-based vaccine design.

tentatively named 2019-nCoV) virus, and to advance effective therapeutic strategies for preventing and controlling the COVID-19 infections (Gao et al., 2020; Heymann, 2020; Huang et al., 2020; Liu et al., 2020; Liu & Wang, 2020; Robson, 2020; Wu & McGoogan, 2020; Zhou et al., 2020).

The evolutionary study identified a bat origin for the spread of 2019-nCoV (Wan et al., 2020; Zhou et al., 2020), indicating it as an airborne transmission (human-to-human), with characteristics fever, diarrhoea, upper or lower respiratory tract symptoms, thrombocytopenia, lymphopenia, and elevated level of C-reactive protein and lactate dehydrogenase or their combination from 3–6 days after acquaintance (Chan et al., 2020). The recent 2019-nCoV or SARS-CoV-2 is a member of the genus Betacoronavirus as like SARS-CoV and MERS-CoV (Lu et al., 2020). Similar to other coronaviruses, it has *a* ~30 kb genome size that encodes for various structural and non-structural proteins. The structural proteins comprise the membrane (M) protein, the nucleocapsid (N) protein, envelope (E) protein, and spike (S) protein. Due to the recent outbreak of the SARS-CoV-2 virus, its antigenic epitopes that elicit an immune response is not well understood, and immunological information is deficient. Initial studies proposed that SARS-CoV-2 is highly related to SARS-CoV based on the phylogenetic analysis of their genome (Lu et al., 2020; Zhou et al., 2020) presumably similar mechanism of cell entry and usage of human cell receptor (Hoffmann et al., 2020; Letko & Munster, 2020; Zhou et al., 2020). The receptor-binding domain (RBD) sequence of 2019-nCoV (SARS-CoV), in association with its receptor binding motif (RBM), used to contact receptor angiotensin-converting enzyme 2 (ACE2), showing similarity with SARS coronavirus that strongly suggest that 2019-nCoV utilizes ACE2 as its receptor. Moreover, compared to other previous strains, 2019-nCoV utilizes some important residues at its receptor binding motif, i.e. Gln493, which helps the virus to interact with human ACE2 (Wan et al., 2020). Thus, this apparent resemblance between SARS-CoV and SARS-CoV-2 will help to understand the protective antibody responses against it and design of vaccine by analyzing previous research against SARS-CoV.

Different studies regarding SARS-CoV recommend a defensive role of humoral and cell-mediated immune responses in case of viral infection. According to the earlier studies, immune responses produced against the most exposed spike S protein of SARS-CoV, have been demonstrated to defend viral infection in mouse models (Deming et al., 2006; Graham et al., 2012; Yang et al., 2004). Moreover, numerous studies have proved that antibody responses produced against the potent antigenic and highly expressed SARS-CoV nucleocapsid (N) protein during infection (Lin et al., 2003), were mostly dominant in SARS-CoV-infected patients (Liu et al., 2004; Wang et al., 2003). Another important thing is that compared to the non-structural proteins, T cell responses generated against the structural proteins have been observed to be the most antigenic in peripheral blood mononuclear cells of recovering SARS-CoV patients (Li et al., 2008). To prevent the viral infection, previously traditional vaccine design approaches were utilized by the researcher. Traditional vaccine design includes biochemical trials that are time-consuming, costly, and necessitate the culture of pathogenic viruses in vitro, causing reactogenic or allergic responses (Li et al., 2014). In contrast, peptide-based vaccines design do not require in vitro study that makes them biologically harmless, and their selectivity enables precise stimulation of immune responses (Dudek et al., 2010; Purcell et al., 2007). Hence, our present study aims to design a peptide vaccine employing bioinformatics tools (Bahrami et al., 2019; Brusic & Petrovsky, 2005; Hegde et al., 2018; Khalili et al., 2014).

2. Materials and methods

The systematic workflow for the step by step procedures which have been followed to predict the most probable peptide-based epitope vaccine for coronavirus is depicted in Figure 1.

2.1. Protein sequence retrieval

Though structural proteins are targeted for vaccine designing, sometimes non structural proteins can also contribute to T-cell and B-cell mediated immunity (Rascón-Castelo et al., 2015). For this reason, to predict the possible SARS CoV-2 specific vaccine, we retrieved the complete genome sequence of SARS-CoV-2 using the NCBI database (GenBank Assembly ID: GCF_009858895.2), (RefSeq: NC_045512.2). Then all the 12 protein sequences of the genome were stored as FASTA format using Refseq accessions for further analysis.

2.2. Antigenic protein identification

All of the selected proteins were submitted to the VaxiJen server (Doytchinova & Flower, 2007), which is stand on auto cross-covariance (ACC) transformation of protein series into normal vectors of prime amino acid properties, a Web-based Server that is used for the prophecy of protective antigens and subunit vaccines. For the highest accuracy, we used the threshold at 0.5 to isolate the antigenic proteins (Oany et al., 2014) to check the antigenicity of each full-length protein. All antigenic proteins were then sorted out accordingly to their relevant score (Hasan et al., 2013).

2.3. T cell epitope identification

For the prediction of human cytotoxic T lymphocyte (CTL) epitopes in the selected protein sequences, NetCTL-1.2 server (http://www.cbs.dtu.dk/services/NetCTL/) (Larsen et al., 2007) is designed by which epitopes were selected from the antigenic sequences (Tenzer et al., 2005). NetCTL is an artificial neural network (ANN) and weight matrix-based tool combining the prediction of peptide MHC-I binding, proteasomal C terminal cleavage, and TAP transport efficiency (Mehla & Ramana, 2016).

For the selected epitopes, a web-based tool 'T-Cell Epitope calculation tool' was employed for the identification of peptide with allele molecules MHC-I (http://tools.iedb.org/mhci/) (Tenzer et al., 2005) alleles molecule binding into the peptide at the Immune Epitope Database (IEDB) using the stabilization matrix method-based prediction method for each peptide. The Stabilized Matrix Method-based (SMM) predicted the half-maximal inhibitory concentration (IC_{50}) values that required peptide binding to specific MHC-I molecules, all the alleles were selected, and the length was set at 9.0 earlier to the prediction.

2.4. Prediction of epitope conservancy

The degree of resemblance between the epitope and the target sequence is predicted by epitope conservancy. This property of epitope assures us of its availability in a range of different strains. The Conservancy of predicted epitopes with all the 18,149 protein sequences from different strains of all SARS-CoV-2, taken from the NCBI Virus database, was analyzed by the Web-based tool IEDB analysis resource (http://tools.immuneepitope.org/conservancy/) (Bui et al., 2007).

2.5. Population coverage calculation

We have employed the IEDB population coverage tool (http://tools.immuneepitope.org/population/) (Angelo et al., 2017) for predicted peptides with their corresponding MHC-I and MHC-II alleles and studied the distribution of human HLA alleles among the predicted epitopes (Bappy et al., 2020). Here we used the allelic frequency of the interacting HLA alleles for the prediction of the population coverage for the equivalent epitope and HLA combinations identified by 90% of the population (PC90). Before the submission, we have to add every epitope and their MHC-I molecules and selected population coverage area (Bui et al., 2007).

2.6. Allergenicity assessment

For the evaluation of allergenicity of our suggested proteins and epitopes, an online server AlgPred (http://crdd.osdd.net/raghava/algpred/) (Saha & Raghava, 2006) was implemented. This server predicts allergenicity through a conjunctional prediction, by using both integration of the Food and Agriculture Organization (FAO)/World Health Organization (WHO) allergenicity evaluation scheme and support vector machines (SVM)-pairwise sequence similarity. Here it can predict the selected proteins as allergens and nonallergens with high sensitivity and specificity without compromising efficiency at the classification of proteins with similar sequences to known allergens. To ensure more accurate prediction, another online tool, named AllerCatPro 1.7 (https://allercatpro.bii.a-star.edu.sg) (Maurer-Stroh et al., 2019), was used to check the allergenicity of the epitopes.

2.7. Modelling and validation of predicted epitopes and HLA alleles

Three-dimensional structure of the best performing epitopes sequences was predicted by using PEPstrMOD server (http://osddlinux.osdd.net/raghava/pepstrmod) (Kaur et al., 2007; Singh et al., 2015). The tertiary structure of the peptide is analyzed by PEPstrMOD modified version of PEPstr employing β-turn and regular secondary structure information as well (Kaur et al., 2007). The modelled structure was optimized for removing specific stereochemical errors present therein in a systematic manner, employing the Protein Preparation Wizard from the Schrödinger suite (Vetrivel et al., 2018).

Phyre2 (Kelley et al., 2015), a widespread accessible web server for the recognition of protein fold that was used in our study for the construction of the homology-based three-dimensional structure of HLA-A*29:02 protein. By employing OPLS 2005 (Banks et al., 2005) force field with VSGB solvation model (Li et al., 2011), Phyre2 generated model further subjected to loop refinement approach where non-template

loops were adjusted by extended sampling. Structure preparation and refinement was performed by utilizing the Protein Preparation Wizard of Schrödinger-Maestro v9.4. Again, the whole protein structure was optimized upon the addition of charges, bonds, and hydrogen at neutral pH. Afterwards, protein minimization was performed via the OPLS 2005 force field, and by setting maximum heavy atom RMSD to 0.30 Å. The force field YAMBER3 (Krieger et al., 2004) was also employed keeping default the simulation parameters characterized by the macro. For obtaining a better and stable quality, the model was additionally subjected to MD refinement utilizing the YASARA software (Krieger et al., 2013) to 500 ps molecular dynamics simulation at a pH 7.4, temperature of 298 K, and solvent density of 0.997. The best snapshot possessing the lowest force field energies was selected to run the further analysis. Finally, the predicted structure was validated by utilizing PROCHECK (Laskowski et al., 1996), VERIFY3D (Eisenberg et al., 1997), ERRAT (Colovos & Yeates, 1993), and QMEAN (Benkert et al., 2008).

2.8. Molecular docking analysis

In this study AutodockVina was used for the docking analysis, and modelled HLA-A*29:02 molecule and best epitopes considered as protein and ligands, respectively, were input as pdbqt format converted from pdb by AutoDock Tools(ADT) of the MGL software package. The grid box generated here in AutodockVina with the size of 65.7388, 69.2971 and 55.1927 respectively for X, Y, Z. AutodockVina was used to determine the binding affinity between the ligands and target protein and the binding affinity of ligands was observed in negative score in kcal/mole as a unit (Trott & Olson, 2010).

2.9. Molecular dynamics simulation

All the molecular dynamics simulations were performed using the YASARA Dynamic program package along with AMBER14 force field (Dickson et al., 2014). Additionally, before the simulation, each protein was exposed to cleaning with hydrogen bond network optimization (Krieger et al., 2012a). Utilizing the transferable intermolecular potential3 points (TIP3P) water model, the simulation system was solvated, keeping the density of 0.997 g L-1. The acid dissociation constant (pKa) of protein titratable amino acids were estimated, and extra Na + and Cl- ions are added to maintain the desired physiological conditions (pH 7.4, 0.9% NaCl) in the solvation system. Then the model investigated was subjected to steepest descent energy minimization (5000 cycles) utilizing the simulated annealing method. PME method is also employed to demonstrate the long-range electrostatic interactions at a cut off distance of 8 Å. Finally, 50 ns MD simulations were accomplished under physiological conditions (Ph 7.4, 298 K, 0.9% NaCl) at constant temperature employing a Berendsen Thermostat and constant pressure (Krieger et al., 2012b). Multiple time-step algorithms were utilized with a 2.50 fs time step interval during the entire simulation (Krieger & Vriend, 2015). For every 0.25 ns, atomic positions were saved for additional analysis. The resulted MD trajectories were further subjected for comprehensive analysis to estimate the structural changes and stability employing RMSD (Root Mean Square Deviation), RMSF (Root Mean Square Fluctuation), the radius of gyration, SASA (Solvent Accessible Surface Area), number of hydrogen bonds, and initial and final protein backbone evaluations using YASARA (Krieger et al., 2002) structure built-in macros and VMD software (Humphrey et al., 1996).

2.10. Binding free energy calculation

After the Molecular Dynamics Simulation, MM-PBSA (Molecular mechanics-Poisson–Boltzmann Surface Area) binding free energy calculation done for all snapshots employing YASARA software using the following formula,

$$\begin{aligned}\text{Binding Energy} = {} & \text{EpotRecept} + \text{EsolvRecept} + \text{EpotLigand} \\ & + \text{EsolvLigand} - \text{EpotComplex} \\ & - \text{EsolvComplex}\end{aligned}$$

Here, YASARA (Krieger et al., 2002) built-in macros was applied to calculate MMPBSA binding energy, using AMBER 14 as a force field, where more positive energies indicate good binding, and negative energies do not indicate any binding (Srinivasan & Rajasekaran, 2016).

2.11. Principal component analysis (PCA)

The principal component analysis is an extensively employed unsupervised data reduction method for describing the variation in the collective energy profile of MD trajectory data (Martens & Naes, 1992; Sittel et al., 2014; Wold et al., 1987) as well as a method that minimizes the complexity of the data and selects the different modes involved in the protein motion by calculating eigenvectors, eigenvalues, and their projection along with the first two principal components (Amadei et al., 1993). Using the first technique, any conformational change during MD can be categorized by evaluating the different docked complexes of drug and protein.

The important properties of a PCA model highlighted by the following equations:

$$X = T_k P_k^T + E$$

where X matrix defines the product of two new matrices, i.e. T_k and P_k, T_k defines the score of matrix indicating how samples correlate to each other, P_k defines the matrix of loadings which enclose data about how variables correlate to each other, k defines the total factors involved in the model, and E defines the matrix of residuals.The unmodeled variances regarded as the residuals. During MD simulation, energy profiles of different complexes of the main protease with the particular epitopes may have variances with the main protease, i.e. apo-protein. These variations can be perceived by the PCA algorithm.

On the other hand, using the second technique of PCA, we can study the direction and degree of the motion of the MD trajectory (Amadei et al., 1993). It is based on the

Table 1. Antigenicity score predicted by VaxiJen v2.0 server at threshold 0.5.

Protein ID	Vaxijen score
YP_009724389.1	0.4624
YP_009724390.1	0.4646
YP_009724391.1	0.4945
YP_009724392.1	0.6025
YP_009724393.1	0.5102
YP_009724394.1	0.6131
YP_009724395.1	0.6441
YP_009724396.1	0.6502
YP_009724397.2	0.5059
YP_009725255.1	0.7185
YP_009725295.1	0.4787
YP_009725296.1	0.8462

construction of a covariance matrix of multifaceted variable sets to shrink the higher-dimension data sets (Ndagi et al., 2017). The atomic shift and protein's conformational difference were characterized, employing PCA by retrieving various forms of protein conformation throughout the MD simulation (Ndagi et al., 2017). The solvent molecules and ions of the protein trajectory were eliminated earlier to process MD trajectory for PCA. Then, PCA was accomplished on Cα atoms. The principal components (PC1, PC2 and PC3) equivalent to the feature vectors of the covariance matrix were estimated to produce a covariance matrix (Farrokhzadeh et al., 2019; Narang et al., 2018). The covariance matrix C is determined by the following collective formula:

$$C_{ij} = \langle x_i - \langle x_j \rangle \rangle \ (i,\ j\ = 1,\ 2,\ 3,\ \ldots,\ n)$$

In the formula above, i and j symbolized the ith and jth Cartesian coordinates of Cα atoms, respectively. Then, xi and xj signified the average of time during the MD simulations in the entire conformations. Moreover, n was the number of protein backbone Cα atoms (Balmith & Soliman, 2017; Yang et al., 2012).

All calculations were executed on the R platform (Islam et al., 2019) utilizing in-house developed codes (R Core Team, 2019), and plots were created via the ggplot2 package (Wickham, 2009). Moreover, data were preprocessed employing autoscale function before the application of the PCA algorithm (Martens & Naes, 1992). The 50 ns of MD trajectory data were utilized for investigating the PCA.

2.12. Dynamic cross-correlation matrix

To demonstrate the time-correlated motions in the protein, the dynamic cross-correlation matrix (DCCM) analysis of the Ca atoms was accomplished by YASARA dynamics. The cross-correlation matrix Cij, which indicates the dislocations of the Cα atoms compared to average positions, was described by the following equation (Ghosh & Vishveshwara, 2007; Lange et al., 2005):

$$C_{ij} = \frac{\langle \Delta R_I \cdot \Delta R_J \rangle}{\sqrt{\langle \Delta R_i \cdot \Delta R_i \rangle \langle \Delta R_j \cdot \Delta R_j \rangle}}$$

where ΔR_i and ΔR_j denotes the dislocation of atom i and j, respectively. In DCCM, the values of Cij fluctuated from −1 to 1, where +1 denotes a correlated motion between the residue i and residue j, and −1 denotes an anti-correlated motion.

Table 2. Epitopes predicted by NetCTL server 1.2 at threshold 0.5.

Protein ID	Epitope	Value
YP_009724392.1	LTALRLCAY	2.6158
	VSLVKPSFY	1.7149
	VFLLVTLAI	1.5566
YP_009724393.1	SSDNIALLV	2.9325
	ATSRTLSYY	2.6146
	LAAVYRINW	1.9258
	YIIKLIFLW	1.7624
	YFIASFRLF	1.7536
YP_009724394.1	KVSIWNLDY	2.6352
YP_009724395.1	QELYSPIFL	1.7455
	VFITLCFTL	1.5865
YP_009724396.1	QSCTQHQPY	2.5149
	NYTVSCLPF	1.5906
	EYHDVRVVL	1.5818
	GSLVVRCSF	1.5177
YP_009724397.2	LSPRWYFYY	2.3408
	ELIRQGTDY	1.9108
	SSPDDQIGY	1.8805
	KAYNVTQAF	1.8413
	KKADETQAL	1.7029
YP_009725296.1	SLIDFYLCF	1.6138

2.13. Prediction of antigenic B cell epitopes

Presumably, immunogenic epitopes in a given protein sequence, are required for the scheme of vaccines and immunodiagnostics, which may effectively reduce wet lab effort. The B cell epitope is the portion of the antigen depicted by being hydrophilic, accessible (Badawi et al., 2016), and in a flexible region of an immunogen (Hasan et al., 2013). It interacts with B Lymphocytes, which can differentiate into an antibody-secreting plasma cell and memory cells. The prediction of B-cell epitopes was performed to find potential antigens that initiate an immune response (Hossain et al., 2018).

