JMIR Bioinformatics and Biotechnology

Methods, devices, web-based platforms, open data and open software tools for big data analytics, understanding biological/medical data, and information retrieval in biology and medicine.
Volume 1 (2020), Issue 1   ISSN 2563-3570   Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

Contents

Tutorial

Original Papers

Tutorial

Molecular Docking Using Chimera and Autodock Vina Software for Nonbioinformaticians

Sania Safdar Butt[1], MS; Yasmin Badshah[1], MPhil; Maria Shabbir[1], PhD; Mehak Rafiq[2], PhD

[1]Atta Ur Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan

[2]Research Centre for Modelling and Simulation, National University of Sciences and Technology, Islamabad, Pakistan

Corresponding Author:
Mehak Rafiq, PhD
Research Centre for Modelling and Simulation
National University of Sciences and Technology
H-12
Islamabad
Pakistan
Phone: 92 5190855733
Email: mehak@rcms.nust.edu.pk

Abstract

In the field of drug discovery, many methods of molecular modeling have been employed to study complex biological and chemical systems. Experimental strategies are integrated with computational approaches for the identification, characterization, and development of novel drugs and compounds. In modern drug designing, molecular docking is an approach that explores the confirmation of a ligand within the binding site of a macromolecule. To date, many software and tools for docking have been employed. AutoDock Vina (in UCSF [University of California, San Francisco] Chimera) is one of the computationally fastest and most accurate software employed in docking. In this paper, a sequential demonstration of molecular docking of the ligand fisetin with the target protein Akt has been provided, using AutoDock Vina in UCSF Chimera 1.12. The first step involves target protein ID retrieval from the protein database, the second step involves visualization of the protein structure in UCSF Chimera, the third step involves preparation of the target protein for docking, the fourth step involves preparation of the ligand for docking, the fifth step involves docking of the ligand and the target protein as Mol.2 files in Chimera by using AutoDock Vina, and the final step involves interpretation and analysis of the docking results. By following the guidelines and steps outlined in this paper, researchers with no previous background in bioinformatics research can perform computational docking in an easier and more user-friendly manner.

(JMIR Bioinformatics Biotechnol 2020;1(1):e14232) doi:10.2196/14232

KEYWORDS

virtual screening; molecular docking; Autodock Vina; UCSF Chimera; nonbioinformaticians; binding affinity; protein database; computer-aided docking, free offline docking; drug discovery

Introduction

In the modern era of pharmaceutical research, many methods of molecular modeling have been employed to study complex chemical and biological systems in a variety of programs of drug discovery. It is very important to integrate experimental strategies into computational approaches in the identification, characterization, and development of novel and propitious compounds. Molecular docking is an approach used extensively in modern drug designing and development; it explores the conformations of ligands within the macromolecular target binding site, providing an estimation of receptor-ligand binding free energy for all different conformations. Small molecular compounds (ligands) are docked into the binding site of the receptor, following which the binding affinity of the complex is estimated. This constitutes a significant part of the structure-based drug design process. For a thorough understanding and estimation of the ligand/protein complex, the ability to visualize the binding interactions and geometries by using a fast and accurate protocol for docking is required [1].

To date, a variety of algorithms for docking are available, which can lead to a better understanding of the benefits and drawbacks of these methods. However, most of the free tools rely on the knowledge of the command-line interface. For biologists, this is a laborious process and hence they avoid it. The proper estimation of each method can lead to the development of plausible strategies and the origination of reproducible and relevant results.

Autodock and AutoDock Vina (The Scripps Research Institute) are some of the most widely used, free, open-source tools for molecular docking simulations [2]. AutoDock is a collection of command-line programs that can be employed to predict binding conformations of a small flexible ligand to a macromolecular target whose structure is known. This technique combines the rapid grid-based method used for energy evaluation with conformation searching and simulated annealing.

AutoDock 4 was used for molecular docking previously. The new AutoDock Vina has a more accurate binding algorithm that can speed up the rate by approximately 2 orders of magnitude as compared to AutoDock 4. In addition, AutoDock 4 has significantly improved predictions of binding mode, assessed by the training tests employed in the AutoDock 4. By the use of multithreading on the multicore machines, faster processing can be achieved from parallelism. AutoDock Vina clusters the results for the user in a transparent way and automatically calculates the grid maps.

The UCSF (University of California, San Francisco) Chimera software is used for visualization as well as analysis of density maps, 3D microscopy, molecular structures, and the associated data [3]. The challenges in the scope, size, and types of data used with the experimental cutting-edge methods are addressed by this software. It provides advanced options for high-quality rendering (reliable calculations of the molecular surface, interactive ambient occlusion, etc) and provides professional approaches to the design and distribution of the software. Chimera is a freely available software for noncommercial use and shows advances particularly in its performance, extensibility, visualization, and usability.

Chimera is segmented into major components: a core that has a role in providing visualization and basic services and extensions that have a higher-level functionality. Two major extensions of Chimera are very important: the first one is the multiscale, which can visualize the molecular assemblies of large-scale components such as the viral coats, and the second one is collaborative interface, which allows sharing of the chimera session interactively, despite being at separate locales. The other extensions of chimera include the Multalign Viewer, which shows multiple sequence alignments and the associated structures, the Movie that replays the trajectories of molecular dynamics, the Volume Viewer that is responsible for displaying and analyzing the volumetric data, and ViewDock that screens the docked ligand orientations. Chimera is available for all operating systems. It can be freely used by academic and nonprofit users.

For the purpose of this protocol, Akt and flavonoid fisetin are used. Protein kinase B, also known as Akt, is a serine/threonine-specific protein that regulates cell growth and survival [4]. In various cancers, the PI3K or Akt signaling cascade is upregulated and linked with enhanced progression and proliferation of cancer cells. Akt is an important part of signaling cascades for cell endurance and growth throughout the progression and proliferation in cancer. It controls the cell cycle, growth, and survival by indirectly altering cyclin D1 levels and directly activating inhibitors of cyclin-dependent kinases (WAF1/p21 and KIP1/p27) [5].

The plant-derived flavonoid named fisetin present in various edible natural sources is reported to possess antiproliferative potential [6]. Invasion, proliferation, and metastatic growth are inhibited significantly by the use of various concentrations of fisetin, especially in lung cancer. Current research has reported that the PI3K/Akt cascade is a direct target of fisetin in human cells, which is a hallmark for growth and survival [7].

The tools employed in Chimera are robust, simple, and interactive, and the computations involved take a few seconds. The major benefit of Chimera is that it integrates a large collection of interactive methods. These tools also play a role in the preparation of input and examination of results from more specialized, complex, and noninteractive algorithmic analysis software. Both the interactive and the noninteractive analyses are beneficial.