B-cell epitopes were predicted through the B-cell epitope prediction tools of IEDB (http://tools.iedb.org/bcell/). For the B-cell epitope prediction with high accuracy IEDB analysis resource was used for the multiple following aspects including:

2.13.1. Prediction of linear B-cell epitopes

Bepipred linear epitope prediction tool (Larsen et al., 2006) uses a combinatorial algorithm comprising both hidden Markov model and propensity scale methods antigenic propensity and thus performs significantly better than any of the other methods (Haste Andersen et al., 2006).

2.13.2. Prediction of surface accessibility

Emini surface accessibility prediction tool used for the identification of accessible surface epitopes from the conserved region (Emini et al., 1985).

2.13.3. Prediction of epitopes antigenicity sites

The Kolaskar and Tongaonkar antigenicity prediction method (Kolaskar & Tongaonkar, 1990) was employed for determining the antigenic sites stand on a table that reflects the occurrence of amino acid residues in experimentally known

segmental epitopes. This method can forecast antigenic peptides with approximately 75% accuracy.

2.13.4. Prediction of hydrophilicity parameters

Parker Hydrophilicity Prediction applies all the hydrophilicity parameters extensively used in all of the algorithms to predict which amino acid residues are antigenic (Kolaskar & Tongaonkar, 1990).

2.13.5. Flexibility prediction

Karplus and Schulz flexibility prediction tools used to portend the flexibility of the epitope region, which is correlated with antigenicity (Karplus & Schulz, 1985).

2.13.6. Beta turn prediction

Chou and Fasman beta-turn prediction analysis tools were used because of the antigenic parts of a protein belonging to the beta-turn regions (Chou & Fasman, 1978).

The results from all tools were cross-referenced, assessment strengthened the possibility of peptide vaccine candidate's cell epitope identification, and common findings were taken as the B-cell epitopes.

3. Results

3.1. Antigenic protein prediction

The antigenicity of the SARS-CoV-2 virus proteins was assessed by using VaxiJen v2.0 server, setting a threshold value of 0.5 for the highest accuracy. The results are listed in Table 1. The proteins YP_009724389.1, YP_009724390.1, YP_009724391.1, and YP_009725295.1 gave a VaxiJen score of <0.5, therefore were excluded from further analysis.

3.2. T Cell epitope identification

Cytotoxic T-cell lymphocyte (CTL) epitopes are known to play a vital role in the effectiveness of a vaccine, and designating the defining those epitopes are identified to significantly reduce the cost and time of research comparative to wet-lab experiments (Zhang et al., 2004). Two distinct online accessible servers NetCTL 1.2 and IEDB were used to predict the potent epitopes of each protein. The epitopes predicted by NetCTL 1.2, with a combinatorial score above 1.5 were selected (Table 2), and we found 21 epitopes which were correlated with the results for MHC-I binding in IEDB. The interacting alleles of common epitopes with an affinity of $IC_{50} < 250$ are depicted in Table 3 including the MHC-I processing total scores. The lower the IC_{50} value of a given epitope, the higher the potency. Though, T-cell epitopes are presented by both class I (MHC I) and II (MHC II) MHC molecules which are recognized by CD8 and CD4 T-cells, respectively, we only predicted the MHC I bound T-cell epitopes in the present study. The MHC I peptide-binding cleft remains closed and can only bind peptides with 9 to 11 amino acid residues, whose N- and C-terminal ends are attached to the MHC I conserved residues via hydrogen bonds. Also, the peptide-binding groove of MHC I has deep binding pockets possessing tight physicochemical preferences that ease binding predictions. Conversely, the peptide-binding groove of MHC II molecules remains open which allows the N- and C-terminal ends of peptide to stretch far-off the binding groove. Moreover, MHC II-bound peptides contain 9–22 amino acid residues, though only a core of nine residues can sit into the MHC II binding groove. As a result, peptide-MHC II binding prediction methods often seek to predict these peptide-binding cores. The binding pockets of MHC II molecule are also superficial and less significant than those of the MHC I molecules. Therefore, in silico peptide-binding prediction of MHC II molecules is less reliable than that of the MHC I molecules (Sanchez-Trincado et al., 2017).

The MHC-I processing total score represents the intrinsic potential of the antigen to be cleaved by the proteasome in the cells, its transportation to the MHC protein on the cells outside using TAP proteins and its ability to bind with MHC itself, to be presentable to helper T-cells later-on bringing about an immunogenic response (Blum et al., 2013). The higher the score, the greater the processing capabilities.

The 9 mer epitope KAYNVTQAF from the protein YP_009724397.2 interacted with most MHC Class I alleles that include, HLA-C*03:02, HLA-B*15:25, HLA-C*12:03, HLA-A*32:01, HLA-C*16:01, HLA-C*03:03, HLA-C*03:04, HLA-B*15:01, HLA-C*14:02, HLA-B*58:01, HLA-C*12:02, HLA-B*15:02, HLA-B*35:01, HLA-C*02:09, HLA-C*02:02, HLA-B*57:01, HLA-C*15:02 with a good affinity (Table 3).

Besides, epitope LTALRLCAY from protein YP_009724392.1 and ATSRTLSYY and YFIASFRLF from protein YP_009724393.1 showed promising interactions with 11, 9, and 9 types of MHC class I allele, respectively.

However, the epitope QSCTQHQPY predicted by NetCTL 1.2 from YP_009724396.1 protein; SSPDDQIGY and KKADETQAL from YP_009724397.2 did not meet the cut-off IC50 value; therefore, no information of those was included in Table 3.

3.3. Epitope conservancy analysis

Particular amino acid residues in viral proteins are critical for conducting essential functions, and these are residues least likely to evolve or vary, even under immune pressure. As such, these regions/epitopes are the desired target in epitope-based vaccine design. These regions are considered 'conserved' because they evolve slowly compared to 'variables' that evolve rapidly (Bui et al., 2007). All the sequenced proteins of SARS CoV-2 from different regions of the world were retrieved from NCBI Virus database and used to conduct the conservancy test. We have completed this conservancy test using the IEDB Epitope Conservancy Analysis tool to ensure that epitopes are present in viruses irrespective of disease state or strain of the virus. Surprisingly, 18 epitopes were found to be highly ($>$99.5%) conserved throughout the pandemic, irrespective of different survival factors, i.e. weather, humidity, region, temperature, etc. and two of the remaining three epitopes showed standard conservancy ($>$90%) where one epitope is below that value. The results are shown in Table 3.

Table 3. Potential T cell epitopes with predicted MHC restriction alleles and conservancy.

Protein ID	Epitope	**Interacting MHC-I allele with an affinity < 250 nm (Total score of proteasome score, Tap score, MHC score, Processing core and MHC-I binding) [MHC-I IC_{50}]**	Epitope conservancy (%)
YP_009724392.1	LTALRLCAY	HLA-B*15:25 (1.1) [38.9]	99.87%
		HLA-B*15:01 (1) [49.5]	
		HLA-C*12:03 (0.7) [97.2]	
		HLA-A*30:02 (0.67) [106]	
		HLA-A*29:02 (0.62) [118.9]	
		HLA-A*01:01 (0.6) [123.3]	
		HLA-B*15:02 (0.58) [129.3]	
		HLA-C*03:02 (0.47) [166.7]	
		HLA-A*26:01 (0.37) [211.8]	
		HLA-B*35:01 (0.36) [214.9]	
		HLA-C*16:01 (0.33) [229.7]	
	VSLVKPSFY	HLA-A*30:02 (0.84) [54.9]	100.00%
		HLA-A*29:02 (0.45) [135]	
	VFLLVTLAI	HLA-A*23:01 (−0.82) [216.2]	100.00%
YP_009724393.1	SSDNIALLV	HLA-C*05:01 (−0.2) [18.5]	100.00%
		HLA-C*15:02 (−0.65) [52.2]	
		HLA-A*01:01 (−0.87) [86.6]	
		HLA-C*08:01 (−1.28) [225]	
		HLA-C*08:02 (−1.31) [242.2]	
	ATSRTLSYY	HLA-A*30:02 (1.48) [13.3]	99.67%
		HLA-A*11:01 (1.08) [32.9]	
		HLA-A*01:01 (0.92) [48.2]	
		HLA-A*29:02 (0.84) [57.6]	
		HLA-B*15:25 (0.59) [102.4]	
		HLA-B*15:01 (0.33) [186.5]	
		HLA-A*03:01 (0.25) [221.3]	
		HLA-C*02:02 (0.21) [244.7]	
		HLA-C*02:09 (0.21) [244.7]	
	LAAVYRINW	HLA-B*58:01 (0.99) [8.6]	99.48%
		HLA-B*57:01 (0.59) [21.7]	
		HLA-B*53:01 (−0.16) [122.5]	
	YIIKLIFLW	HLA-B*58:01 (−0.11) [54.1]	100.00%
		HLA-A*32:01 (−0.71) [214.6]	
		HLA-A*23:01 (−0.72) [217.6]	
	YFIASFRLF	HLA-A*23:01 (1.74) [5.4]	100.00%
		HLA-C*14:02 (1.51) [9.2]	
		HLA-A*24:02 (1.48) [9.9]	
		HLA-A*29:02 (1.01) [28.8]	
		HLA-C*07:02 (0.68) [62.3]	
		HLA-C*03:02 (0.22) [180.2]	
		HLA-C*16:01 (0.19) [189.1]	
		HLA-C*12:03 (0.19) [191.9]	
		HLA-B*15:25 (0.17) [200.4]	
YP_009724394.1	KVSIWNLDY	HLA-A*30:02 (1.08) [28.8]	100.00%
		HLA-A*29:02 (0.89) [43.9]	
		HLA-A*11:01 (0.33) [161.1]	
		HLA-B*15:25 (0.27) [183.2]	
		HLA-A*03:01 (0.2) [215.5]	
YP_009724395.1	QELYSPIFL	HLA-B*40:01 (0.3) [38.8]	99.69%
		HLA-B*40:02 (0.18) [50.7]	
	VFITLCFTL	HLA-A*23:01 (0.36) [43.6]	100.00%
		HLA-A*24:02 (−0.21) [159.8]	
YP_009724396.1	QSCTQHQPY	***No alleles within range**	95.53%
	NYTVSCLPF	HLA-C*14:02 (0.5) [4.60]	67.32%
		HLA-A*23:01 (0.3) [28.11]	
		HLA-C*07:02 (0.7) [37.86]	
		HLA-C*03:03 (32) [45.67]	
		HLA-B*15:02 (3.4) [50.03]	
		HLA-C*12:03 (97) [103.09]	
		HLA-A*24:02 (0.3) [118.75]	
	EYHDVRVVL	HLA-C*14:02 (0.48) [56.4]	99.87%
	GSLVVRCSF	HLA-B*58:01 (0.1) [207.6]	99.94%
YP_009724397.2	LSPRWYFYY	HLA-A*29:02 (1.11) [15.2]	100.00%
		HLA-A*30:02 (0.59) [50.8]	
		HLA-A*01:01 (0.45) [70.4]	
		HLA-C*16:01 (0.05) [176.5]	

(*continued*)

Table 3. Continued.

Protein ID	Epitope	Interacting MHC-I allele with an affinity < 250 nm (Total score of proteasome score, Tap score, MHC score, Processing core and MHC-I binding) [MHC-I IC_{50}]	Epitope conservancy (%)
	ELIRQGTDY	HLA-B*15:02 (−0.03) [243.3]	99.94%
	SSPDDQIGY	*No alleles within range	100.00%
	KAYNVTQAF	HLA-C*03:02 (1.68) [6]	99.94%
		HLA-B*15:25 (1.54) [8.3]	
		HLA-C*12:03 (1.54) [8.4]	
		HLA-A*32:01 (1.48) [9.5]	
		HLA-C*16:01 (1.44) [10.6]	
		HLA-C*03:03 (1.18) [19.1]	
		HLA-C*03:04 (1.18) [19.1]	
		HLA-B*15:01 (1.12) [21.9]	
		HLA-C*14:02 (0.99) [29.7]	
		HLA-B*58:01 (0.98) [30.1]	
		HLA-C*12:02 (0.95) [32.3]	
		HLA-B*15:02 (0.81) [44.6]	
		HLA-B*35:01 (0.56) [79.8]	
		HLA-C*02:09 (0.53) [86.4]	
		HLA-C*02:02 (0.53) [86.4]	
		HLA-B*57:01 (0.46) [100.1]	
		HLA-C*15:02 (0.16) [198.5]	
	KKADETQAL	*No alleles within range	99.87%
YP_009725296.1	SLIDFYLCF	HLA-B*15:25 (0.97) [39.2]	100.00%
		HLA-B*15:01 (0.8) [57.2]	
		HLA-A*29:02 (0.59) [93.5]	
		HLA-A*23:01 (0.37) [154.9]	
		HLA-B*15:02 (0.16) [248.3]	

3.4. Prediction of population coverage

There are thousands of different human MHC (HLA) alleles present, and each type is expressed at a different frequency in different ethnicities of the world. Epitopes that bind to several different types of HLA alleles are considered the best candidate for vaccine design only when their combined frequency in a population group shows good coverage (near to 100%). Population coverage analysis for individual epitopes with the corresponding alleles was done for 21 different population groups using the IEDB Population coverage tool (Figure 2) with a combined analysis approach. The projected population coverage for the whole world was found at 96.12%. All other population group results showed that the respective epitopes have the potential to cover >90% of the population in the respective regions.

The coverage for China, Italy, Spain, and the USA, the most affected countries by SARS CoV-2, showed a favourable value of 97.48%, 96.80%, 96.88%, and 96.14%, respectively (Figure 3).

3.5. Allergenicity assessment

The prediction of allergenicity has become a necessary step in vaccine design in recent years because of the ability of antigens called allergens to elicit hypersensitive reactions such as rhinitis, asthma, and atopic eczema and more severe acute and fatal anaphylactic shock. The allergenicity of the antigenic proteins was evaluated by Algpred employing the Hybrid (SVMc + IgE epitope + ARPs BLAST + MAST) Approach (Table S1) (Sanchez-Trincado et al., 2017). For further confirmation about the allergenicity of the epitopes, Allercatpro (Maurer-Stroh et al., 2019) was used. The result is shown in Table 4, where eight epitopes are non-allergen among the 21 and l3 are predicted as a probable allergen. Interestingly, the best four epitopes based on MHC-I binding ability, three epitopes named KAYNVTQAF, LTALRLCAY, and ATSRTLSYY gave non allergen results while the rest one YFIASFRLF become a probable allergen.

3.6. Comparative modelling of the protein structure

Three-dimensional structure of the protein has a significant role in the understanding of protein dynamics and stability. Therefore, we modelled the HLA-A*29:02 protein structure by utilized Phyre-2 webserver to generate three-dimensional protein structure in an intensive mode. As a consequence, 273 residues (100%) have been modelled at >90% accuracy. Based on confidence level and resemblance, the final structure was selected from the twenty models generated by Phyre2. By building backbone, adding side chains, and loop modelling, this server generates the protein structure from a given protein sequence by using Hidden Markov Model (HMM) (Kelley et al., 2015). Further *ab initio* approach is utilized in the intensive mode to construct the backbone, missing region, and side-chain (Kelley et al., 2015). Afterwards, the modelled structure was validated via PROCHECK, ERRAT, VERIFY3D, PROSA II, and QMEAN 4 analysis (Figure 4).

PROCHECK analysis predicted that the hypothetical model obtained 89.6% residues in the favoured region; while 9.1% residues in the additional allowed region, 0.8% in generously allowed regions and 0.4% in disallowed regions. The plots demonstrate the phi(F)-psi(ψ) torsion angles for every residue

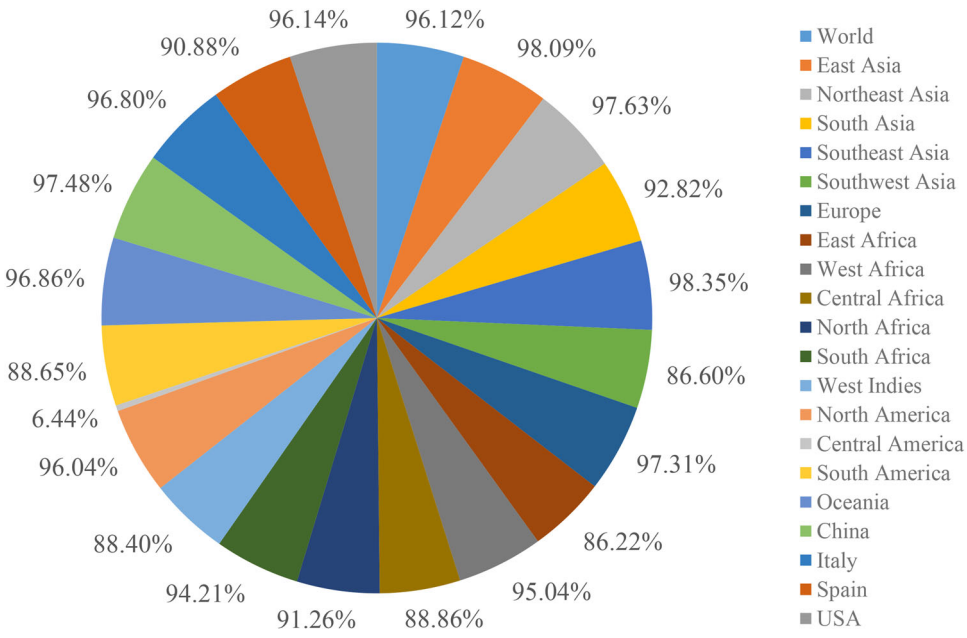


Figure 2. Population coverage of different regions and ethnicity.

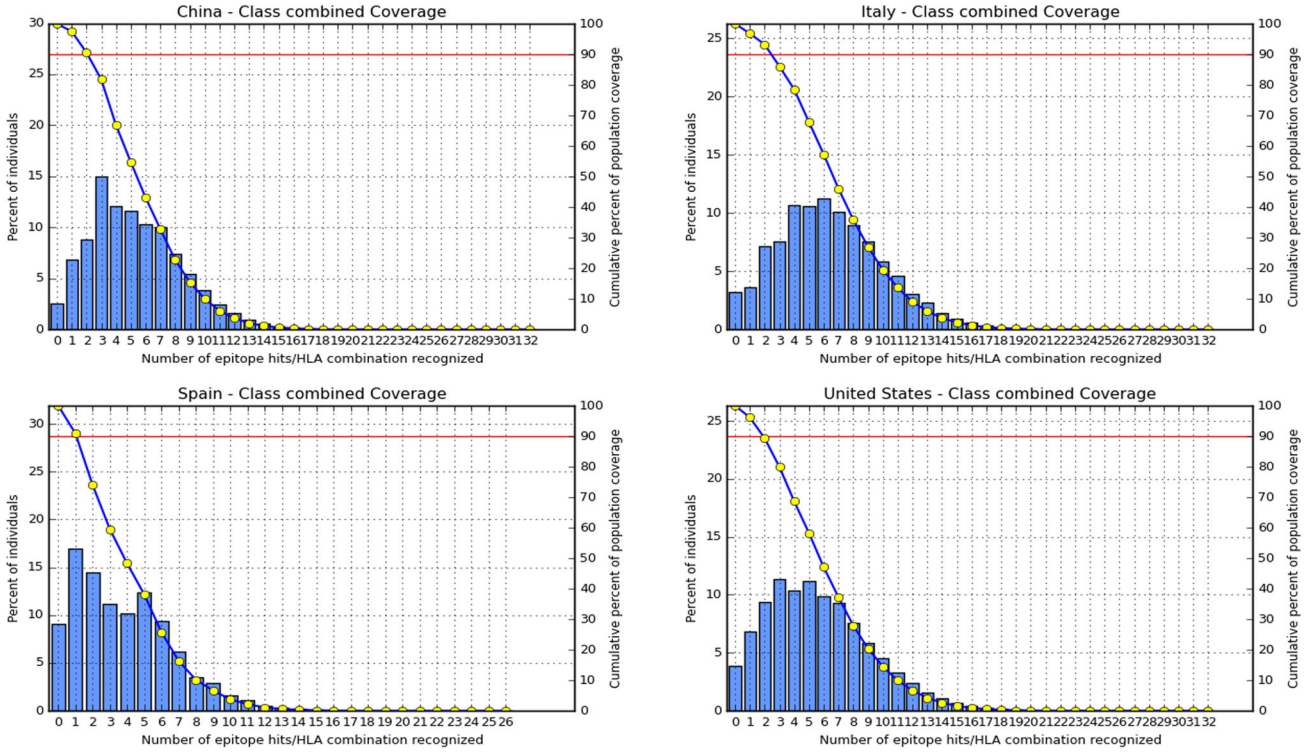


Figure 3. Population coverage results in most affected countries by SARS-CoV-2 based on MHC restriction data. Here the line (-0-) showed the cumulative percentage of population coverage of all the epitopes.

of a protein (Figure 4a). The Ramachandran plot generated by Rampage predicted that the final model had 96.7% residues in favoured regions, 3.3% in allowed regions and 0.0% in outlier regions (Figure 4b).

Though outstanding quality models are likely to have over 90% residues in the most favoured regions of PROCHECK's Ramachandran plot and around 98% in the favoured regions of Rampage's Ramachandran plot, the final model was considered of good quality since it scored 89.6% and 96.7% respectively.

A comparative analysis is done utilizing ProSA II web algorithm, which predicted that the final structure gained a Z-score of −8.94 and thus also falls within the range of scores established on similarly sized proteins, with an NMR quality (Figure 4c and 4d). ERRAT analysis estimated the score of 93.4615 for the final model, where a score of more than

Table 4. Allergenicity prediction result of the epitopes.

Protein ID	Epitope	Allergenicity (Allergen/Non-allergen)
YP_009724392.1	LTALRLCAY	Probable Non-allergen
	VSLVKPSFY	Probable Allergen
	VFLLVTLAI	Probable Allergen
YP_009724393.1	SSDNIALLV	Probable Allergen
	ATSRTLSYY	Probable Non-allergen
	LAAVYRINW	Probable Allergen
	YIIKLIFLW	Probable Allergen
	YFIASFRLF	Probable Allergen
YP_009724394.1	KVSIWNLDY	Probable Non-allergen
YP_009724395.1	QELYSPIFL	Probable Non-allergen
	VFITLCFTL	Probable Non-allergen
YP_009724396.1	QSCTQHQPY	Probable Non-allergen
	NYTVSCLPF	Probable Allergen
	EYHDVRVVL	Probable Allergen
	GSLVVRCSF	Probable Allergen
YP_009724397.2	LSPRWYFYY	Probable Allergen
	ELIRQGTDY	Probable Non-allergen
	SSPDDQIGY	Probable Allergen
	KAYNVTQAF	Probable Non-allergen
	KKADETQAL	Probable Allergen
YP_009725296.1	SLIDFYLCF	Probable Allergen

80.00 revealed the better quality (Monterrubio-López et al., 2015) (Figure 4e). Similarly, VERIFY3D graph (Figure 4f) predicted that 95.24% of residues of our examined model had an averaged 3D-1D score of 0.2, where a good model represents a cuff-of percentage of 80% (Daydé-Cazals et al., 2016). Assessment of protein model from QMEAN4 (Figure 4g and 4h) denoted a Z-score of −0.48, and the total score of 0.79 ± 0.05. The results confirmed the higher quality of the model, where the standard score ranges from 0 to 1 (Benkert et al., 2008).

3.7. Molecular docking analysis

After analyzing antigenicity, epitope conservancy, and allergenicity, the three best epitopes KAYNVTQAF, LTALRLCAY, and ATSRTLSYY from protein YP_009724397.2, YP_009724393.1, and YP_009724392.1 were chosen for molecular docking study against HLA-A*29:02 protein. Here we utilized AutoDock Vina software for performing molecular docking simulation to understand the binding patterns of the epitopes in HLA molecules environment. To understand the intermolecular interaction between small and macromolecule, molecular docking simulation is the most common platform for the structure-based drug design.

The outcomes of the docking analysis showed that the epitope ATSRTLSYY gave the highest binding affinity value (−11.3 kcal/mol) and the epitope LTALRLCAY showed the lowest value (−10.8 kcal/mol). In comparison, the epitope KAYNVTQAF displayed almost the similar binding affinities pattern (−11.2 kcal/mol) compare to ATSRTLSYY epitope with the HLA-A*29:02 protein. More negatives values of docking score indicate higher binding affinity (Junaid et al., 2018; 2019). Docked protein-ligand best pose complex was visualized using BIOVIA Discovery Studio as illustrated in Figure 5. As the ATSRTLSYY and KAYNVTQAF epitopes showed the highest binding affinity, it was subjected for subsequent molecular dynamics (MD) simulation, to investigate the dynamic stability, their mechanism of binding, the behavior of structural conformation of the HLA-A*29:02-epitope

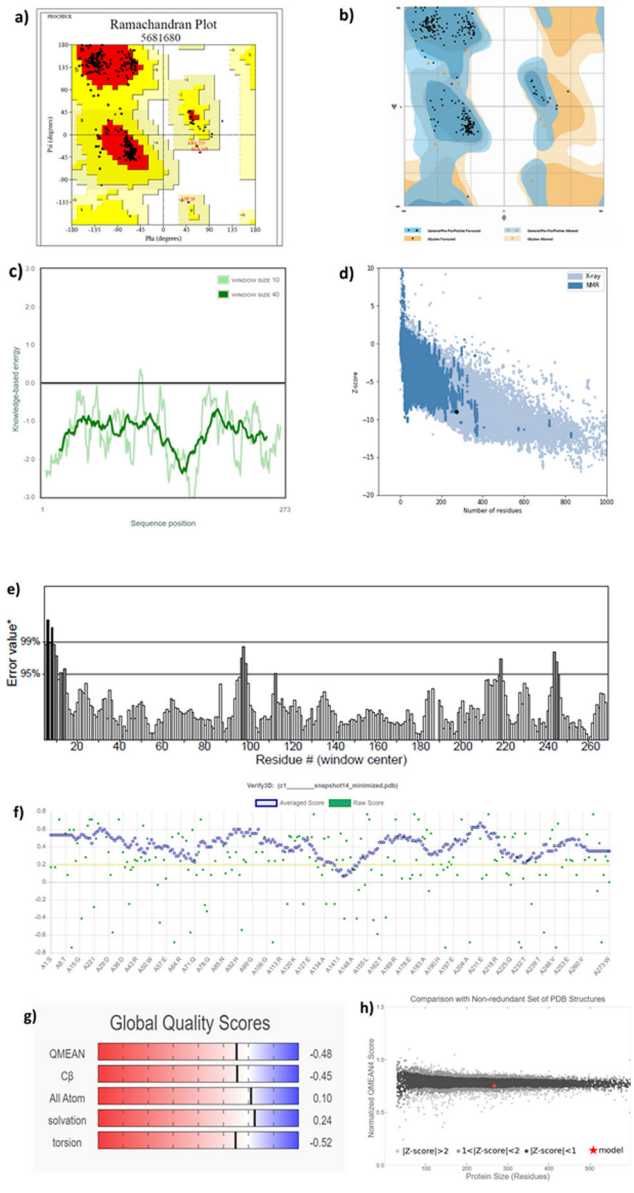


Figure 4. Validation of the in silico modelled X structure. The modelled structure was validated by PROCHECK, Rampage, Prosa II, ERRAT, Verify3D and QMEAN4. (a) PROCHECK's Ramachandran plot represents that 89.6% of residues located in most favoured regions, 9.1% in additional allowed regions, 0.8% in generously allowed regions and 0.4% in disallowed regions. (b) Rampage's Ramachandran plot shows 96.7% of residues in favoured regions, 3.3% in allowed regions and 0.0% in outlier regions. (c, d) Structure validation by ProSA web algorithm. The graph displays that the modelled structure's score, designated by the black dot, falls within the range of scores identified on similarly sized proteins, with an NMR quality. (e) According to ERRAT, the structure achieved 93.4615% overall quality factor. (f) VERIFY3D plot of the model of X protein. (g, h) Structure validation by Qmean, which shows the quality of a structure while compared to a non-redundant set of PDB structures of the same size. The image shows that the modelled structure's score, denoted by the red 'X', falls within the range of scores of reference structures of the same size, and it is therefore of good quality.

complexes. The apo form of HLA-A*29:02 protein was also underlying for MD along with epitope complexes.

3.8. Molecular dynamics simulation analysis

We performed MD simulations study to validate the findings from molecular docking. The 50 ns MD simulations were carried out to realize the perturbations at the atomic level in

both HLA (protein) and epitopes (ligand) structure of HLA-epitope complex.

3.8.1. Root mean square deviation (RMSD)

The RMSD of the atoms of protein backbone regarded as an important parameter for the evaluation of equilibrium and stabilization of MD trajectories.

We calculated the RMSD pattern of protein and ligand from the protein-ligand complex and plotted in Figure 6. The structural stability of the Apo form of HLA (indicated by red colour) protein and in complex with two epitopes were calculated in terms of RMSD and depicted in Figure 6(a). The results demonstrated that the Apo form of HLA remained stable from 0 to 34 ns showing an average 3 Å RMSD value and revealing some small fluctuations in its backbone structure. After 34 ns, it showed a more substantial fluctuation from 2 Å to 12 Å and again decreased to 10 Å at the end of the simulation. The HLA-A*29:02 (indicated by green) of HLA-A*29:02_KAYNVTQAF complex also showed a similar range of stability like Apo representing average 4 Å RMSD from 0 to 30 ns. After 30 ns, it showed a significant fluctuation from 2 Å to 14 Å and decreased to 13 Å at the end of the simulation. On the other hand, the HLA-A*29:02 (indicated by blue) of HLA-A*29:02_ATSRTLSYY complex remained stable throughout the simulation time, representing an RMSD value of 2 Å with very less fluctuation.

From Figure 6(b), it can be shown that epitope ATSRTLSYY (indicated by blue) showed a similar pattern of RMSD found for HLA-A*29:02 of HLA-A*29:02_ATSRTLSYY (blue). On the other hand, the epitope KAYNVTQAF (green) of HLA-A*29:02_KAYNVTQAF complex was seen to fluctuate in greater magnitude while compared to the RMSD pattern of HLA-A*29:02 (indicated by green) of HLA-A*29:02_KAYNVTQAF complex. From this analysis, it can be inferred that upon the binding of ATSRTLSYY epitope to the HLA-A*29:02, there was no change in the stability of HLA.

3.8.2. Radius of gyration (Rg)

We also calculated the Rg value for both HLA and epitopes (Figure 7) to analyze the impact of epitope binding to HLA and to predict the alteration in terms of compactness (Kamaraj et al., 2015). It was demonstrated that a stably folded protein maintains a stable Rg value and lower Rg value denotes good compactness. During protein unfolding, Rg value changes with time.

From Figure 7(a), it can be depicted that the Rg value for Apo form of HLA increased from 22.5 Å to 30.5 Å and remained stable in the range of 24 Å to 25 Å during 5 to 35 ns simulation time. The Rg value for HLA-A*29:02 of HLA-A*29:02_KAYNVTQAF complex also showed a gradual increase in Rg value from 23 Å to 29 Å. Though, from 5 ns to 27 ns, it showed a 24 Å of Rg value and remained in an equilibrium state. The HLA-A*29:02 from HLA-A*29:02_ATSRTLSYY complex showed 23 Å of Rg value and remained in equilibrium state except for a little fluctuation from 15 to 22 ns. Hence, by revealing lower Rg value, it indicated the better compactness and higher stability for HLA-A*29:02 upon ATSRTLSYY epitope binding.

From Figure 7(b), it can be seen that ATSRTLSYY epitope (blue) from HLA-A*29:02_ATSRTLSYY complex showed a similar pattern of Rg found for HLA-A*29:02 of HLA-A*29:02_ATSRTLSYY complex without any fluctuation in its structure throughout the simulation, that revealed the healthy binding pattern between HLA-A*29:02 and ATSRTLSYY epitope.

3.8.3. Solvent accessible surface area (SASA)

We performed SASA analysis which demonstrated at Figure 8, to assess the protein surface area that is easily accessible to the solvent through estimating the hydrophilic and hydrophobic residues of a protein (Kamaraj et al., 2015), where the higher value of SASA indicates relative expansion. Any alteration in SASA reveals a deviation in protein structure and hence the function of the protein (Doss & Rajith, 2013). Therefore, both radii of gyration and solvent accessible surface area were calculated for the assessment of protein packing and conservation.

Usually, hydrophobic residues present in protein structure mostly responsible for the increase of SASA value. The SASA values for Apo form of HLA as well as each of the two epitope-HLA complexes were calculated and depicted in Figure 8(a). The average SASA values for Apo-HLA-A*29:02, HLA-A*29:02_ATSRTLSYY (indicated by blue), and HLA-A*29:02_KAYNVTQAF (green) were 17,000 $Å^2$, 16,500 $Å^2$ and 16,800 $Å^2$, respectively (Figure 8a). The average SASA value showed that the HLA-A*29:02_ATSRTLSYY showed less value as compared to HLA-A*29:02 and HLA-A*29:02_KAYNVTQAF complex. From the SASA values, we have concluded that the binding of ATSRTLSYY epitope to Apo-HLA-A*29:02 induced conformational stability and better compactness during binding to Apo-HLA-A*29:02.

Also, the SASA values of residues are a vital parameter to comprehend the conformational changes based on residues. So we have plotted a graph based on residues versus SASA value in Figure 8(b). The average SASA values for Apo-HLA-A*29:02 (red), HLA-A*29:02_ATSRTLSYY (indicated by blue), and HLA-A*29:02_KAYNVTQAF (green) were 100, 150, and 140 $Å^2$, respectively showing higher SASA value for HLA-A*29:02_ATSRTLSYY which represents that our predicted ATSRTLSYY epitope leads to the stability after ligand binding.

3.8.4. Hydrogen bonding analysis

Hydrogen bonds play a significant role in the prediction of protein-ligand stability as they are transient non-bonded interactions between complexes and responsible for protein stability. In this study (Figure 9), we have estimated the number of hydrogen bonds for HLA before and upon epitope binding (Figure 9a), for epitopes upon binding to HLA (Figure 9b), and for combined HLA-epitope complexes (Figure 9c). The average number of hydrogen bonds for HLA in three conditions (Apo, with A*29:02_KAYNVTQAF, and A*29:02_ATSRTLSYY) (Figure 9a) were found 82, 83, and 82, respectively. Similarly, for two epitopes KAYNVTQAF (green), and ATSRTLSYY (blue) (Figure 9b), the average number of hydrogen bonds were 6 and 20, respectively. Again, in the complex of HLA-A*29:02_KAYNVTQAF (green) and HLA-

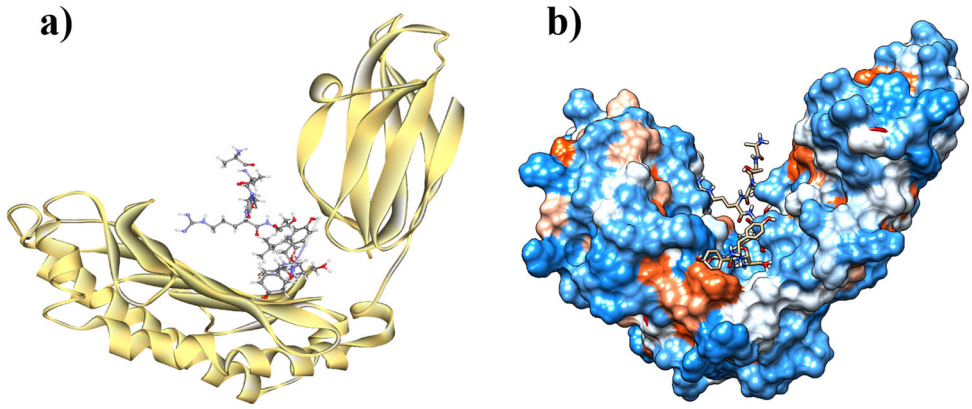


Figure 5. Predicted pose from the docking analysis showed the binding orientation map of HLA-A*29:02_ATSRTLSYY in (a) normal and (b) solid surface.