Methods

Requirements for Docking

Docking requires the following: (1) Windows 7, 8, or 10 or Mac operating system and Linux, and (2) UCSF Chimera 1.12.

Instructions

The stepwise instructions for docking are provided below:

Retrieval

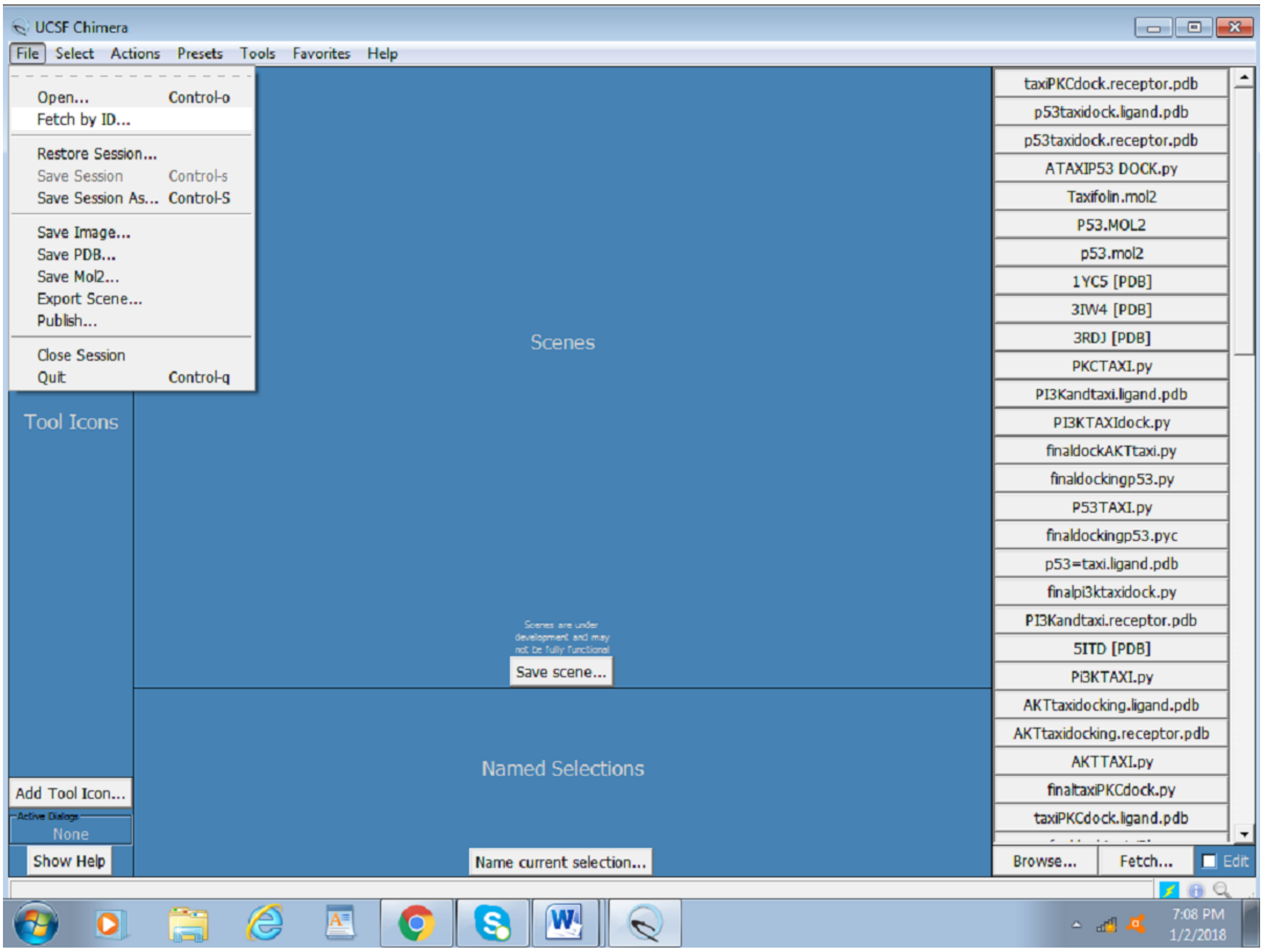
Retrieve the required target protein structure from the major database Protein Data Bank (PDB) [8,9] as a PDB file.

Use of UCSF Chimera for Docking the Target Protein

UCSF Chimera is an extensible program that is meant mainly for visualization and analysis of the molecular structures. However, in this paper, we are operating Autodock Vina in Chimera for docking purposes.

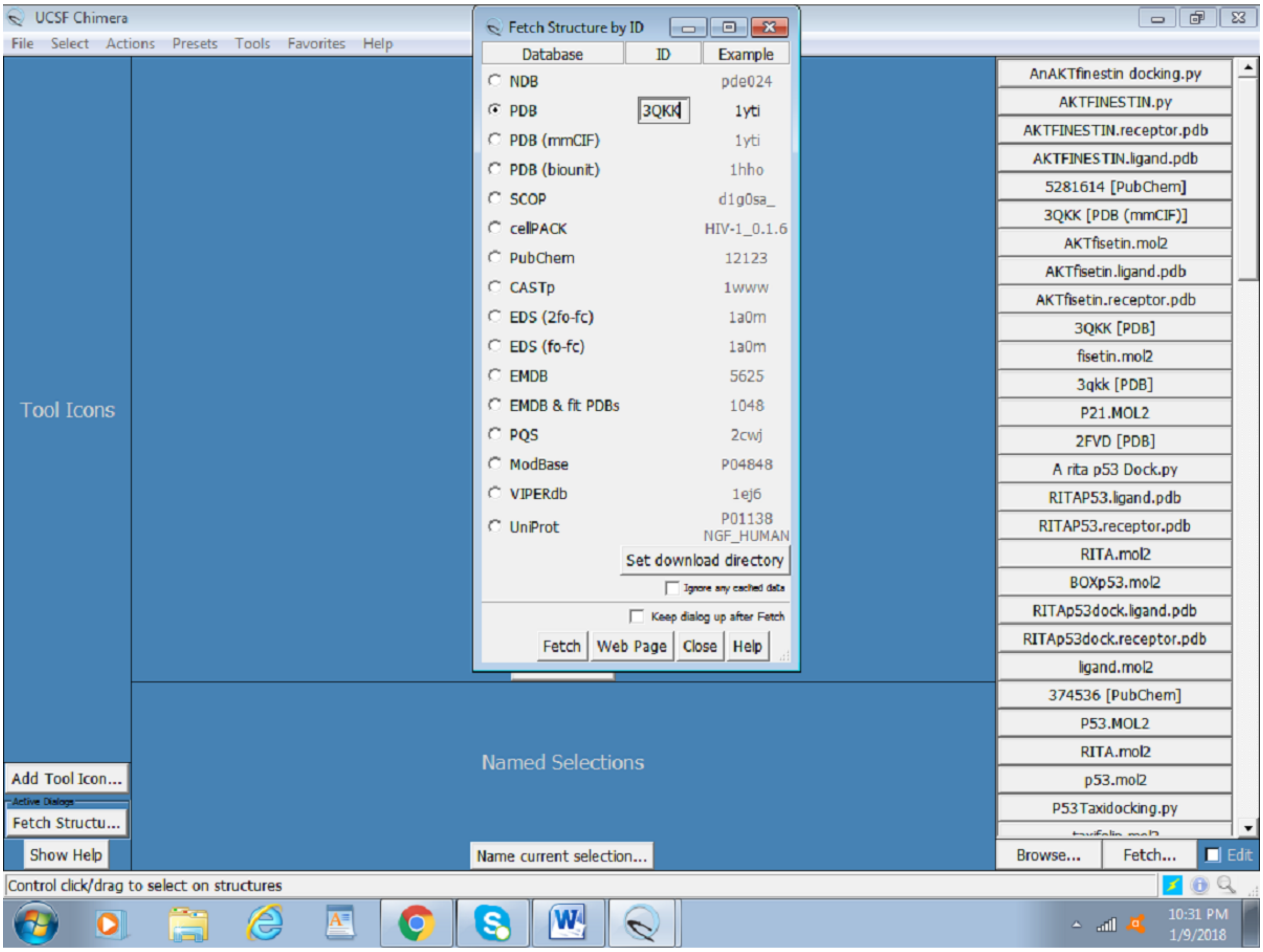
1. Click on the file and fetch by ID, as shown in Figure 1.
2. Input the PDB ID of the protein (Akt: 3QKK). Figure 2 presents a screenshot of how to obtain the protein structure through PDB ID in Chimera. Any protein can be fetched by inserting the PDB ID of the protein.
3. When the protein is fetched, its structure is downloaded through the website; hence, a working internet connection is required, or the PDB file can be downloaded beforehand and simply be opened thorough File > open. Figure 3 displays the Akt structure retrieved in UCSF Chimera.
4. Create a working directory for the docking project that is convenient to access, such as Users/Desktop/Docking/. Start saving all your prepared files there, for example, save 3QKK as Akt.pdb.

Figure 1. Screenshot of the process to fetch/deliver a protein structure in Chimera.



Scenes are under development and may not be fully functional

Figure 2. Screenshot of the retrieved protein structure of Akt from the RCSB protein database.



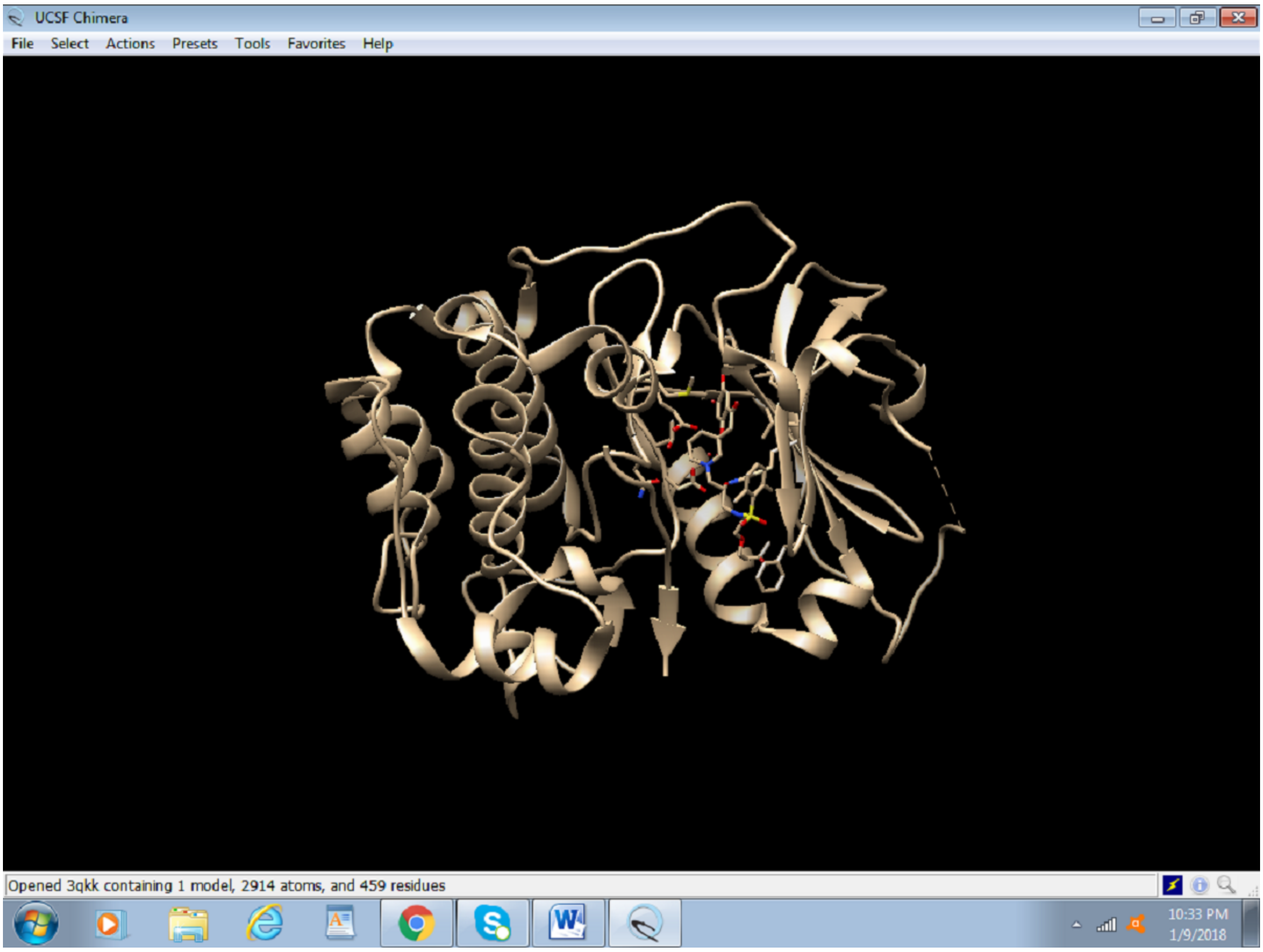
Fetch Structure by ID

Database	ID	Example
○ NDB		pde024
● PDB	3QKK	1yti
○ PDB (mmCIF)		1yti
○ PDB (biounit)		1hho
○ SCOP		d1g0sa_
○ cellPACK		HIV-1_0.1.6
○ PubChem		12123
○ CASTp		1www
○ EDS (2fo-fc)		1a0m
○ EDS (fo-fc)		1a0m
○ EMDB		5625
○ EMDB & fit PDBs		1048
○ PQS		2cwj
○ ModBase		P04848
○ VIPERdb		1ej6
○ UniProt		P01138 NGF_HUMAN

□ Ignore any cached data

□ Keep dialog up after Fetch

Figure 3. Screenshot of the retrieval of protein database structure of the protein in UCSF (University of California, San Francisco) Chimera.