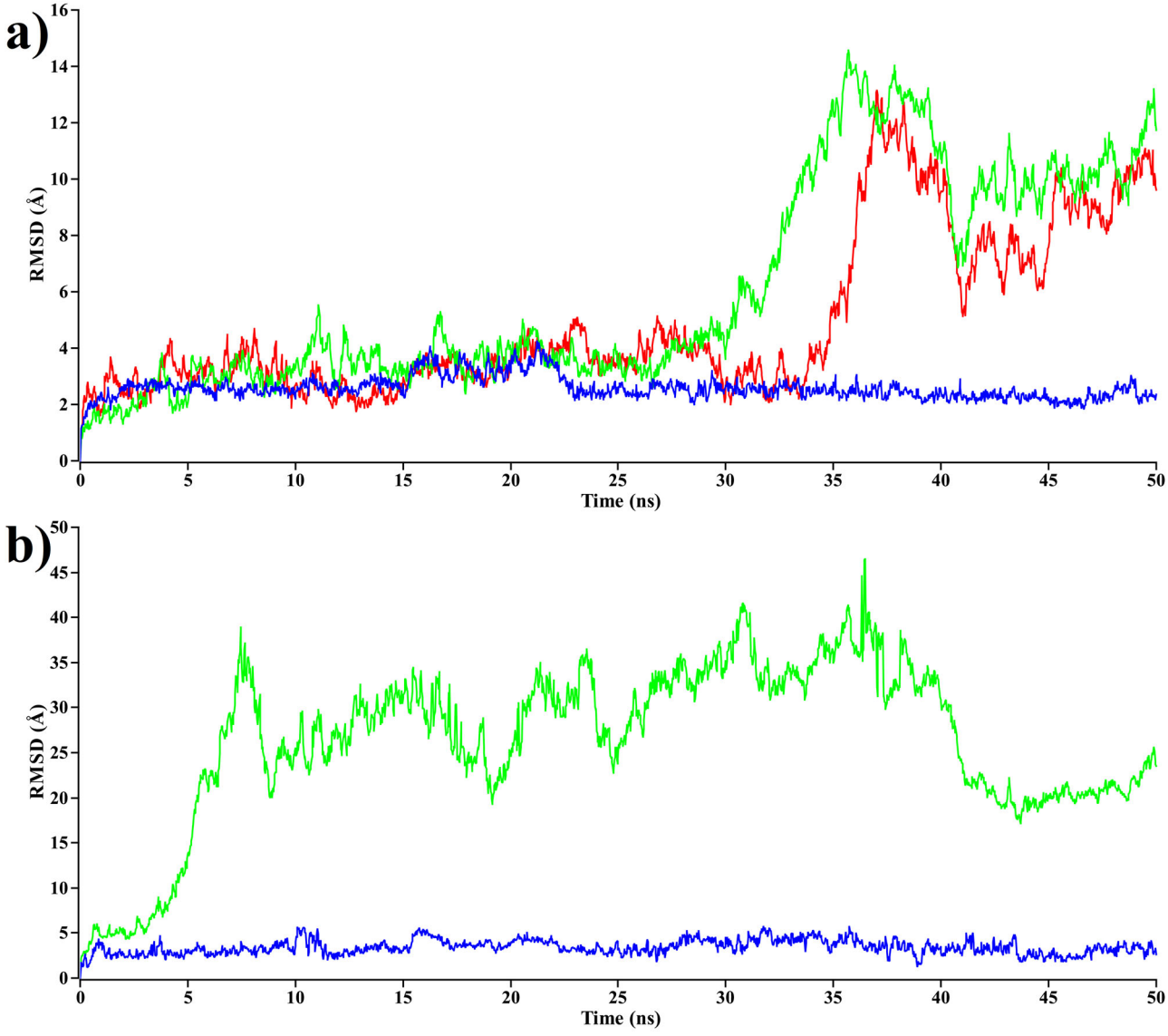


Figure 6. The time series of the RMSD of backbone atoms (C, Cα, and N) for (a) protein, and (b) ligand. Here, red, green, and blue lines denote HLA-A*29:02_Apo, HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY complex respectively.

A*29:02_ATSRTLSYY (blue) (Figure 9c), the average number of hydrogen bond was 6 and 11. From the above analysis, it can be concluded that epitope ATSRTLSYY made a higher number of hydrogen bonds compared to epitope KAYNVTQAF during binding with HLA-A*29:02, revealing the excellent stability of the HLA-A*29:02_ATSRTLSYY complex.

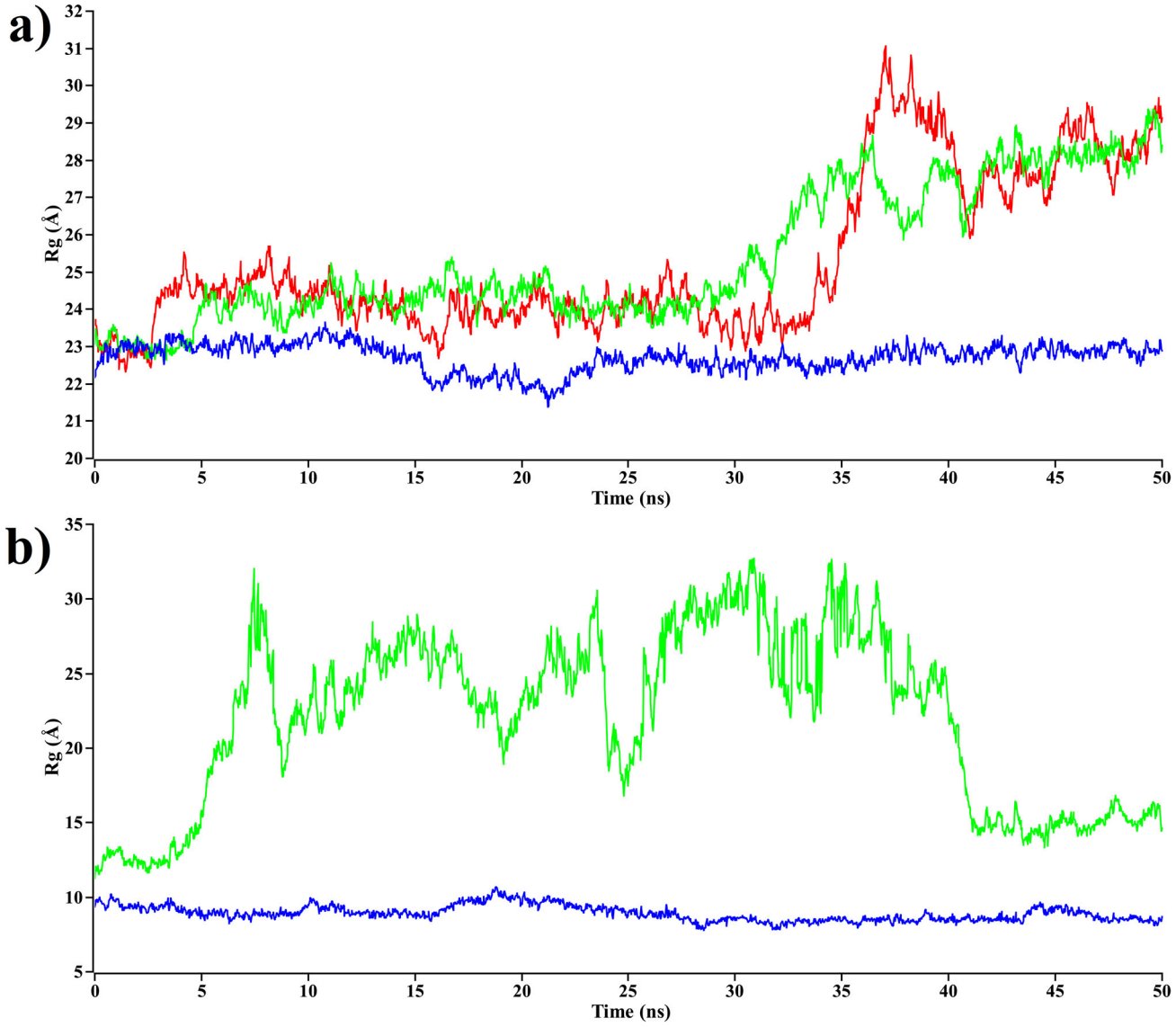


Figure 7. The structural changes of (a) protein, and (b) ligand through the radius of gyration analysis. Here, red, green, and blue lines denote HLA-A*29:02_Apo, HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY complex respectively.

3.8.5. Root mean square fluctuations (RMSF)

Besides RMSD analysis, we also measured the level of fluctuation at the atomic level (Figure 10) for Apo-HLA-A*29:02, and HLA-A*29:02 upon KAYNVTQAF and ATSRTLSYY binding. Figure 10 illustrated that the average RMSF value for Apo-HLA-A*29:02, HLA-A*29:02 of HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02 of HLA-A*29:02_ATSRTLSYY were 6.5, 7.5 and 2.5 Å, respectively. This analysis revealed that binding of epitope ATSRTLSYY to HLA-A*29:02 reduced the fluctuation and induced compactness in terms of stability while compared to Apo-HLA-A*29:02 protein backbone.

3.9. MM-PBSA binding free energy analysis

Poisson-Boltzmann surface area (PBSA) is one of the most appealing solvation systems in computer-aided drug designing methods and widely used to calculate the binding energy of protein-ligand complexes (Wang et al., 2010). The analysis of binding free energy is essential for the evaluation of the binding capacity of the ligands towards its receptor, as it estimates the quantitative value for binding affinity calculation.

Therefore, we subjected each HLA-epitope (protein-ligand) complex to the MM-PBSA binding energy calculation to identify how structural changes influence epitope binding. For this purpose, we measured the binding energy of each snapshot generated from MD simulations. The results are depicted in Figure 11, where more positive energies signified good binding, but negative values of energies do not mean any binding, instead these two terms used for comparative binding analysis between ligands according to the theory of nuclear physics (Blatt & Weisskopf, 1991).

This theory stated that the energy necessary for disassembling of a whole into separate parts is usually positive (Lovering et al., 2005). On average, the binding free energy of HLA-A*29:02_ATSRTLSYY (blue), and HLA-A*29:02_KAYNVTQAF (green) complexes were −46,000 kJ/mol and −50,000 kJ/mol, respectively. From the graph, it can be seen that there is a significant difference between the

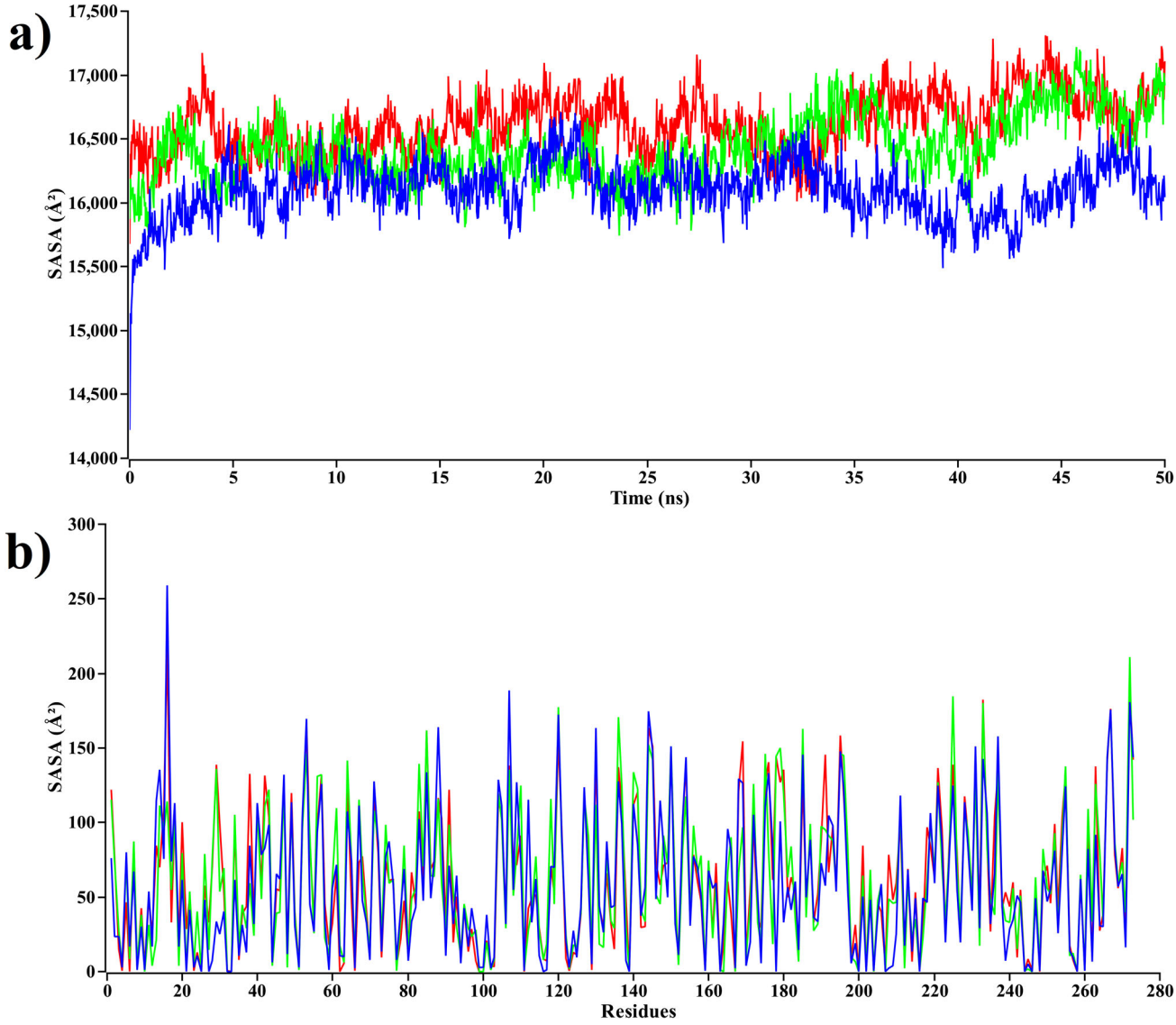


Figure 8. The structural changes of protein utilizing (a) solvent accessible surface area (SASA), and (b) residue-SASA analysis. Here, red, green, and blue lines denote HLA-A*29:02_Apo, HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY complex respectively.

binding energy of two complexes and both maintained an equilibrated nature throughout the simulation without much fluctuation or overlap on each other. Thus, from the analysis, we have found that HLA-A*29:02_ATSRTLSYY complex revealed higher binding energy compared to HLA-A*29:02_KAYNVTQAF complex, representing the better binding affinity and stable complex formation for ATSRTLSYY epitope.

3.10. Secondary structure element analysis

We investigated the deviations in secondary structure elements for both HLA-Apo and HLA-epitopes complexes at position level by using VMD and plotted in Figure 12. From the analysis, it was found that there were no significant changes in both forms of HLA, revealing the proper binding of epitope without causing structural changes compared to HLA-A*29:02_Apo, which illustrated at Figure 12(a). Only at position 177, during 0 to 15 ns simulation time, there was the appearance of overlapping turn and 3-helix, as well as loss of 3-helix from 15 ns to the end of simulation, were found for HLA-A*29:02_KAYNVTQAF (Figure 12b) while compared to HLA-A*29:02 _Apo (Figure 12a). On that particular position, HLA-A*29:02_ATSRTLSYY (Figure 12c) showed quite similar patterns compared to HLA-A*29:02_Apo. Also, at residue position 17, there was the loss of turn found for HLA-A*29:02_ATSRTLSYY from 17 ns to the end of simulation which may be contributed to its fluctuation on that particular position shown in RMSF plot.

β-turns represent the most remarkable structures of local protein alongside the α-helices and the β-strands. Since they play a role in the orientation of α-helices and β-strands, they exert a critical role in the final protein topology (Bornot & de Brevern, 2006). The secondary element analysis data revealed the quite same pattern of secondary structure components for HLA-A*29:02_Apo and HLA-A*29:02_KAYNVTQAF. Furthermore, some differences in HLA-A*29:02_ATSRTLSYY was observed in some particular positions which revealed its different nature of binding compared to HLA-A*29:02_Apo and HLA-A*29:02_KAYNVTQAF and may contribute in its different level of stable nature.

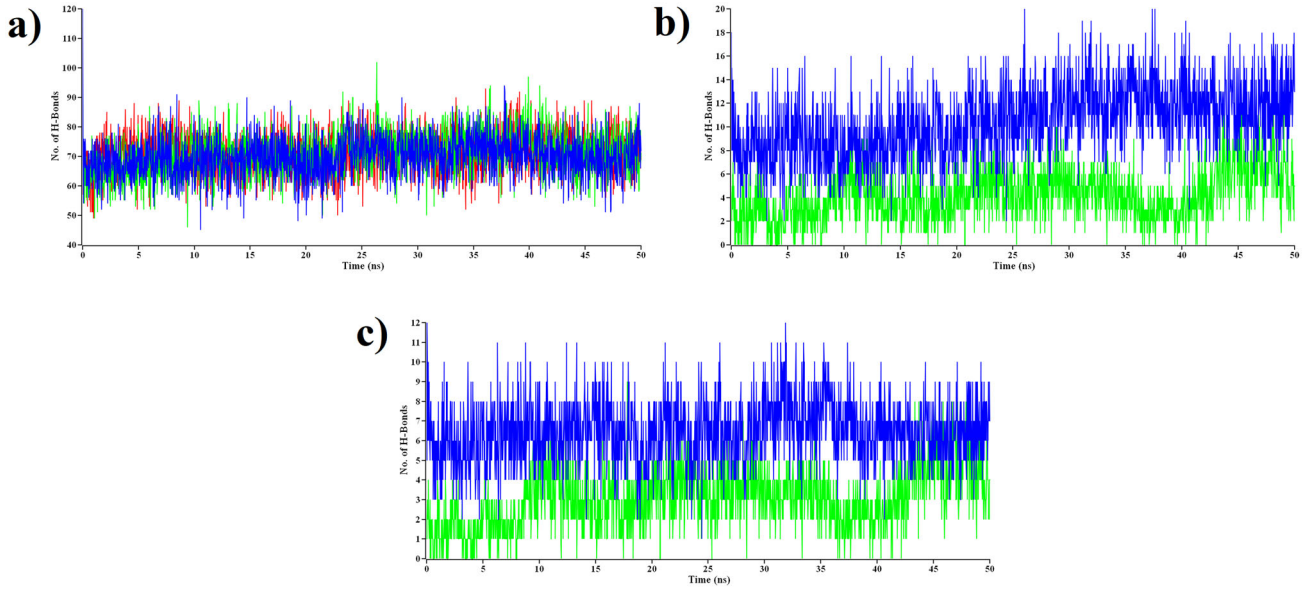


Figure 9. The total number of hydrogen bonds formed in (a) protein, (b) ligand and (c) between the protein and ligand in the complex state during the simulation. Here, red, green, and blue lines denote HLA-A*29:02_Apo, HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY complex respectively.