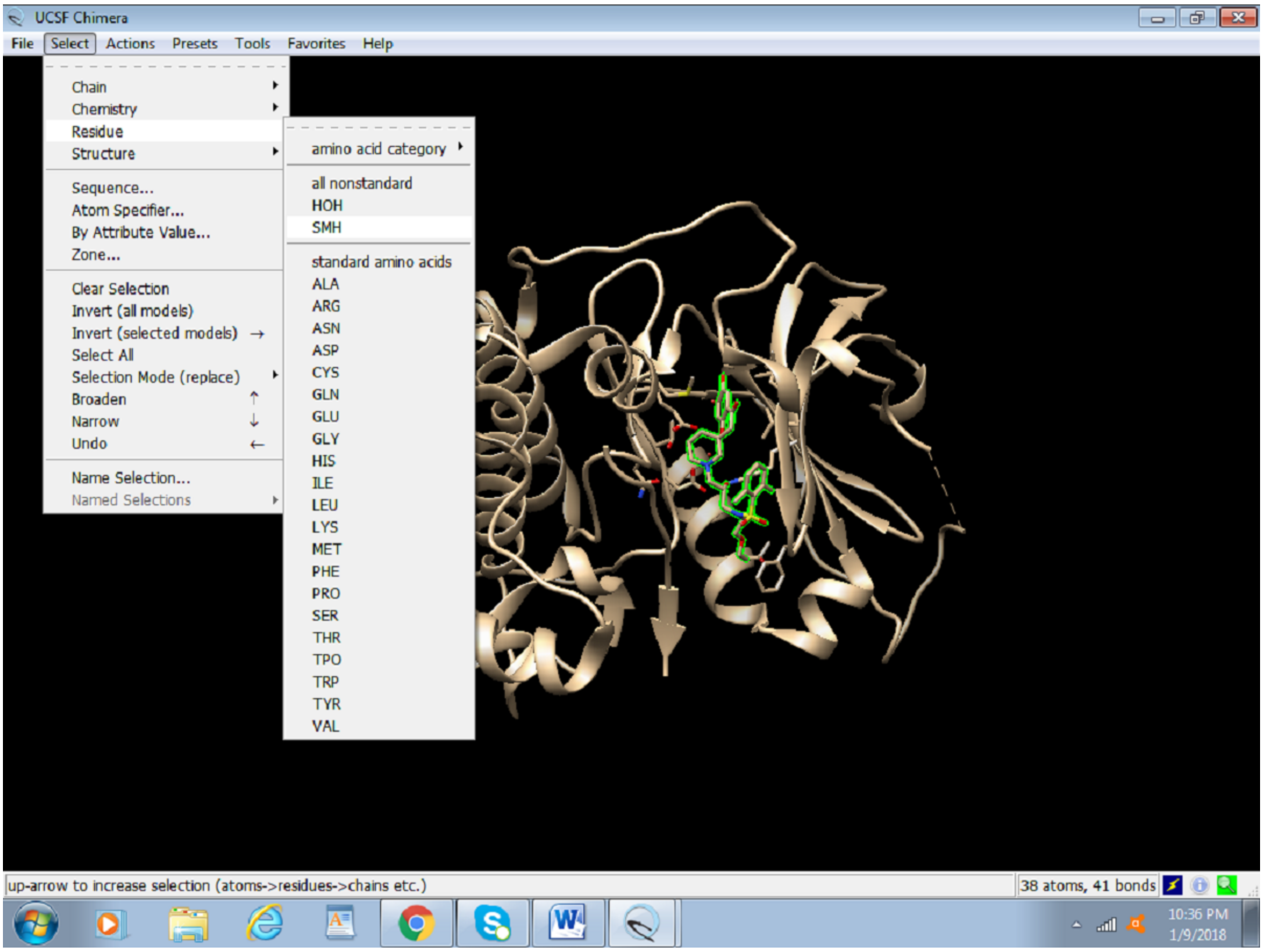


Opened 3qkk containing 1 model, 2914 atoms, and 459 residues

Preparing the Target Protein for Docking

1. To easily define the active site, the already present inhibitor needs to be identified. To do so, select the inhibitor by click on Select > Residue > SMH (nonstandard residue), as seen in Figure 4. In this screenshot, Akt bears an HOH group and SMH residues as nonstandard residues. Due to the selection, SMH appears to be highlighted in green.

Figure 4. Screenshot of selecting nonstandard residues.

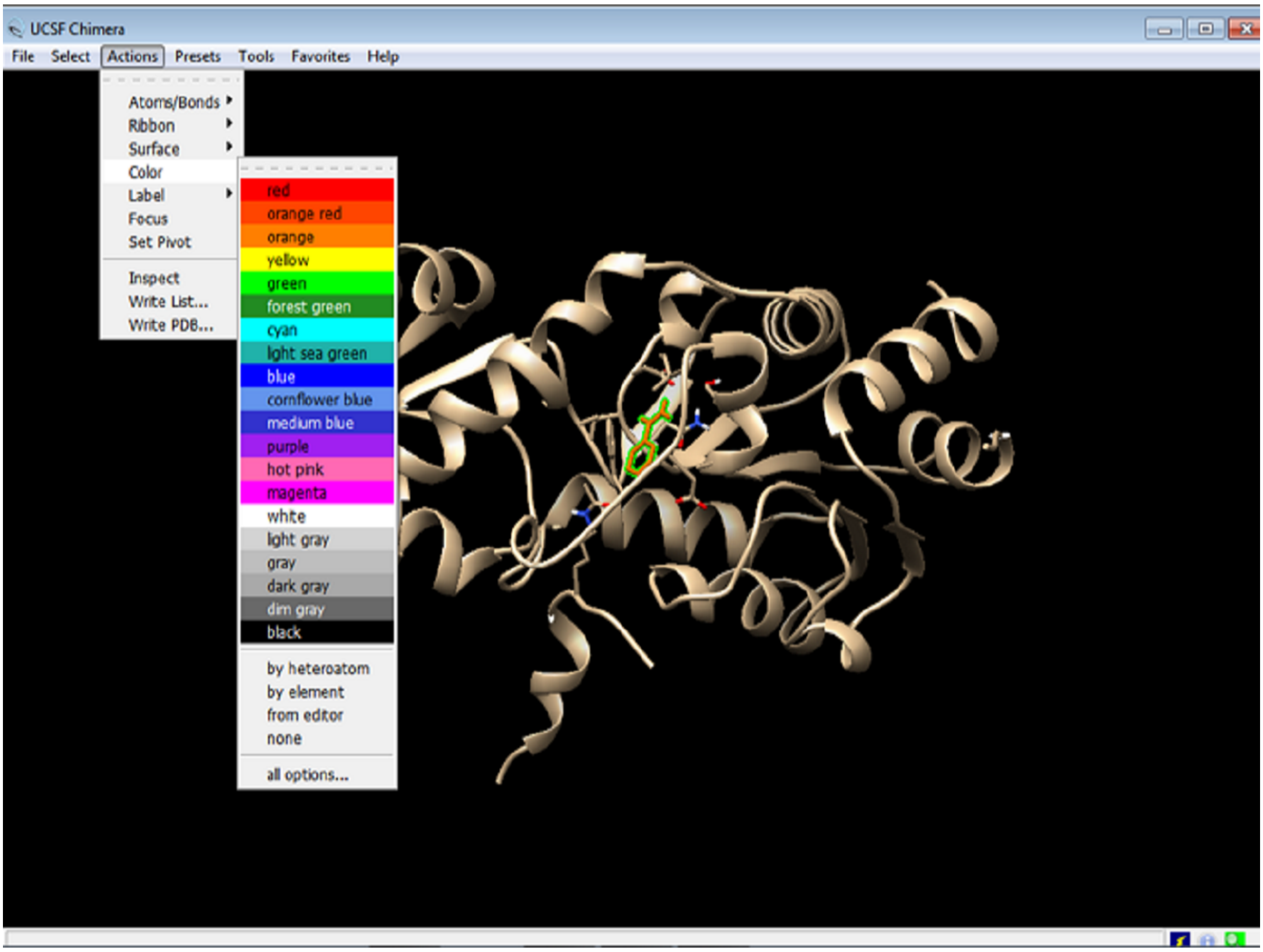


up-arrow to increase selection (atoms->residues->chains etc.)

38 atoms, 41 bonds

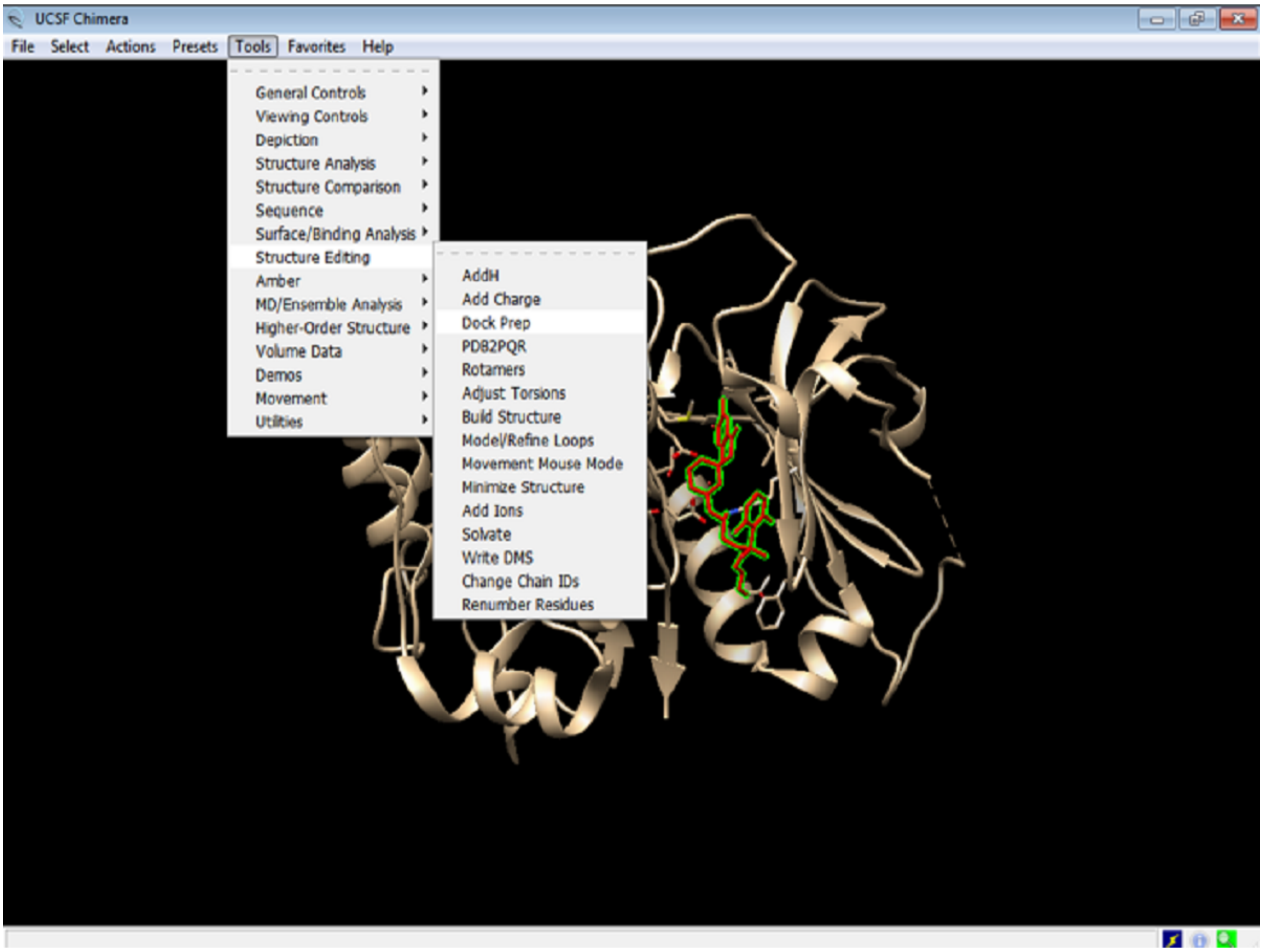
2. After selecting the nonstandard (inhibitor) residues, the residues must be accorded a color. To distinguish the chosen residue from the rest of the protein (Figure 5), change the color by clicking on Actions > Color > red (any color of your choice).