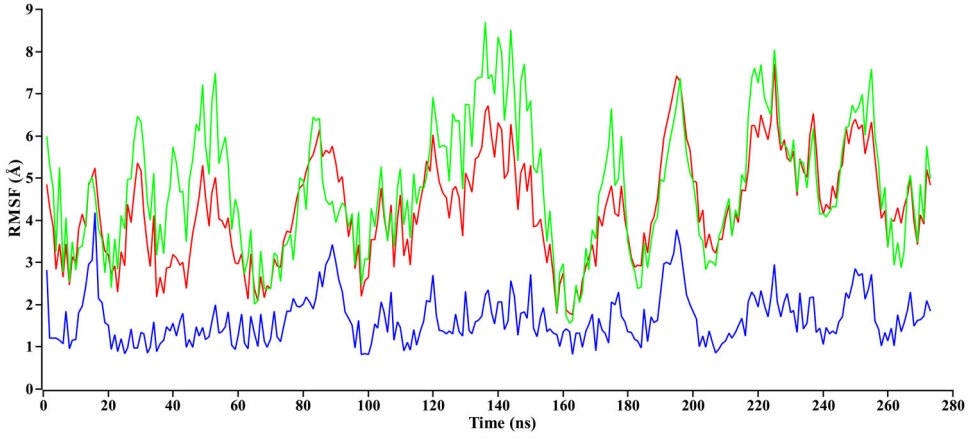


Figure 10. The structural changes of protein using root means square fluctuations (RMSF) analysis. Here, red, green, and blue lines denote HLA-A*29:02_Apo, HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY complex respectively.

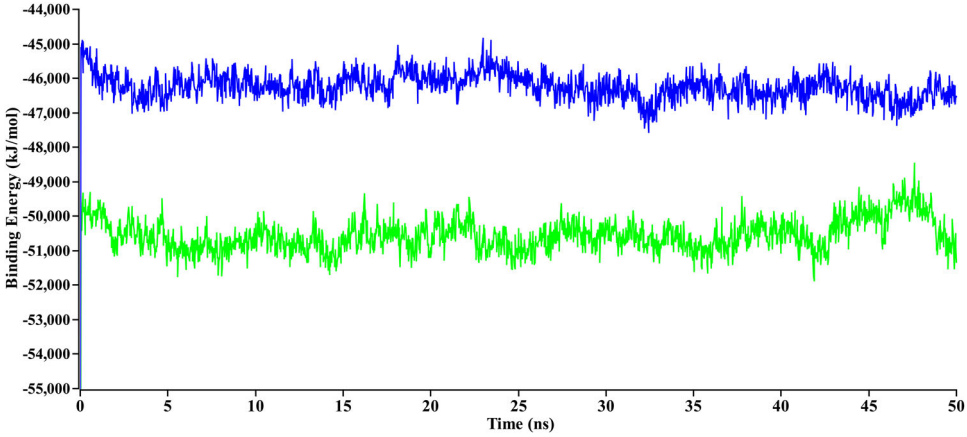


Figure 11. Binding free energy (in kJ mol^{-1}) of each snapshot was calculated by molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) analysis, representing the change in binding stability of each HLA-epitope complex during simulations. Here, green, and blue lines denote HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY complex, respectively.

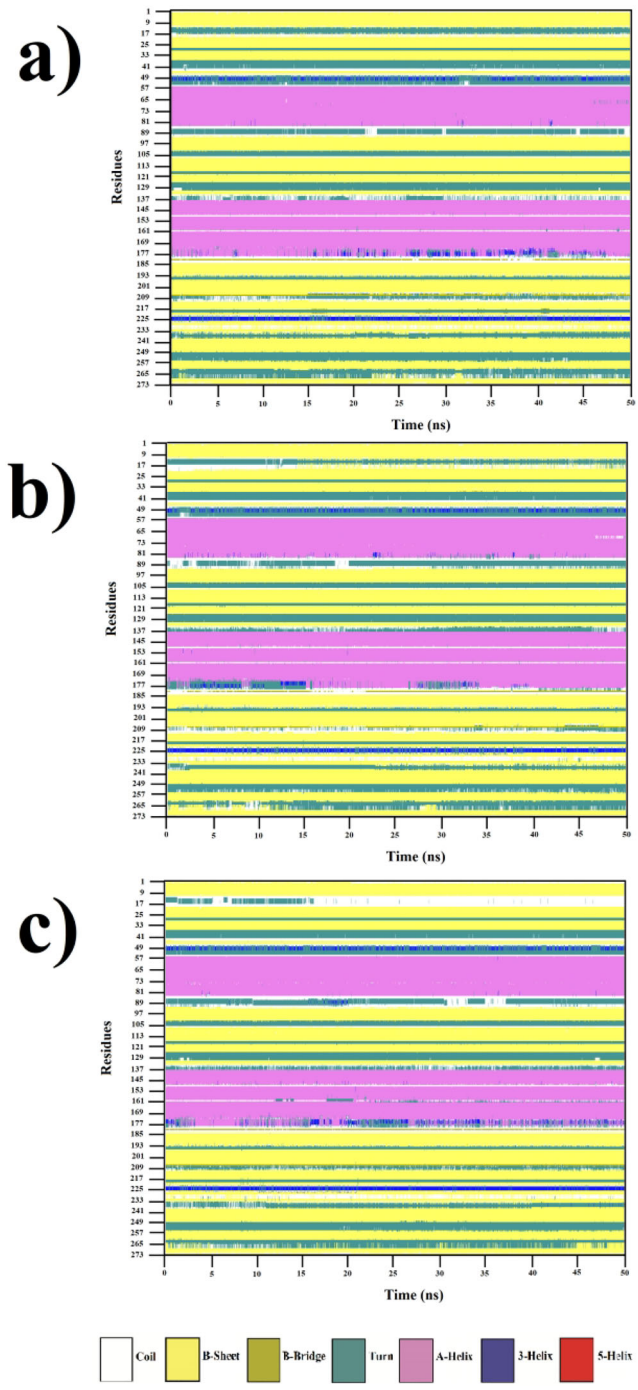


Figure 12. Secondary structural elements (SSE) of HLA protein were calculated during the 50 ns simulation. Here, three secondary structure plot in the figure alphabetically denote HLA-A*29:02_Apo, HLA-A*29:02_KAYNVTQAF and HLA-A*29:02_ATSRTLSYY complex, respectively. Here, white, yellow, dark yellow, teal, pink, blue, and red colour represents the formation of secondary structure including coils, β-sheet, β-bridge, turn, α-helix, 3-helix, and π-helix or 5-helix, respectively.

3.11. Principal component analysis (PCA) of molecular dynamics

Principal component analysis (PCA) is utilized to evaluate the structural and energy data retrieved from MD simulation on protein-ligand complexes and apo-protein. Bond angle energies, bond energies, planarity energies, dihedral angle energies, electrostatic energies and Van der Waals energies were included as variables. The PCA score plot (Figure 13) discloses three cluster formations. The three clusters did not overlap each other (Figure 13a). The clusters representing the Apo-HLA and HLA-A*29:02_ATSRTLSYY located almost in same score plot and same plane while a vast difference in the planarity of HLA-A*29:02_KAYNVTQAF was observed. Also, the energy distribution of HLA-A*29:02_ATSRTLSYY was condensed due to ligand binding while compared with Apo-HLA. Moreover, the broader distribution of energy was found for HLA-A*29:02_KAYNVTQAF complex compared to Apo-HLA. On the other hand, the loading plot of the PCA (Figure 13b) exhibits that the dihedral angle energies, RMSD, and planarity show a positive correlation with these three groups. The less fluctuating nature of HLA-

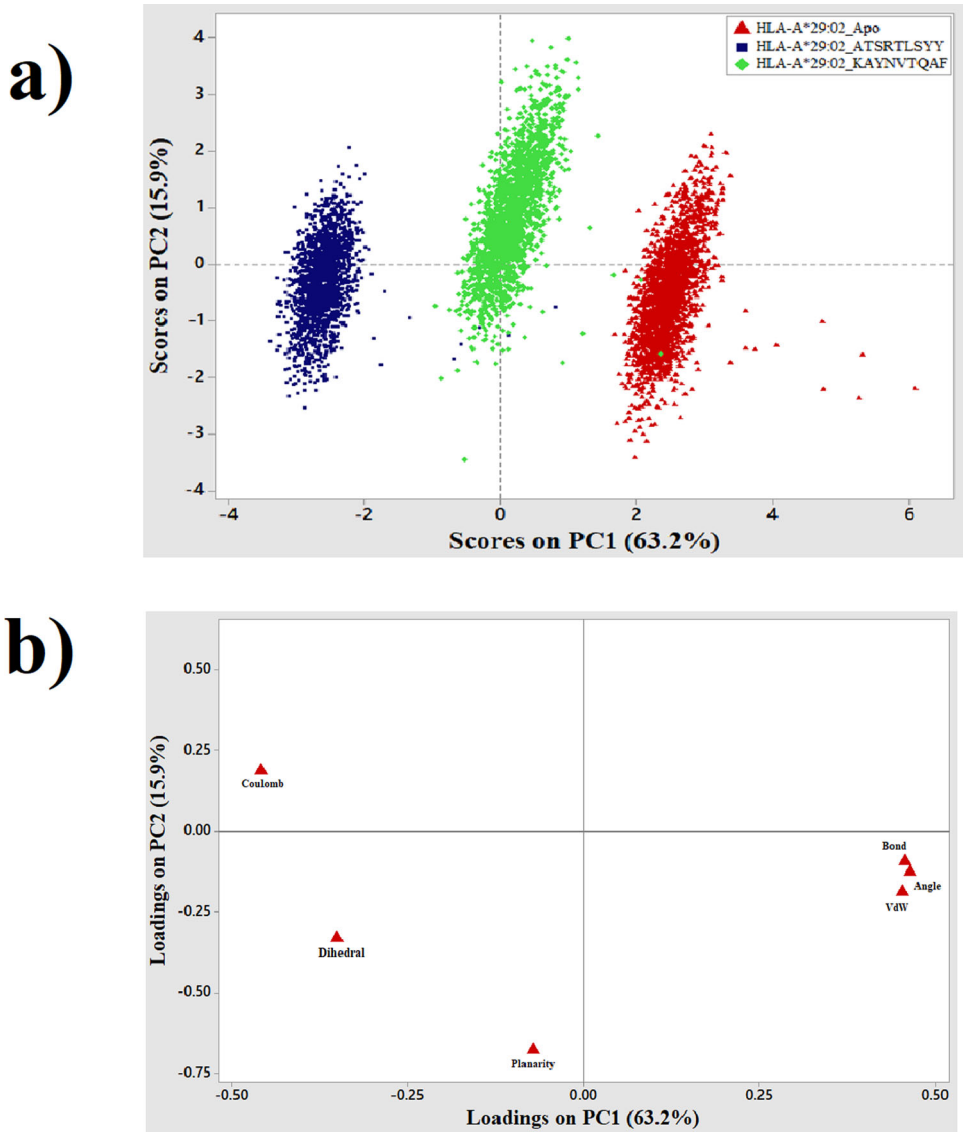


Figure 13. (a) The score plot presented two data clusters in different colours, where each dot represents a one-time point. The clustering is attributable as apo-protein (red), HLA-A*29:02_KAYNVTQAF (green), A*29:02_ATSRTLSYY (blue) (b) Loading plot from principal components analysis of the energy and structural data.

A*29:02_ATSRTLSYY was also proved from the PCA analysis, which revealed its lesser distribution in the score plot, and epitope KAYNVTQAF showed the highest deviation from the same plane upon ligand binding.

On the other hand, using the second technique of PCA (described in method section), we evaluated the direction and degree of changes in the collective motion of HLA protein upon epitope binding (Ndagi et al., 2017; Yesudhas et al., 2016).

The systems were projected along the direction of the first three principal components or eigenvectors (PC1, PC2 and PC3), obtaining the PCA scatter plots for the HLA system and HLA-epitope system, as shown in Figure 14. The PCA scatter plots displayed two conformational states on the subspace, where the blue dots meant the unstable conformational state and the red dots meant the stable conformational state. The intermediate state between the two conformations was indicated by white dots. The protein systems switched between two conformational states (blue and red) with periodic jumps. In the scatter plots, the differences in the motion between the three systems indicated that the conformational state of the HLA-A*29:02_Apo (red) protein due to ATSRTLSYY and KAYNVTQAF epitopes binding show significant changes in its correlated motion and compactness (Figure 14a).

The HLA-A*29:02_KAYNVTQAF (Figure 14b) scatters showed irregular distribution between unstable and stable state to each side of the diagonal while compared to HLA-A*29:02_Apo and HLA-A*29:02_ATSRTLSYY system (Figure 14c). Conversely, compared to both HLA-A*29:02_Apo system and HLA-A*29:02_KAYNVTQAF (Figure 14b) system, the PC1 and PC2 of HLA-A*29:02_ATSRTLSYY (Figure 14c) scatters were distributed more centrally to each side of the diagonal, indicating that the conformational state of HLA-A*29:02_ATSRTLSYY system was more stable than before.

In 50 ns simulation trajectories, the top 20 PCs of the HLA-A*29:02_Apo system, HLA-A*29:02_KAYNVTQAF, and HLA-A*29:02_ATSRTLSYY system accounted for 78.6% and 78%, and 44.8% of the total variation, respectively. It was observed that the first few eigenvectors contribute mostly to elucidate the

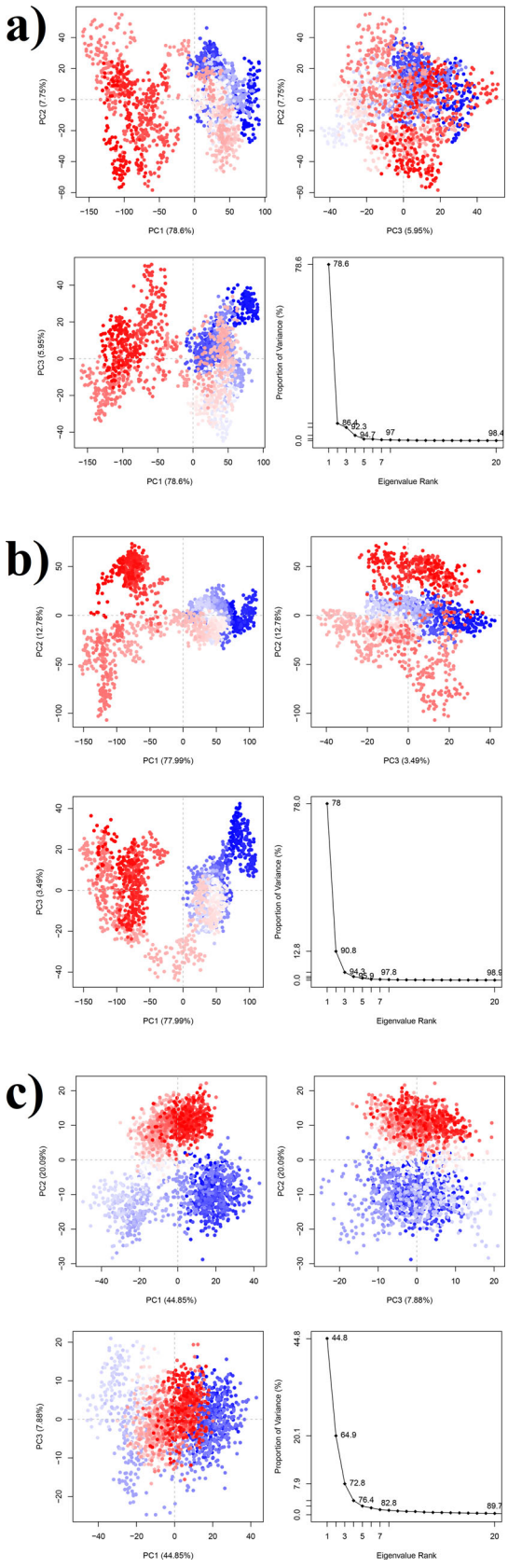


Figure 14. Principle Component Analysis (PCA) regarding the HLA protein in three different systems, (a) HLA-A*29:02_Apo, (b) HLA-A*29:02_KAYNVTQAF, and (c) HLA-A*29:02_ATSRTLSYY. Each panel represents the two-dimensional plots between eigenvectors (EV) 1, 2, and 3 in three cross profiles, PC1–PC2, PC1–PC3 and PC2–PC3 for all systems. Throughout the x and y axes, each dot denotes the one conformation of the complex. The spread of blue and red colour dots described the degree of conformational changes in the simulation, where the colour scale from blue to white to red is equivalent to simulation time. The blue indicates initial time step, and white is intermediate and final time step is represented by red colour.