Figure 5. Screenshot of changing the color of the nonstandard (inhibitor) residues.



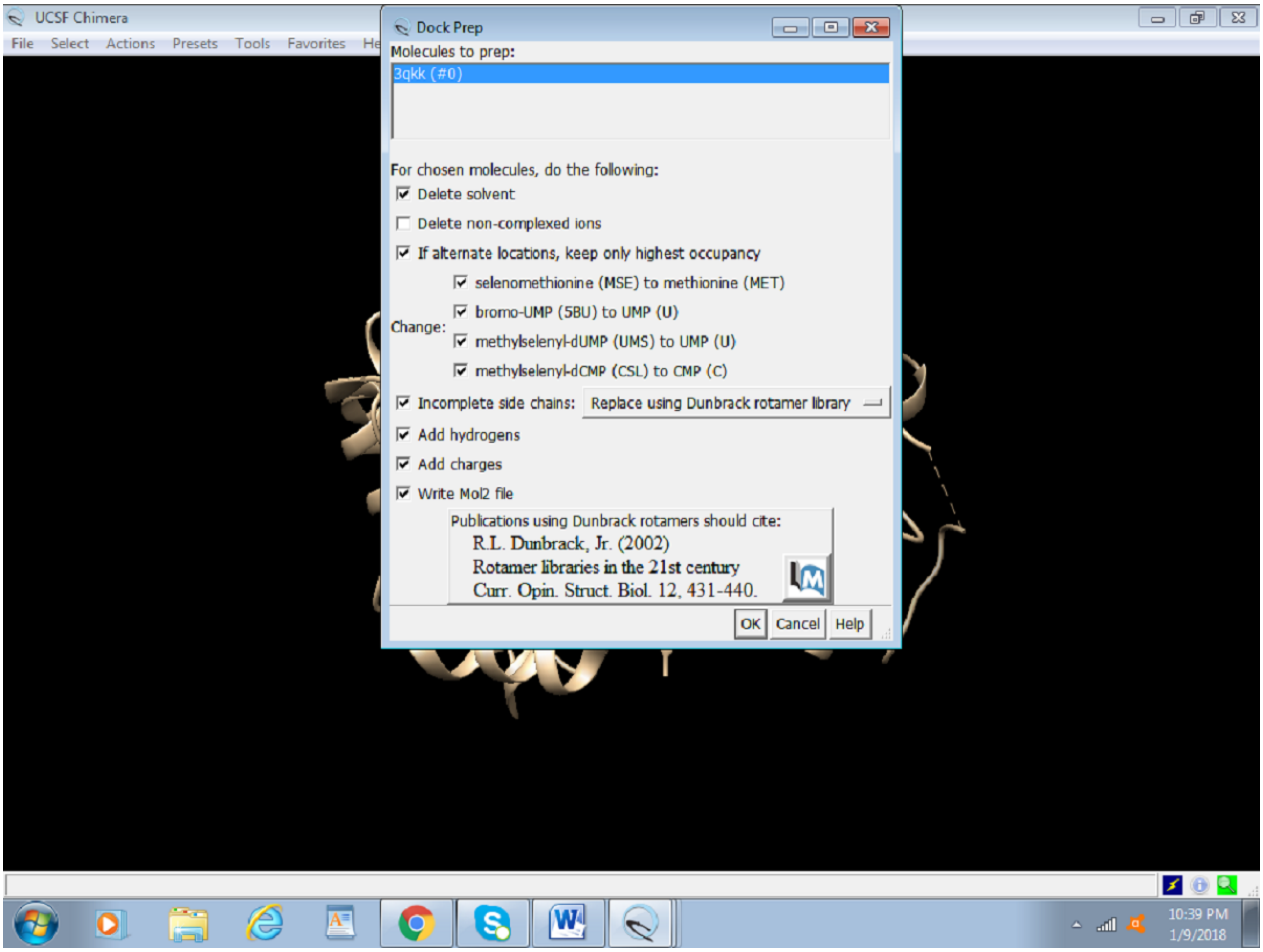
3. The protein needs to be optimized for docking. Click on Tools > Structure Editing > Dock Prep (Figure 6). The required dock prep tools are all available within Chimera. These dock prep tools are available in the structure editing file menu option.

Figure 6. Screenshot of an illustration of preparation of the protein for docking (ie, Dock Prep).



4. In the dock prep box, select all options except “Delete non-complexed ions” and click OK (Figure 7).

Figure 7. Screenshot of an example of the Dock Prep box that pops up.



Molecules to prep:
3qkk (#0)

For chosen molecules, do the following:

☑ Delete solvent
□ Delete non-complexed ions
☑ If alternate locations, keep only highest occupancy

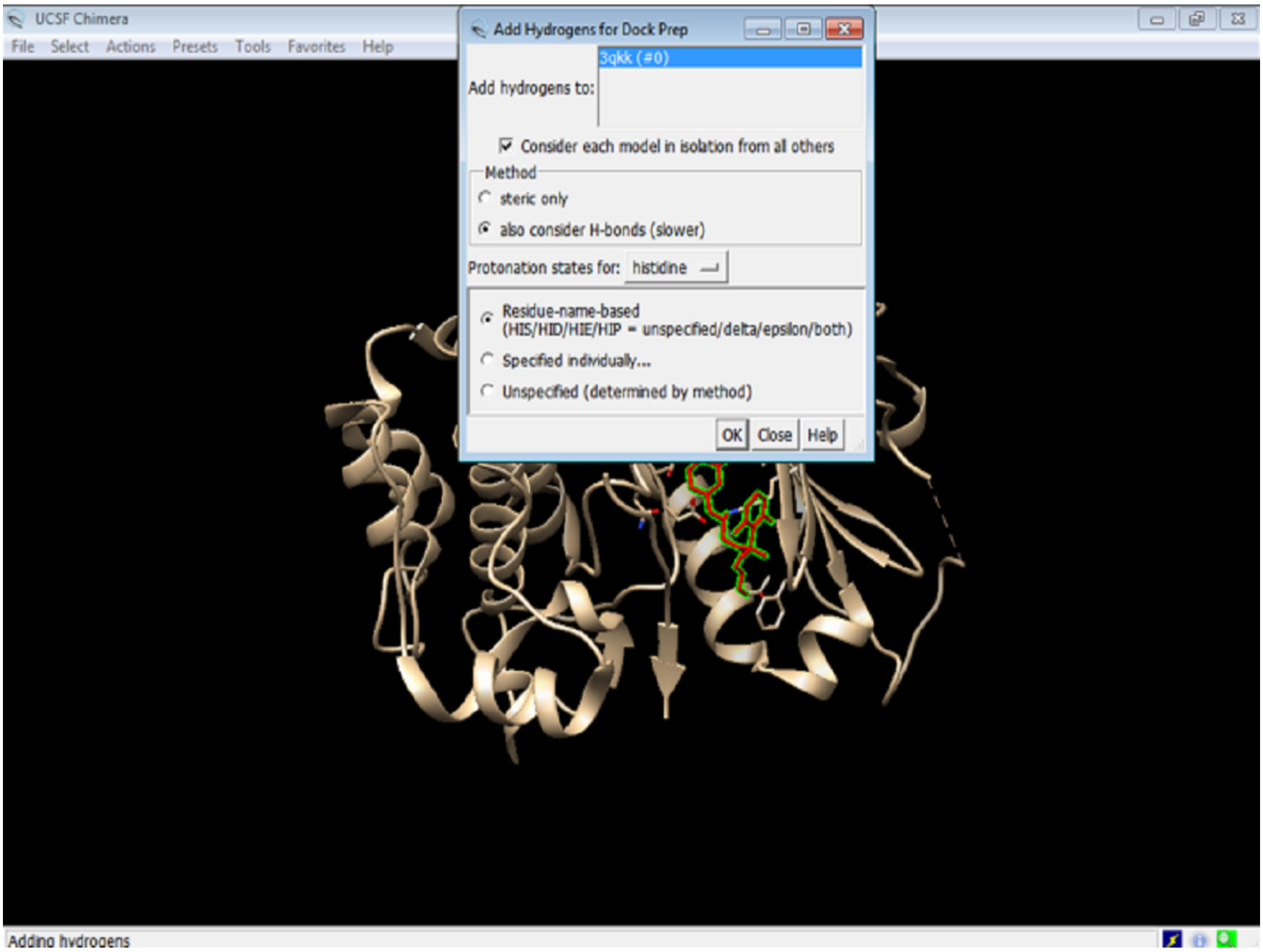
Change:
☑ selenomethionine (MSE) to methionine (MET)
☑ bromo-UMP (5BU) to UMP (U)
☑ methylselenyl-dUMP (UMS) to UMP (U)
☑ methylselenyl-dCMP (CSL) to CMP (C)

☑ Incomplete side chains: Replace using Dunbrack rotamer library
☑ Add hydrogens
☑ Add charges
☑ Write Mol2 file

Publications using Dunbrack rotamers should cite:
R.L. Dunbrack, Jr. (2002)
Rotamer libraries in the 21st century
Curr. Opin. Struct. Biol. 12, 431-440.

5. Add hydrogen to the proteins by selecting the appropriate following options and click OK (Figure 8). We allow the program to make the best choice according to the model by selecting the abovementioned options.

Figure 8. Screenshot of adding hydrogen atoms to the protein.



Add Hydrogens for Dock Prep

Add hydrogens to: 3qkk (#0)

☑ Consider each model in isolation from all others

Method

○ steric only

● also consider H-bonds (slower)

Protonation states for: histidine

● Residue-name-based (HIS/HID/HIE/HIP = unspecified/delta/epsilon/both)

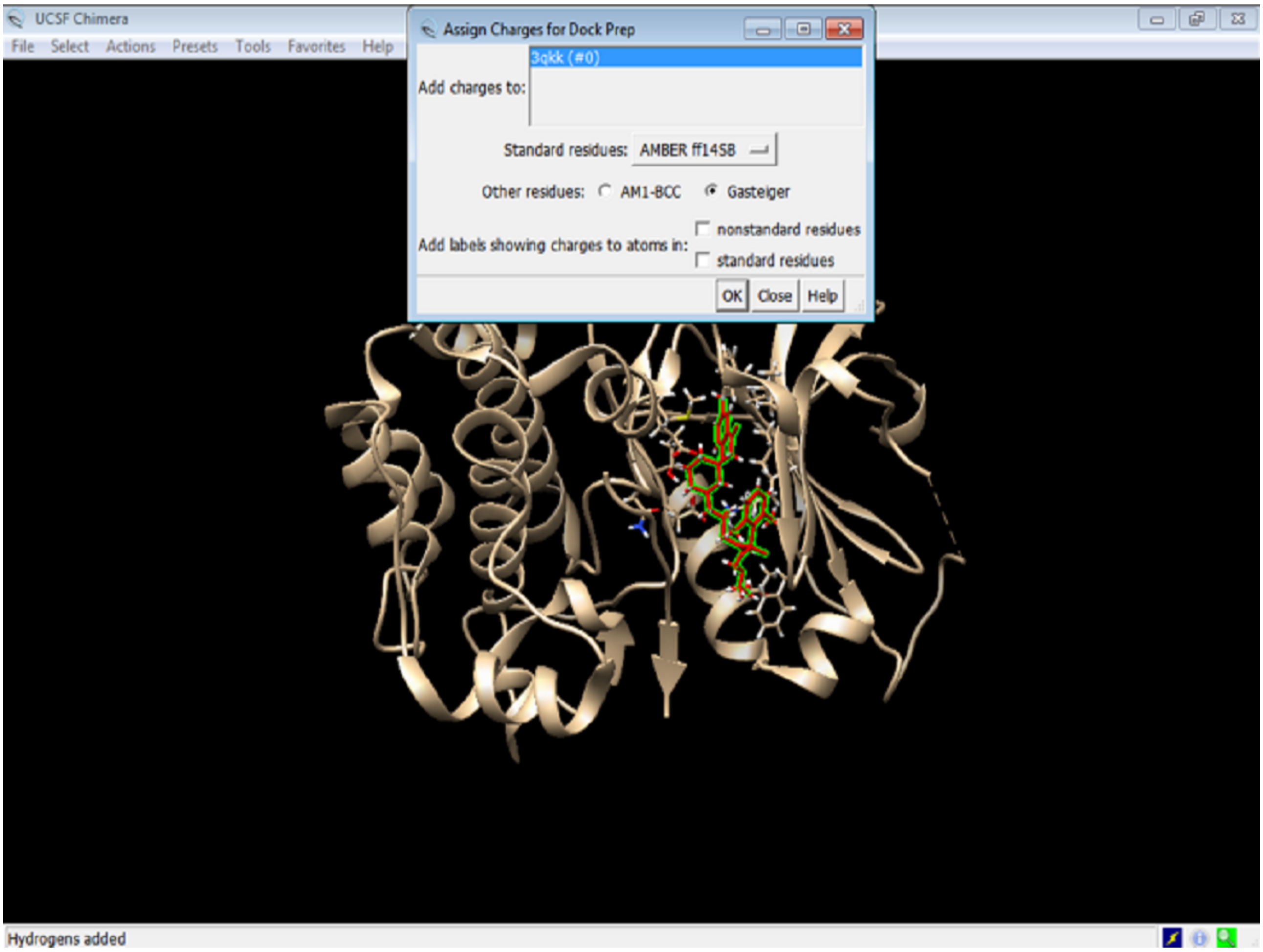
○ Specified individually...