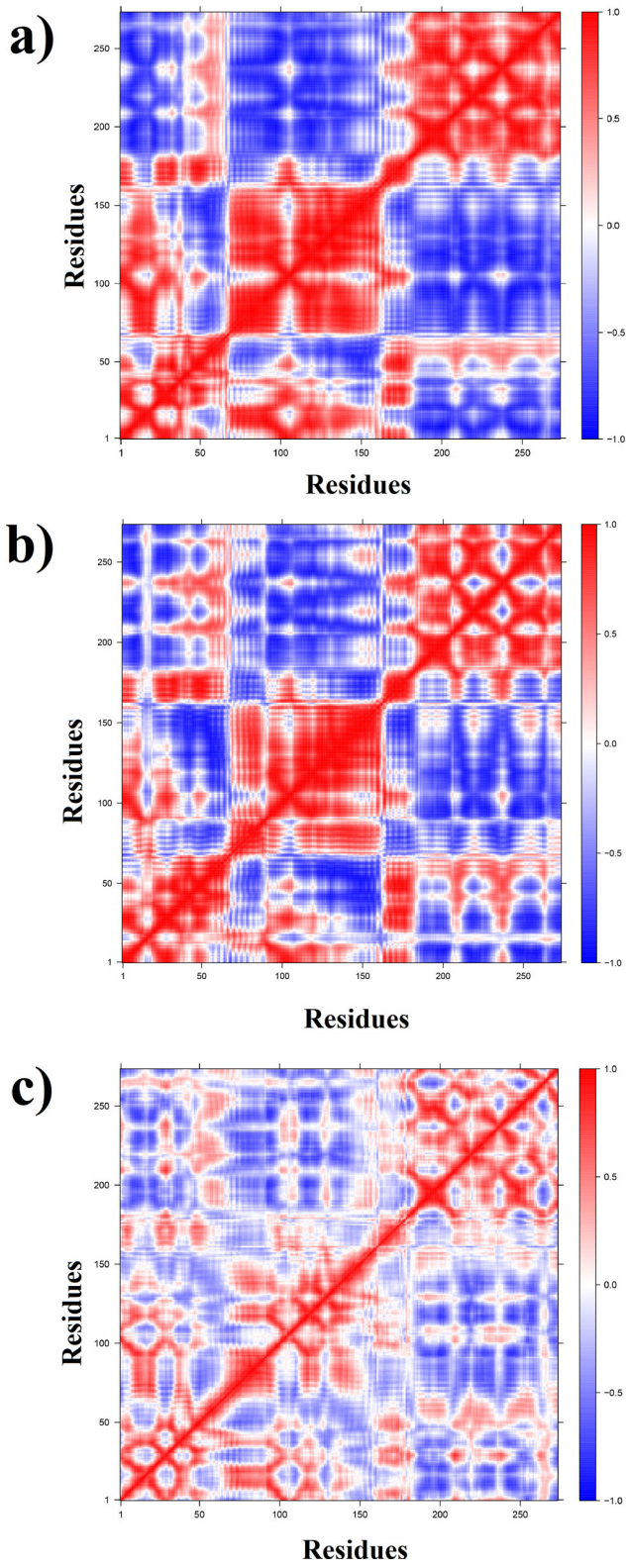


Figure 15. Calculated dynamic cross-correlation matrix of Cα atoms around their mean positions for 50 ns molecular dynamics simulations. Extents of correlated motions and anti-correlated motions are colour-coded from red to blue, which represent positive and negative correlations, respectively. (a) Apo-HLA-A*29:02; (b) HLA-A*29:02_KAYNVTQAF, and (c) HLA-A*29:02_ATSRTLSYY.

overall dynamics of the protein. Hence, only the first two eigenvectors were taken into account to describe the correlated motion. In HLA-A*29:02_Apo system (Figure 14a), the contributions of the first two PCs (PC1, PC2) to the variance were 78.6%,

and 5.9%, respectively. In HLA-A*29:02_KAYNVTQAF (Figure 14b) system, the contributions of the first two PCs to the variance were 78%, and 12.8%, respectively. In HLA-A*29:02_ATSRTLSYY (Figure 14c) system, the contributions of the first two PCs to the variance were 44.8% and 20.1%, respectively.

It could be seen that HLA-epitope system occupied higher phase space and exhibited higher flexibility than HLA-A*29:02_Apo system. The HLA-A*29:02_ATSRTLSYY showed relatively lower phase space while compared to both HLA-A*29:02_Apo and HLA-A*29:02_KAYNVTQAF system which was consistent with the conclusion of the RMSF, Rg, and the number of hydrogen bond analysis.

3.12. Dynamic cross-correlation matrix (DCCM) analysis

Motions of the global domain in a protein involve the collective movement of Cα backbone atoms and affects the shifting of protein conformations from one functional state to another. To investigate this displacement patterns, we have created and examined a dynamic cross-correlation matrix (DCCM). It represents the collective fluctuation of Cα atoms and unveils the relationships between residues and domains by quantifying their relative motions. Positive values indicate a positive correlation, meaning residues displace in the same direction (parallel direction), while negative values define negative correlations accounting for the opposite displacement of the residues (anti-parallel direction). Figure 15 illustrated the DCCM for (a) Apo form of HLA-A*29, (b) HLA-A*29 in a complex with KAYNVTQAF, and (c) HLA-A*29 in a complex with ATSRTLSYY.

The red colour represents positive correlation, the white colour represents no correlation, and local displacement and the blue colour represent the negative correlation. The depth of the colour correlates with stronger positive or negative correlation, i.e. it signifies the degree of correlation. The result for HLA-A*29:02_Apo (Figure 15a) discloses a robust negative correlation in the overall conformation, which has been significantly reduced and replaced by slight positive or no correlation at all, in the complexed form of HLA-A*29:02_KAYNVTQAF in Figure 15(b). This means the overall movement of residues which was in the opposite direction in the ligand unbound system of (Figure 15a) showed weaker correlation/intensity in its movement and places vanished altogether. This implies more significant relativity between this residue where the degree of negative correlation decreased, accounting for a more stable structure in Figure 15(b). The intensity of positively correlated motion also decreased when ligand bind with the protein (Figure 15b), meaning that the residues moved in the same direction together. This signifies that the overall conformation became slightly larger to accommodate the ligand in its binding site.

Conversely, while compared to the Apo-HLA-A*29:02, HLA-A*29:02_ATSRTLSYY (Figure 15c) complex showed more positive correlation upon epitope binding signifies by the reduced and faded blue colour and appearance of more red colour in the plot. Again, the overall conformation of HLA-A*29:02_ATSRTLSYY became compact occupying smaller spaces while compared to the overall conformation of HLA-A*29:02_KAYNVTQAF, revealing more non-bond interactions formed between HLA and epitope. Thus the intensity of positively correlated motion increased when ATSRTLSYY bind with the HLA-A*29:02.

The greater extent of positive correlative motions specified the better stability of the protein, as the conformational freedom is reduced in the 3D spaces, which induce the development of strong intramolecular interactions within the protein or protein-ligand complex. Additionally, enzyme dynamics is intrinsically linked with protein activity and stability, and residues with correlated motions are expected to be correlated functionally (Hosen et al., 2019).

3.13. B cell epitope identification

B-cell epitopes are essential for B-cell directed immune response in viral infections. We predicted B-cell epitopes within each antigenic protein of SARS-CoV-2 using the IEDB B-cell epitope prediction tool, employing all the methods provided in the server.

3.13.1. YP_009724392.1

Firstly we implemented the Kolaskar and Tongaonkar antigenicity measurement tool that makes use of physicochemical properties of amino acid residues and their occurrence frequencies in previously experimented segmental epitopes to predict antigenic elements on proteins (Kolaskar & Tongaonkar, 1990). Higher the score, the greater the role in stimulating an immune response. In our study, the threshold value was 0.940, and window size was fixed at 7. The estimated antigenic potential of the protein was found to be 1.119 at average, with a minimum of 0.947 and a maximum at 1.262. Findings are shown in Table 5.

Usually, the antigenic regions of a protein contain an accessible and uncovered hydrophilic portion on the surface that plays a significant role in binding with the B-cells. It has been established that beta-turns, more so than other secondary structures, are considerably more hydrophilic and more likely to elicit the desired response. We checked these two parameters, accessibility, and hydrophilicity, via the Emini Surface Accessibility (ESA) and Parker Hydrophilicity prediction methods. The accessibility profile was acquired by applying the formula $\boldsymbol{Sn = (n + 4 + i)\ (0.37) - 6}$ where $\boldsymbol{Sn}$ represents the surface probability, $\boldsymbol{n}$ is the fractional surface probability where $\boldsymbol{i}$ may vary from 1 to 6 (Emini et al., 1985). The threshold for the ESA method was fixed at 1.000, the average value, with the minimum value calculated at 0.088 and the maximum value found at 4.316. Two peptides in the region 4–9 and 54–71 were found to be more accessible (Table 6). Chou and Fasman beta-turn analyzing algorithm was applied to predict the beta-turn, i.e. the most hydrophilic part exposed on the surface. The regions 4–9, 12–13, 40–48, 51–71 were found to be more disposed to persuade beta-turns in the peptide.

The epitopes that connect with alleles, antibodies, or cells are usually elastic. The flexible nature assessed by the Karplus & Schulz Flexibility Prediction system associates the flexibility of a protein to its antigenicity (Karplus & Schulz, 1985) was used. The values above 0.920 were considered as

Table 5. Kolaskar & Tongaonkar Antigenicity prediction.

Protein ID	No.	Start	End	Peptide	Length
YP_009724393.1	1	7	39	TITVEELKKLLEQWNLVIGFLFLTWICLLQFAY	33
	2	45	74	FLYIIKLIFLWLLWPVTLACFVLAAVYRIN	30
	3	78	105	GGIAIAMACLVGLMWLSYFIASFRLFAR	28
	4	117	199	NILLNVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGR CDIKDLPKEITVATSRTLSYYKLGASQRVAGDSGFAAYSRY	83
YP_009724395.1	1	4	12	ILFLALITL	9
	2	68	77	PDGVKHVYQL	10
	3	98	115	SPIFLIVAAIVFITLCFT	18
YP_009724396.1	1	28	36	HQPYVVDDP	9
	2	57	63	LIELCVD	7
	3	81	87	VSCLPFT	7
	4	96	103	GSLVVRCS	8
YP_009724397.2	1	12	20	APRITFGGP	9
	2	35	48	ARSKQRRPQGLPNN	14
	3	51	59	SWFTALTQH	9
	4	61	77	KEDLKFPRGQGVPINTN	17
	5	81	93	DDQIGYYRRATRR	13
	6	102	116	KDLSPRWYFYYLGTG	15
	7	118	125	EAGLPYGA	8
	8	128	146	DGIIWVATEGALNTPKDHI	19
	9	154	175	NAAIVLQLPQGTTLPKGFYAEG	22
	10	179	188	GSQASSRSSS	10
	11	204	209	GTSPAR	6
	12	216	233	DAALALLLLDRLNQLESK	18
	13	238	259	GQQQQGQTVTKKSAAEASKKPR	22
	14	263	276	TATKAYNVTQAFGR	14
	15	288	318	DQELIRQGTDYKHWPQIAQFAPSASAFFGMS	31
	16	321	342	GMEVTPSGTWLTYTGAIKLDDK	22
	17	347	368	KDQVILLNKHIDAYKTFPPTEP	22
YP_009725255.1	1	13	19	IYSLLLC	7

Table 6. Emini surface accessibility prediction.

Protein ID	No.	Start	End	Peptide	Length
YP_009724392.1	1	4	9	FVSEET	6
	2	54	71	PSFYVYSRVKNLNSSRVP	18
YP_009724393.1	1	11	20	EELKKLLEQW	10
	2	38	44	AYANRNR	7
	3	103	109	FARTRSM	7
	4	162	167	KDLPKE	6
	5	171	188	ATSRTLSYYKLGASQRVA	18
	6	195	202	AYSRYRIG	8
	7	204	215	YKLNTDHSSSSD	12
YP_009724395.1	1	38	45	GTYEGNSP	8
	2	49	54	LADNKF	6
	3	73	78	HVYQLR	6
	4	89	96	RQEEVQEL	8
YP_009724396.1	1	25	33	CTQHQPYVV	9
	2	41	46	FYSKWY	6
	3	65	70	AGSKSP	6
	4	107	113	DFLEYHD	7
YP_009724397.2	1	4	11	NGPQNQRN	8
	2	36	42	RSKQRRP	7
	3	87	92	YRRATR	6
	4	185	197	RSSSRSRNSSRNS	13
	5	237	242	KGQQQQ	6
	6	254	264	ASKKPRQKRTA	11
	7	277	282	RGPEQT	6
	8	295	300	GTDYKH	6
	9	340	346	DDKDPNF	7
	10	365	377	PTEPKKDKKKKAD	13
	11	384	390	QRQKKQQ	7
YP_009725255.1	1	21	26	MNSRNY	6

more flexible, and peptides 1–19, 27–39, and 47–71 satisfied this criterion. The average flexibility value was 0.965, minimum 0.894, and a maximum of 1.081.

Parker's hydrophilicity prediction tool was used to predict the hydrophilicity of the whole protein. In this method, a hydrophilic scale based on peptide retention times during high-performance liquid chromatography (HPLC) on a reversed-phase column was constructed and used by assigning to each peptide (Parker et al., 1986). The prediction tool demonstrated that at threshold −2.220, areas 2–14 and 37–42 are most hydrophilic when the limit and average was −2.220, and the minimum value was −6.843, and the maximum value was 4.929.

Finally, the BepiPred-2.0, Sequential B-Cell Epitope Prediction tool was used, which predicts B-cell epitopes from a protein sequence, using a Random Forest algorithm trained on epitopes and non-epitope amino acids determined from crystal structures. A sequential prediction smoothing is performed afterwards (Jespersen et al., 2017). The results indicate that predicted epitopes lie in the region 6–9 and 57–71 (Table 7).

By cross-referencing all the previous data obtained for YP_009724392.1, it is speculated that a peptide of desirable length in region 57–71 would be a likely candidate (Figure S1) inducing B-cell mediated immune response.

3.13.2. YP_009724393.1

For the membrane glycoprotein YP_009724393.1, we also conducted all the tests above. The Kolaskar and Tongaonkar antigenicity tested 7–39, 45–74, 78–105, and 117–199 as the most antigenic area (Table 5). The Emini Surface Accessibility method predicted 11–20, 38–44, 103–109, 162–167, 171–188, 195–202, and 204–215 peptides as mostly surface accessible (Table 6). At the same time, the Chou and Fasman visualized 4–8, 39–44, 74–78, 107–118, 123–126, 128–129, 133–134, 154–166 and 174–217 regions as containing beta-turns. The Karplus & Schulz tool was used for predicting flexibility providing peptides of sequence 4–19, 41–44, 49–79, 104–108, 113–117, 124–138, 146–148, 157–176, 180–193, and 197–216 as most elastic. Finally, the Bepipred Linear Epitope

Table 7. Bepipred linear epitope prediction 2.0.

Protein ID	No.	Start	End	Peptide	Length
YP_009724392.1	1	6	9	SEET	4
	2	57	71	YVYSRVKNLNSSRVP	15
YP_009724393.1	1	5	20	NGTITVEELKKLLEQW	16
	2	40	41	AN	2
	3	132	137	PLLESE	6
	4	161	163	IKD	3
	5	180	191	KLGASQRVAGDS	12
	6	199	218	YRIGNYKLNTDHSSSSDNIA	20
YP_009724394.1	1	44	57	LTENKYSQLDEEQP	14
YP_009724395.1	1	17	25	LYHYQECVR	9
	2	33	51	EPCSSGTYEGNSPFHPLAD	19
	3	71	96	VKHVYQLRARSVSPKLFIRQEEVQEL	26
YP_009724396.1	1	23	45	QSCTQHQPYVVDDPCPIHFYSKW	23
	2	48	56	RVGARKSAP	9
	3	63	78	DEAGSKSPIQYIDIGN	16
	4	91	95	QEPKL	5
	5	106	111	EDFLEY	6
YP_009724397.2	1	4	15	NGPQNQRNAPRI	12
	2	17	48	FGGPSDSTGSNQNGERSGARSKQRRPQGLPNN	32
	3	59	105	HGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLS	47
	4	119	127	AGLPYGANK	9
	5	137	163	GALNTPKDHIGTRNPANNAAIVLQLPQ	27
	6	165	216	TTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD	52
	7	226	267	RLNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKA	42
	8	276	299	RRGPEQTQGNFGDQELIRQGTDYK	24
	9	343	348	DPNFKD	6
	10	358	402	DAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDD	45
	11	404	416	SKQLQQSMSSADS	13
YP_009725255.1	1	25	26	NY	2
YP_009725296.1	1	33	39	ELQDHNE	7

Prediction 2.0 technique foresaw 5–20, 40–41, 132–137, 161–163, 180–191, and 199–218 as the potential antigenic proteins (Table 7). Based on all these calculations, 180–188 is the most potent B-cell candidate epitope from YP_009724393.1. Parker's prediction tool also displays this section in the hydrophilic region (Figure S2).

3.13.3. YP_009724394.1

The ORF6 protein YP_009724394.1 was also tested with the above methods to find out the most suitable epitope. Kolaskar and Tongaonkar revealed that the overall structure of the protein is fully antigenic itself (Table 5). The Emini Surface Accessibility method detected 39–40 and 43–58 regions as the most accessible (Table 6) while Chou and Fasman method recognized 22–34 and 36–58 areas as the possible beta-turn sites. The Karplus & Schulz tool identified 38–57 as the area most flexible. Parker's hydrophilicity prediction showed 9–11, 22–23, and 39–58 as the most hydrophilic sections. At last, Bepipred Prediction reported 44–57 amino acids as potential epitopes (Table 7). This epitope satisfies all the essential features and can be considered as a good contender for B-cell targeted response. Details are given in Figure S3.

3.13.4. YP_009724395.1

The Kolaskar and Tongaonkar antigenic results showed that 4–12, 68–77, 98–115 regions are most antigenic (Table 5). The Emini prediction described 17–22, 38–46, 48–54, 73–83, and 86–97 peptides having the highest accessible options (Table 6). According to the Chou and Fasman method, 17–18, 20–26, 33–55, 58–62, 66–74, 80–84 and 96–99 peptides are likely having beta-turns. The Bepipred Prediction method 2.0 revealed 17–25, 33–51, and 71–96 sequences as the potential epitope (Table 7). Lastly, Parker's hydrophilic prediction marked 13–31, 33–55, 58–61, 63–84, 88–97, and 116–117, as the most water-loving regions. It can be concluded that sequence 71–77 (Figure S4) is the most likely to prompt a B-cell response.

3.13.5. YP_009724396.1

All the previous test also conducted for this protein revealed regions, 28–36, 57–63, 81–87, and 96–103 for Kolaskar and Tongaonkar antigenic test (Table 5); 25–33, 41–46, 65–70 and 107–113 for the Emini Surface Accessibility method (Table 6); 20–29, 31–41, 43–46, 51–55, 63–72, 75–88, 90–96, 99–100, and 104–105 for Chou and Fasman approach; 18–37, 50–56, 62–71, 76–79, 90–97 and 106–109 as the most flexible amino acid residues by Karplus & Schulz tool; 15–37, 50–54, 61–70, 77–81, 89–95, 103–105, 109–110, 112–113 hydrophilic sections by Parker's prediction tool; finally Bepipred linear epitope 2.0 predicted 23–45, 48–56, 63–78, 91–95, 106–111 as the possible epitopes (Table 7). Examining these data (Figure S5), the epitope at the position 28–33 is the most favourable B cell target.

3.13.6. YP_009724397.2

The Kolaskar and Tongaonkar antigenic tests revealed many sections of the protein as antigenic (Table 5). There are many sections as well, which were demonstrated as being accessible, i.e. 36–42, 185–197, 254–264 (Table 6). Chou and Fasman predicted numerous positions where beta turns might be present such as 4–51, 57–89, 94–109, etc. The

Kolaskar and Tongaonkar test also predicted many antigenic regions as shown in the graph (Supplementary-8). Karplus & Schulz's tool calculated 4–12, 19–50, 60–63, and a few other sections as flexible. The Bepipred model predicted only one epitope of length five ranging from 37–41 as the potential epitope (Table 7). Parker's hydrophilicity prediction confirmed that this region is hydrophilic (Figure S6). Scrutinizing these epitopes with the predicted result of the other methods satisfied all the parameters and possessed the requirements to sufficiently interact with B-cells.

3.13.7. YP_009725255.1

Bepipred analysis tool predicted an epitope of 2 amino acid long. Since such a small epitope is not sufficient to elicit B-cell response, we have not conducted any further study on it (Figure S7).

3.13.8. YP_009725296.1

The antigenic regions (Table 5) of this protein were determined by the Kolaskar and Tongaonkar test. Amino acids in 32–40 sequences are only surface accessible, as shown by Emini Surface Accessibility method (Table 6). 5–11, 31–40 are the most hydrophilic parts as examined by Chou and Fasman method. Sections 4–6 and 32–39 are surface accessible according to Karplus & Schulz method and 4–8 and 32–40 peptides as most hydrophilic by parkers test. All these results aligned with the Bepipred linear epitope prediction 2.0, which predicted peptide ELQDHNE, in regions 33–39 (Table 7) as the potential B-cell epitope (Figure S8).

4. Discussion

In the present study, overall, we tried to suggest some promising epitopes, which are highly conserved throughout the analysis, as a potential vaccine. At first, the whole proteome of the virus was retrieved and scanned for potential immunogenicity, and we found 21 different epitopes from 7 different proteins that happen to induce an immune response. Their antigenicity and allergenicity were also checked. All these epitopes were then examined for their conservancy using IEDB Epitope Conservancy Analysis tool among the 18,149 SARS CoV-2 proteins, sequenced from different regions of the world that were stored in the NCBI Virus database. Distribution of MHC class I HLA alleles varies across different ethnic groups and geographic areas around the world. Hence, development of efficient vaccine require the consideration of population coverage. The coverage for China, Italy, Spain, and the USA, the most affected countries by SARS CoV-2, showed a favourable value of 97.48%, 96.80%, 96.88%, and 96.14%, respectively. Eighteen epitopes were found to be $>$99.5.% conserved throughout the pandemic, irrespective of different survival factors, i.e. weather, humidity, region, temperature, etc. Based on these results, we selected the top three epitopes namely KAYNVTQAF, LTALRLCAY and ATSRTLSYY, that showed good interactions with the maximum number of MHC alleles and are not allergens for further analysis through Molecular Docking. T cell epitopes, e.g. ATSRTLSYY and KAYNVTQAF of the Membrane Glycoprotein and Nucleocapsid Phosphoprotein as well as LTALRLCAY were antigenic in nature according to VaxiJen 2.0 (Doytchinova & Flower, 2007). The server predicted these proteins to be antigenic based on an overall protective antigen prediction score of 0.5102, 0.5059, and 0.6025, respectively, which is beyond the threshold 0.4. The NetCTL 1.2 web server (Larsen et al., 2007) was used for both the structures to predict $CD8^+$ T cell epitopes based on MHC binding affinity, C-terminal cleavage affinity, TAP transport efficiency, and NetCTL prediction scores. The 3D model of the HLA-A*29:02 and the epitopes were generated and validated before docking using different servers and software. Our study found the highest docking score of -11.3, -11.2 and -10.8 for ATSRTLSYY, KAYNVTQAF, and LTALRLCAY, respectively. Besides, we found that KAYNVTQAF epitope made a maximum of seventeen interactions with MHC allele and ATSRTLSYY epitope made nine interactions with MHC allele. Along with highest docking score and a higher number of interactions, and these two epitopes are also part of two different structural proteins of the virus namely Membrane Glycoprotein and Nucleocapsid Phosphoprotein, which are generally the target of choice for vaccines designing. Hence, we selected these two epitopes for a further 50 ns MD simulations study to validate the docking result.

Molecular Dynamics (MD) of these epitopes were performed with HLA-A*29:02 up to 50 ns. We have conducted RMSD, RMSF, SASA, Rg, PCA and DCCM analysis of the MD data that supported the in silico immunogenic interaction of the epitopes. RMSD analysis inferred that upon the binding of ATSRTLSYY epitope to the HLA-A*29:02, there was no change in the stability of HLA-A*29:02 since the epitope and HLA-A*29:02 showed a similar pattern of RMSD while analyzed them separately from their docking complex. RMSF, Rg and SASA analysis also revealed the strong binding pattern between HLA-A*29:02 and ATSRTLSYY epitope. The number of hydrogen bond analysis concluded that the epitope ATSRTLSYY made a higher number of hydrogen bonds compared to epitope KAYNVTQAF during binding with HLA-A*29:02, revealing the good stability of the HLA-A*29:02_ATSRTLSYY complex. Moreover, binding free energy analysis also revealed a similar result obtained from the above analyses. The HLA-A*29:02_ATSRTLSYY complex revealed higher binding energy compared to HLA-A*29:02_KAYNVTQAF complex, representing the better binding affinity and stable complex formation for ATSRTLSYY epitope.

The secondary element analysis data revealed the quite same pattern of secondary structure components for HLA-A*29:02_Apo and HLA-A*29:02_KAYNVTQAF, compared to HLA-A*29:02_ATSRTLSYY which revealed its different nature of binding and may contribute in its different level of stable nature.

Principal component analysis disclosed that the energy distribution of HLA-A*29:02_ATSRTLSYY was condensed due to ligand binding while compared with Apo-HLA-A*29:02 and HLA-A*29:02_KAYNVTQAF complex. Again, loading plot of the PCA exhibits the less fluctuating nature of HLA-A*29:02_ATSRTLSYY which revealed it lesser distribution and

less deviation from the same plane in the score plot upon ligand binding compared to HLA-A*29:02_KAYNVTQAF complex. Also, the HLA-A*29:02_ATSRTLSYY showed relatively lower phase space while compared to both HLA-A*29:02_Apo and HLA-A*29:02_KAYNVTQAF system which was consistent with the conclusion of the RMSF, Rg, and the number of hydrogen bond analysis.

Finally, DCCM analysis predicted that the overall conformation of HLA-A*29:02_ATSRTLSYY became compact occupying smaller spaces compared to the overall conformation of HLA-A*29:02_KAYNVTQAF, revealing more non-bond interactions and positively correlated motion between HLA and epitope and support the all previous analyses. Thus, from the overall analysis, we can infer that the epitope ATSRTLSYY is the most potent T-cell epitope among our three selected T-cell epitopes based on MD simulation studies analysis.

Furthermore, we also identified B-cell epitopes for each of the antigenic proteins of SARS CoV-2 by employing IEDB (http://tools.iedb.org/bcell/) B cell epitope prediction tools. The B cell epitope represents a suitable part of an antigen that is identified by either the elicited antibody or the specific B cell receptor in a humoral immune response (Parker, 1993). The uncovering of antigenic B cell epitopes is the key stage in the epitope-dependent vaccines design. Based on linear B cell epitope prediction method BepiPred, a peptide of desirable length in region 57–71 would be a likely candidate (Figure S1) inducing B-cell mediated immune response in YP_009724392.1 protein, which produced ATSRTLSYY T-cell epitope. In YP_009724393.1 protein, 180–188 is the most potent B-cell candidate, which produced LTALRLCAY T-cell epitope. Parker's prediction tool also displays this section in the hydrophilic region (Figure S2). The Bepipred model predicted only one epitope of length five ranging from 37–41 in the YP_009724397.2 protein as the potential epitope (Figure S6). Similarly, Chou-Fasman beta-turn prediction method, Emini surface accessibility and Parker Hydrophilicity prediction methods, Karplus & Schulz Flexibility Prediction system, and Kolaskar and Tongaonkar antigenicity prediction method also detected several regions of B cell epitope that could be of a significant interest for research community in the development of potent vaccines against SARS-CoV-2. Overall, after completing this evaluation, we found that seven of them possess potential regions that can elicit a good immune response. One protein (YP_009725296.1) showed two amino acid residue long epitope for B-cell, which is not viable to induce antibody production or B-cell mediated immunity. That is why we have excluded the possibility of this protein. Along with the T-cell epitopes, these B-cell epitopes can play a crucial role in fighting SARS-CoV-2 virus and can be a good choice to construct a multi-epitope vaccine.

5. Conclusion

This study has suggested 21 T-Cell epitopes as therapeutic targets where the best three of them are tested for vaccine potentiality via molecular docking with Human Leukocyte Antigen (HLA). Based on the docking result, Molecular Dynamics was performed for the best two epitopes (KAYNVTQAF & ATSRTLSYY). Interestingly, the epitope (KAYNVTQAF) with most MHC allele binding ability, showed less binding affinity in docking comparative to the second-best epitope (ATSRTLSYY). Furthermore, molecular dynamics revealed ATSRTLSYY as the most potential epitope. Also, this epitope showed almost fully (99.67%) conserved nature among the 1531 Membrane Glycoprotein sequenced from different regions of the world till April 2020. This work also identified B-Cell epitopes from seven distinct antigenic proteins of SARS-CoV-2 which are capable of inducing the immune response strongly. We hope that our findings will help to prioritize the target for treatment, which will reduce the wet lab effort as well as smoothen the experiments of *in vitro* and *in vivo* studies.

Acknowledgements

Authors are thankful to Molecular Modeling Drug-design and Discovery Laboratory, Pharmacology Research Division, BCSIR Laboratories Chattogram, Bangladesh Council of Scientific and Industrial Research, Chattogram, Bangladesh, for providing software supports during perform protein preparation (Schrodinger Suite) and molecular dynamics simulation (YASARA).

Disclosure statement

The authors declare no competing interests

Author contributions

Conceptualization: M.J., and M.M.A.E.; methodology: M.M.A.E., and M.J.; data curation: M.M.A.E., A.N., A.S., S.M.A.N. and M.S.H.; formal analysis & investigation: M.M.A.E., M.J., Y.A., and S.S.A.; validation, M.J., and M.M.A.E.; resources: S.M.Z.H.; writing—original draft preparation: Y.A., M.M.A.E., M.J., A.N., A.S., S.S.A., and S.M.A.N.; writing—review and editing, Y.A., M.M.A.E., and M.J.; visualization, M.J., and M.M.A.E; supervision: M.J.; suggestion for journal selection: M.A.M.; project administration, M.J. and S.M.Z.H. All authors read and approved the final manuscript.

ORCID

Md. Muzahid Ahmed Ezaj http://orcid.org/0000-0002-1139-985X
Md. Junaid http://orcid.org/0000-0003-3588-4937
Yeasmin Akter http://orcid.org/0000-0002-5417-8918
Afsana Nahrin http://orcid.org/0000-0003-2646-3735
Aysha Siddika http://orcid.org/0000-0001-8596-0630
Syeda Samira Afrose http://orcid.org/0000-0002-2057-8227
S. M. Abdul Nayeem http://orcid.org/0000-0002-3167-0651
Md. Sajedul Haque http://orcid.org/0000-0002-6343-4901
Mohammad Ali Moni http://orcid.org/0000-0003-0756-1006
S. M. Zahid Hosen http://orcid.org/0000-0001-5789-3418

Reference

Amadei, A., Linssen, A. B., & Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, *17*(4), 412–425. https://doi.org/10.1002/prot.340170408

Angelo, M. A., Grifoni, A., O'Rourke, P. H., Sidney, J., Paul, S., & Peters, B. (2017). Human CD4+ T cell responses to an attenuated tetravalent dengue vaccine parallel those induced by natural infection in magnitude, HLA restriction, and antigen specificity. *Journal of Virology*, *91*, e02147-16.

Badawi, M. M., Fadl Alla, A., Alam, S. S., Mohamed, W. A., Osman, D., & Alrazig Ali, S. (2016). Immunoinformatics predication and in silico modeling of epitope-based peptide vaccine against virulent Newcastle disease viruses. *American Journal of Infectious Diseases and Microbiology*, *4*, 61–71.

Bahrami, A. A., Payandeh, Z., Khalili, S., Zakeri, A., & Bandehpour, M. (2019). Immunoinformatics: In silico approaches and computational design of a multi-epitope, immunogenic protein. *International Reviews of Immunology*, *38*(6), 307–322. https://doi.org/10.1080/08830185.2019.1657426

Balmith, M., & Soliman, M. E. (2017). Non-active site mutations disturb the loop dynamics, dimerization, viral budding and egress of VP40 of the Ebola virus. *Molecular BioSystems*, *13*(3), 585–597. https://doi.org/10.1039/c6mb00803h

Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., Felts, A. K., Halgren, T. A., Mainz, D. T., Maple, J. R., Murphy, R., Philipp, D. M., Repasky, M. P., Zhang, L. Y., Berne, B. J., Friesner, R. A., Gallicchio, E., & Levy, R. M. (2005). Integrated modeling program, applied chemical theory (IMPACT). *Journal of Computational Chemistry*, *26*(16), 1752–1780. https://doi.org/10.1002/jcc.20292

Bappy, S. S., Sultana, S., Adhikari, J., Mahmud, S., Khan, M. A., & Kaderi Kibria, K. (2020). Extensive immunoinformatics study for the prediction of novel peptide-based epitope vaccine with docking confirmation against envelope protein of Chikungunya virus: A computational biology approach. *Journal of Biomolecular Structure and Dynamics*, *24*, 1–30.

Benkert, P., Tosatto, S. C., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, *71*(1), 261–277. https://doi.org/10.1002/prot.21715

Blatt, J. M., & Weisskopf, V. F. (1991). *Theoretical nuclear physics*. Courier Corporation.

Blum, J. S., Wearsch, P. A., & Cresswell, P. (2013). Pathways of antigen processing. *Annual Review of Immunology*, *31*, 443–473. https://doi.org/10.1146/annurev-immunol-032712-095910

Bornot, A., & de Brevern, A. G. (2006). Protein beta-turn assignments. *Bioinformation*, *1*(5), 153–155.

Brusic, V., & Petrovsky, N. (2005). Immunoinformatics and its relevance to understanding human immune disease. *Expert Review of Clinical Immunology*, *1*(1), 145–157. https://doi.org/10.1586/1744666X.1.1.145

Bui, H.-H., Sidney, J., Li, W., Fusseder, N., & Sette, A. (2007). Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics*, *8*(1), 361. https://doi.org/10.1186/1471-2105-8-361

Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C. C.-Y., Poon, R. W.-S., Tsoi, H.-W., Lo, S. K.-F., Chan, K.-H., Poon, V. K.-M., Chan, W.-M., Ip, J. D., Cai, J.-P., Cheng, V. C.-C., Chen, H., Hui, C. K.-M., & Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet (London, England)*, *395*(10223), 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9

Chou, P. Y., & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual Review of Biochemistry*, *47*(1), 251–276. https://doi.org/10.1146/annurev.bi.47.070178.001343

Colovos, C., & Yeates, T. (1993). ERRAT: An empirical atom-based method for validating protein structures. *Protein Science*, *2*(9), 1511–1519. https://doi.org/10.1002/pro.5560020916

Daydé-Cazals, B., Fauvel, B., Singer, M., Feneyrolles, C., Bestgen, B., Gassiot, F., Spenlinhauer, A., Warnault, P., Van Hijfte, N., Borjini, N., Chevé, G., & Yasri, A. (2016). Rational design, synthesis, and biological evaluation of 7-azaindole derivatives as potent focused multi-targeted kinase inhibitors. *Journal of Medicinal Chemistry*, *59*(8), 3886–3905. https://doi.org/10.1021/acs.jmedchem.6b00087

Deming, D., Sheahan, T., Heise, M., Yount, B., Davis, N., Sims, A., Suthar, M., Harkema, J., Whitmore, A., Pickles, R., West, A., Donaldson, E., Curtis, K., Johnston, R., & Baric, R. (2006). Vaccine efficacy in senescent mice challenged with recombinant SARS-CoV bearing epidemic and zoonotic spike variants. *PLoS Medicine*, *3*(12), e525. https://doi.org/10.1371/journal.pmed.0030525

Dickson, C. J., Madej, B. D., Skjevik, A. A., Betz, R. M., Teigen, K., Gould, I. R., & Walker, R. C. (2014). Lipid14: The amber lipid force field. *Journal of Chemical Theory and Computation*, *10*(2), 865–879. https://doi.org/10.1021/ct4010307

Doss, C. G. P., & Rajith, B. (2013). A new insight into structural and functional impact of single-nucleotide polymorphisms in PTEN gene. *Cell Biochemistry and Biophysics*, *66*(2), 249–263. https://doi.org/10.1007/s12013-012-9472-9

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

Dudek, N. L., Perlmutter, P., Aguilar, M. I., Croft, N. P., & Purcell, A. W. (2010). Epitope discovery and their use in peptide based vaccines. *Current Pharmaceutical Design*, *16*(28), 3149–3157. https://doi.org/10.2174/138161210793292447

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in Enzymology*, *277*, 396–404.

Emini, E. A., Hughes, J. V., Perlow, D., & Boger, J. (1985). Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, *55*(3), 836–839. https://doi.org/10.1128/JVI.55.3.836-839.1985

Farrokhzadeh, A., Akher, F. B., & Soliman, M. E. (2019). Probing the dynamic mechanism of uncommon allosteric inhibitors optimized to enhance drug selectivity of SHP2 with therapeutic potential for cancer treatment. *Applied Biochemistry and Biotechnology*, *188*(1), 260–281. https://doi.org/10.1007/s12010-018-2914-0

Gao, J., Tian, Z., & Yang, X. (2020). Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *BioScience Trends*, 14*(1), 72-73*.

Ghosh, A., & Vishveshwara, S. (2007). A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(40), 15711–15716. https://doi.org/10.1073/pnas.0704459104

Graham, R. L., Becker, M. M., Eckerle, L. D., Bolles, M., Denison, M. R., & Baric, R. S. (2012). A live, impaired-fidelity coronavirus vaccine protects in an aged, immunocompromised mouse model of lethal disease. *Nature Medicine*, *18*(12), 1820–1826. https://doi.org/10.1038/nm.2972

Hasan, A., Hossain, M., & Alam, J. (2013). A computational assay to design an epitope-based peptide vaccine against Saint Louis encephalitis virus. *Bioinformatics and Biology Insights*, *7*, 347–355.

Haste Andersen, P., Nielsen, M., & Lund, O. (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Science*, *15*(11), 2558–2567. https://doi.org/10.1110/ps.062405906

Hegde, N. R., Gauthami, S., Sampath Kumar, H., & Bayry, J. (2018). The use of databases, data mining and immunoinformatics in vaccinology: Where are we? *Expert Opinion on Drug Discovery*, *13*(2), 117–130. https://doi.org/10.1080/17460441.2018.1413088

Heymann, D. L. (2020). Data sharing and outbreaks: Best practice exemplified. *The Lancet*, *395*(10223), 469–470. https://doi.org/10.1016/S0140-6736(20)30184-7

Hoffmann, M., Kleine-Weber, H., Krueger, N., Mueller, M. A., Drosten, C., & Pöhlmann, S. (2020). The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *BioRxiv*. https://doi.org/10.1101/2020.01.31.929042

Hosen, S. Z., Dash, R., Junaid, M., Mitra, S., & Absar, N. (2019). Identification and structural characterization of deleterious non-synonymous single nucleotide polymorphisms in the human SKP2 gene. *Computational Biology and Chemistry*, *79*, 127–136. https://doi.org/10.1016/j.compbiolchem.2019.02.003

Hossain, M. U., Keya, C. A., Das, K. C., Hashem, A., Omar, T. M., Khan, M. A., Rakib-Uz-Zaman, S. M., & Salimullah, M. (2018). An immunopharmacoinformatics approach in development of vaccine and drug candidates for West Nile virus. *Frontiers in Chemistry*, *6*, 246. https://doi.org/10.3389/fchem.2018.00246

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Hui, D. S., I Azhar, E., Madani, T. A., Ntoumi, F., Kock, R., Dar, O., Ippolito, G., Mchugh, T. D., Memish, Z. A., Drosten, C., Zumla, A., & Petersen, E. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases: IJID*, *91*, 264–266. https://doi.org/10.1016/j.ijid.2020.01.009

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*., *14*(1), 33–38. [Database] https://doi.org/10.1016/0263-7855(96)00018-5

Islam, M. J., Khan, A. M., Parves, M. R., Hossain, M. N., & Halim, M. A. (2019). Prediction of deleterious non-synonymous SNPs of human STK11 gene by combining algorithms, molecular docking, and molecular dynamics simulation. *Scientific Reports*, *9*, 1–16.

Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, *45*(W1), W24–W29. https://doi.org/10.1093/nar/gkx346

Junaid, M., Alam, M. J., Hossain, M. K., Halim, M. A., & Ullah, M. O. (2018). Molecular docking and dynamics of Nickel-Schiff base complexes for inhibiting β-lactamase of Mycobacterium tuberculosis. *In Silico Pharmacology*, *6*(1), 6. https://doi.org/10.1007/s40203-018-0044-6

Junaid, M., Islam, N., Hossain, M. K., Ullah, M. O., & Halim, M. A. (2019). Metal based donepezil analogues designed to inhibit human acetylcholinesterase for Alzheimer's disease. *PLoS One*, *14*(2), e0211935. https://doi.org/10.1371/journal.pone.0211935

Kamaraj, B., Rajendran, V., Sethumadhavan, R., Kumar, C. V., & Purohit, R. (2015). Mutational analysis of FUS gene and its structural and functional role in amyotrophic lateral sclerosis 6. *Journal of Biomolecular Structure and Dynamics*, *33*(4), 834–844. https://doi.org/10.1080/07391102.2014.915762

Karplus, P., & Schulz, G. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, *72*(4), 212–213. https://doi.org/10.1007/BF01195768

Kaur, H., Garg, A., Raghava, G. P. S. J. P., & Letters, p. (2007). PEPstr: A de novo method for tertiary structure prediction of small bioactive peptides. *Protein & Peptide Letters*, *14*, 626–631.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Khalili, S., Jahangiri, A., Borna, H., Ahmadi Zanoos, K., & Amani, J. (2014). Computational vaccinology and epitope vaccine design by immunoinformatics. *Acta Microbiologica et Immunologica Hungarica*, *61*(3), 285–307. https://doi.org/10.1556/AMicr.61.2014.3.4

Kolaskar, A., & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Letters*, *276*(1–2), 172–174. https://doi.org/10.1016/0014-5793(90)80535-q

Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A., & Vriend, G. (2004). Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins*, *57*(4), 678–683. https://doi.org/10.1002/prot.20251

Krieger, E., Dunbrack, R. L., Hooft, R. W., & Krieger, B. (2012a). Assignment of protonation states in proteins and ligands: Combining pK_a prediction with hydrogen bonding network optimization. In: *Computational Drug Discovery and Design* (pp. 405–421). Springer.

Krieger, E., Dunbrack, R. L., Jr., Hooft, R. W., & Krieger, B. (2012b). Assignment of protonation states in proteins and ligands: Combining pKa prediction with hydrogen bonding network optimization. *Methods in Molecular Biology (Clifton, NJ)*, *819*, 405–421.

Krieger, E., Koraimann, G., & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *Proteins*, *47*(3), 393–402. https://doi.org/10.1002/prot.10104

Krieger, E., & Vriend, G. (2015). New ways to boost molecular dynamics simulations. *Journal of Computational Chemistry*, *36*(13), 996–1007. https://doi.org/10.1002/jcc.23899

Krieger, E., Vriend, G., & Spronk, C. (2013). YASARA—Yet another scientific artificial reality application. YASARA Org.

Lange, O. F., Grubmüller, H., & Groot, B. L, D. (2005). Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions. *Angewandte Chemie (International ed. in English)*, *44*(22), 3394–3399. https://doi.org/10.1002/anie.200462957

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., & Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*., *8*, 424.

Larsen, J. E. P., Lund, O., & Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Research*, *2*(2), 2 https://doi.org/10.1186/1745-7580-2-2

Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, *8*(4), 477–486. https://doi.org/10.1007/BF00228148

Letko, M. C., & Munster, V. (2020). Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. *BioRxiv*. https://doi.org/10.1101/2020.01.22.915660

Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., & Friesner, R. A. (2011). The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins*, *79*(10), 2794–2812. https://doi.org/10.1002/prot.23106

Li, W., Joshi, M. D., Singhania, S., Ramsey, K. H., & Murthy, A. K. (2014). Peptide vaccine: Progress and challenges. *Vaccines*, *2*(3), 515–536. https://doi.org/10.3390/vaccines2030515

Lin, Y., Shen, X., Yang, R. F., Li, Y. X., Ji, Y. Y., He, Y. Y., Shi, M. D., Lu, W., Shi, T. L., Wang, J., Wang, H. X., Jiang, H. L., Shen, J. H., Xie, Y. H., Wang, Y., Pei, G., Shen, B. F., Wu, J. R., & Sun, B. (2003). Identification of an epitope of SARS-coronavirus nucleocapsid protein. *Cell Research*, *13*(3), 141–145. https://doi.org/10.1038/sj.cr.7290158

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine, 27(2), taaa021*.

Liu, X., Shi, Y., Li, P., Li, L., Yi, Y., Ma, Q., & Cao, C. (2004). Profile of antibodies to the nucleocapsid protein of the severe acute respiratory syndrome (SARS)-associated coronavirus in probable SARS patients. *Clinical and Diagnostic Laboratory Immunology*, *11*(1), 227–228. https://doi.org/10.1128/cdli.11.1.227-228.2004

Liu, X., & Wang, X.-J. (2020). Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. *BioRxiv*. https://doi.org/10.1101/2020.01.29.924100.

Li, C. K.-f., Wu, H., Yan, H., Ma, S., Wang, L., Zhang, M., Tang, X., Temperton, N. J., Weiss, R. A., Brenchley, J. M., Douek, D. C., Mongkolsapaya, J., Tran, B.-H., Lin, C-l S., Screaton, G. R., Hou, J-l., McMichael, A. J., & Xu, X.-N. (2008). T cell responses to whole SARS coronavirus in humans. *The Journal of Immunology*, *181*(8), 5490–5500. https://doi.org/10.4049/jimmunol.181.8.5490

Lovering, A. L., Lee, S. S., Kim, Y.-W., Withers, S. G., & Strynadka, N. C. (2005). Mechanistic and structural analysis of a family 31 alpha-glycosidase and its glycosyl-enzyme intermediate. *The Journal of Biological Chemistry*, *280*(3), 2105–2115. https://doi.org/10.1074/jbc.M410468200

Lu, H., Stratton, C. W., & Tang, Y. W. (2020). Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology*, *92*(4), 401–402. https://doi.org/10.1002/jmv.25678

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet (London, England)*, *395*(10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Martens, H., & Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.

Maurer-Stroh, S., Krutz, N. L., Kern, P. S., Gunalan, V., Nguyen, M. N., Limviphuvadh, V., Eisenhaber, F., & Gerberick, G. F. (2019). AllerCatPro-prediction of protein allergenicity potential from the protein sequence. *Bioinformatics (Oxford, England)*, *35*(17), 3020–3027. https://doi.org/10.1093/bioinformatics/btz029

Mehla, K., & Ramana, J. (2016). Identification of epitope-based peptide vaccine candidates against enterotoxigenic Escherichia coli: A comparative genomics and immunoinformatics approach. *Molecular BioSystems*, *12*(3), 890–901. https://doi.org/10.1039/c5mb00745c

Monterrubio-López, G. P., González-Y-Merchand, J. A., & Ribas-Aparicio, R. M. (2015). Identification of novel potential vaccine candidates against tuberculosis based on reverse vaccinology. *BioMed Research International*, *2015*, 483150. https://doi.org/10.1155/2015/483150

Narang, S. S., Shuaib, S., Goyal, D., & Goyal, B. (2018). Assessing the effect of D59P mutation in the DE loop region in amyloid aggregation propensity of β2-microglobulin: A molecular dynamics simulation study. *Journal of Cellular Biochemistry*, *119*(1), 782–792. https://doi.org/10.1002/jcb.26241

Ndagi, U., Mhlongo, N. N., & Soliman, M. E. (2017). The impact of Thr91 mutation on c-Src resistance to UM-164: Molecular dynamics study revealed a new opportunity for drug design. *Molecular BioSystems*, *13*(6), 1157–1171. https://doi.org/10.1039/c6mb00848h

Oany, A. R., Emran, A.-A., & Jyoti, T. P. (2014). Design of an epitope-based peptide vaccine against spike protein of human coronavirus: An in silico approach. *Drug Design, Development and Therapy*, *8*, 1139–1149.

Parker, D. C. (1993). T cell-dependent B cell activation. *Annual Review of Immunology*, *11*, 331–360. https://doi.org/10.1146/annurev.iy.11.040193.001555

Parker, J., Guo, D., & Hodges, R. (1986). New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: Correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, *25*(19), 5425–5432. https://doi.org/10.1021/bi00367a013

Purcell, A. W., McCluskey, J., & Rossjohn, J. (2007). More than one reason to rethink the use of peptides in vaccine design. *Nature Reviews. Drug Discovery*, *6*(5), 404–414. https://doi.org/10.1038/nrd2224

R Core Team. (2019). *R: A Language and Environment for Statistical Computing (Version 3.5.2)*. R Foundation for Statistical Computing.

Rascón-Castelo, E., Burgara-Estrella, A., Mateu, E., & Hernández, J. J. V. (2015). Immunological features of the non-structural proteins of porcine reproductive and respiratory syndrome virus. *Viruses*, *7*(3), 873–886. https://doi.org/10.3390/v7030873

Robson, B. (2020). Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. *Computers in Biology and Medicine.*, *119*, 103670. https://doi.org/10.1016/j.compbiomed.2020.103670

Saha, S., & Raghava, G. P. S. (2006). AlgPred: Prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Research*, *34*(Web Server issue), W202–W209. https://doi.org/10.1093/nar/gkl343

Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and methods for T- and B-cell epitope prediction. *Journal of Immunology Research*, *2017*, 2680160.

Singh, S., Singh, H., Tuknait, A., Chaudhary, K., Singh, B., Kumaran, S., & Raghava, G. P. S. (2015). PEPstrMOD: Structure prediction of peptides containing natural, non-natural and modified residues. *Biology Direct*, *10*, 73. https://doi.org/10.1186/s13062-015-0103-4

Sittel, F., Jain, A., & Stock, G. (2014). Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *Journal of Chemical Physics*, *141*, *014111*.

Srinivasan, E., & Rajasekaran, R. (2016). Computational investigation of curcumin, a natural polyphenol that inhibits the destabilization and the aggregation of human SOD1 mutant (Ala4Val). *RSC Advances*, *6*(104), 102744–102753. https://doi.org/10.1039/C6RA21927F

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., & Holzhütter, H.-G. (2005). Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cellular and Molecular Life Sciences: CMLS*, *62*(9), 1025–1037. https://doi.org/10.1007/s00018-005-4528-2

Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334

van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J. M., Wolthers, K. C., Wertheim-van Dillen, P. M. E., Kaandorp, J., Spaargaren, J., & Berkhout, B. (2004). Identification of a new human coronavirus. *Nature Medicine*, *10*(4), 368–373. https://doi.org/10.1038/nm1024

Vetrivel, U., Nagarajan, H., & Thirumudi, I. (2018). Design of inhibitory peptide targeting Toxoplasma gondii RON4-human β-tubulin interactions by implementing structural bioinformatics methods. *Journal of Cellular Biochemistry*, *119*(4), 3236–3246. https://doi.org/10.1002/jcb.26480

Wan, Y., Shang, J., Graham, R., Baric, R. S., & Li, F. (2020). Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *Journal of Virology*, *94*(7), e00127–20.

Wang, Y., Li, Y., Ma, Z., Yang, W., & Ai, C. (2010). Mechanism of microRNA-target interaction: Molecular dynamics simulations and thermodynamics analysis. *PLoS Computational Biology*, *6, e1000866*.

Wang, J., Wen, J., Li, J., Yin, J., Zhu, Q., Wang, H., Yang, Y., Qin, E., You, B., Li, W., Li, X., Huang, S., Yang, R., Zhang, X., Yang, L., Zhang, T., Yin, Y., Cui, X., Tang, X., … Liu, S. (2003). Assessment of immunoreactive synthetic peptides from the structural proteins of severe acute respiratory syndrome coronavirus. *Clinical Chemistry*, *49*(12), 1989–1996. https://doi.org/10.1373/clinchem.2003.023184

Wickham, H. (2009). *ggplot2*. Springer.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA*, *323*(13), 1239–1242. https://doi.org/10.1001/jama.2020.2648

Yang, Z.-Y., Kong, W.-P., Huang, Y., Roberts, A., Murphy, B. R., Subbarao, K., & Nabel, G. J. (2004). A DNA vaccine induces SARS coronavirus neutralization and protective immunity in mice. *Nature*, *428*(6982), 561–564. https://doi.org/10.1038/nature02463

Yang, Y., Liu, H., & Yao, X. (2012). Understanding the molecular basis of MK2-p38α signaling complex assembly: Insights into protein-protein interaction by molecular dynamics and free energy studies. *Molecular BioSystems*, *8*(8), 2106–2118. https://doi.org/10.1039/c2mb25042j

Yesudhas, D., Anwar, M. A., Panneerselvam, S., Durai, P., Shah, M., & Choi, S. (2016). Structural mechanism behind distinct efficiency of Oct4/Sox2 proteins in differentially spaced DNA complexes. *PloS One*, *11*.

Zhang, M., Ishii, K., Hisaeda, H., Murata, S., Chiba, T., Tanaka, K., Li, Y., Obata, C., Furue, M., & Himeno, K. (2004). Ubiquitin-fusion degradation pathway plays an indispensable role in naked DNA vaccination with a chimeric gene encoding a syngeneic cytotoxic T lymphocyte epitope of melanocyte and green fluorescent protein. *Immunology*, *112*(4), 567–574. https://doi.org/10.1111/j.1365-2567.2004.01916.x

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., … Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–274. https://doi.org/10.1038/s41586-020-2012-7