○ Unspecified (determined by method)

Adding hydrogens

6. Assign charges to the protein by clicking on the Gasteiger charges (Figure 9) and click OK.

Figure 9. Screenshot of the selection of Gasteiger charges for Akt.



UCSF Chimera
File Select Actions Presets Tools Favorites Help
Assign Charges for Dock Prep
OK
Close
Help

Add charges to: 3qkk (#0)

Standard residues: AMBER ff14SB

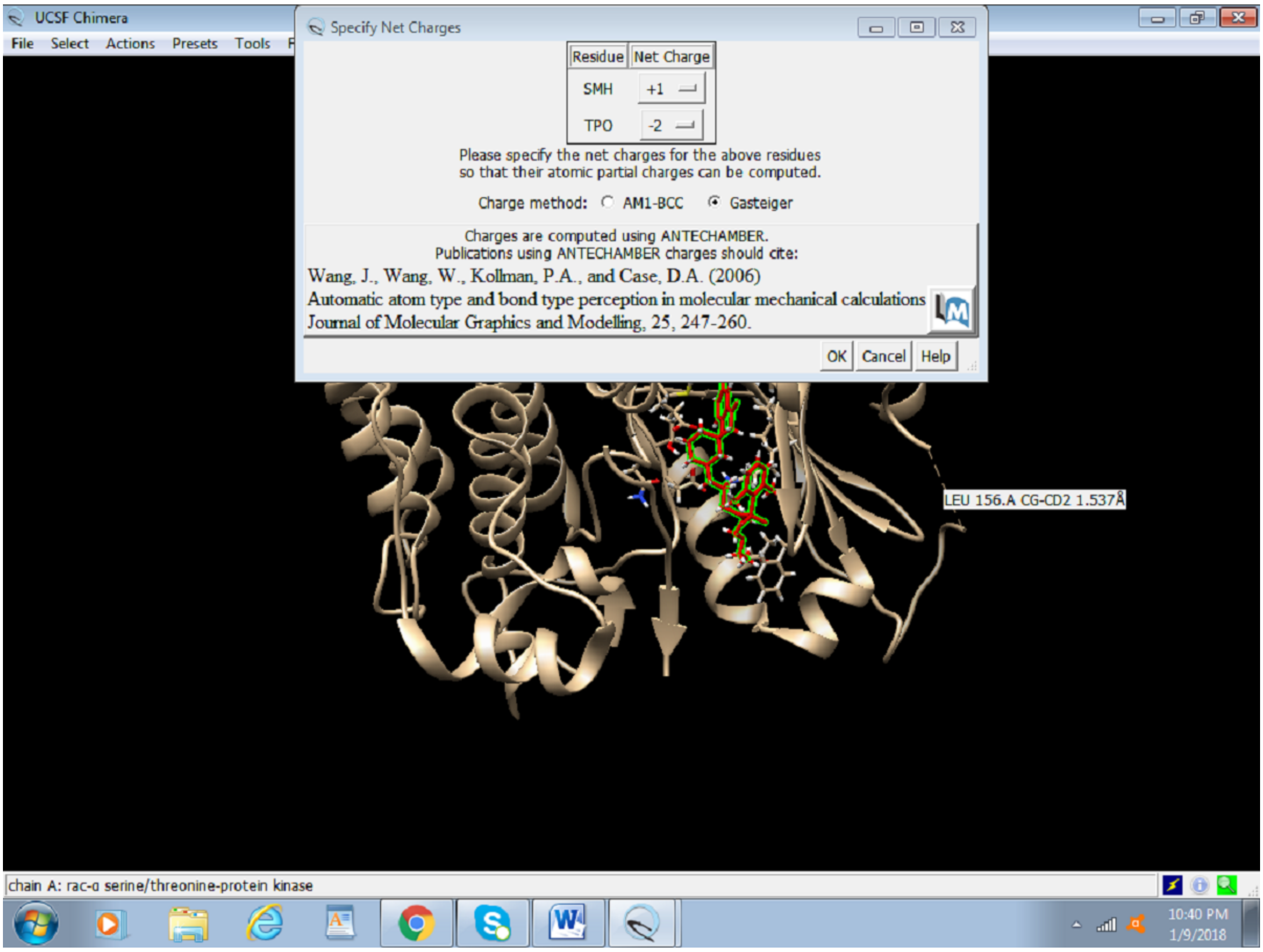
Other residues: ○ AM1-BCC ● Gasteiger

Add labels showing charges to atoms in: □ nonstandard residues □ standard residues

Hydrogens added

7. Select the net charges (Figure 10) and click OK. For most proteins, the net charges equal to zero.

Figure 10. Screenshot of the net charges of a protein.



Residue	Net Charge
SMH	+1
TPO	-2

Please specify the net charges for the above residues so that their atomic partial charges can be computed.

Charge method: ○ AM1-BCC ● Gasteiger

Charges are computed using ANTECHAMBER.
Publications using ANTECHAMBER charges should cite:
Wang, J., Wang, W., Kollman, P.A., and Case, D.A. (2006)
Automatic atom type and bond type perception in molecular mechanical calculations
Journal of Molecular Graphics and Modelling, 25, 247-260.

chain A: rac-α serine/threonine-protein kinase

8. Save this file again as preped_Akt.PDB.

Preparing the Ligand for Docking

Similar to the process of obtaining the protein, drugs with Pubchem compound ID (CID) can be fetched through the software with a working internet connection.

1. Click on Structure Editing > Build Structure > PubChem CID or you can even insert the simplified molecular-input line-entry system (SMILES) of the novel compound being used. Figure 11 shows how to fetch ligands from PubChem using its ID.
2. Enter the PubChem CID and click apply.
3. The ligand needs to be optimized as the protein was optimized. Click on Tools > Structure Editing > Dock Prep, and repeat the same steps followed for preparing the protein. These steps include removing solvents, adding hydrogens, and determining the charge. Figure 12 shows an overview of the dock prep for the ligand.
4. The ligand Fisetin is saved as prep_fisetin.mol2 file in the working directory earlier created (Figure 13).

Figure 11. Screenshot of fetching the ligand compound fisetin through its PubChem ID in Chimera.

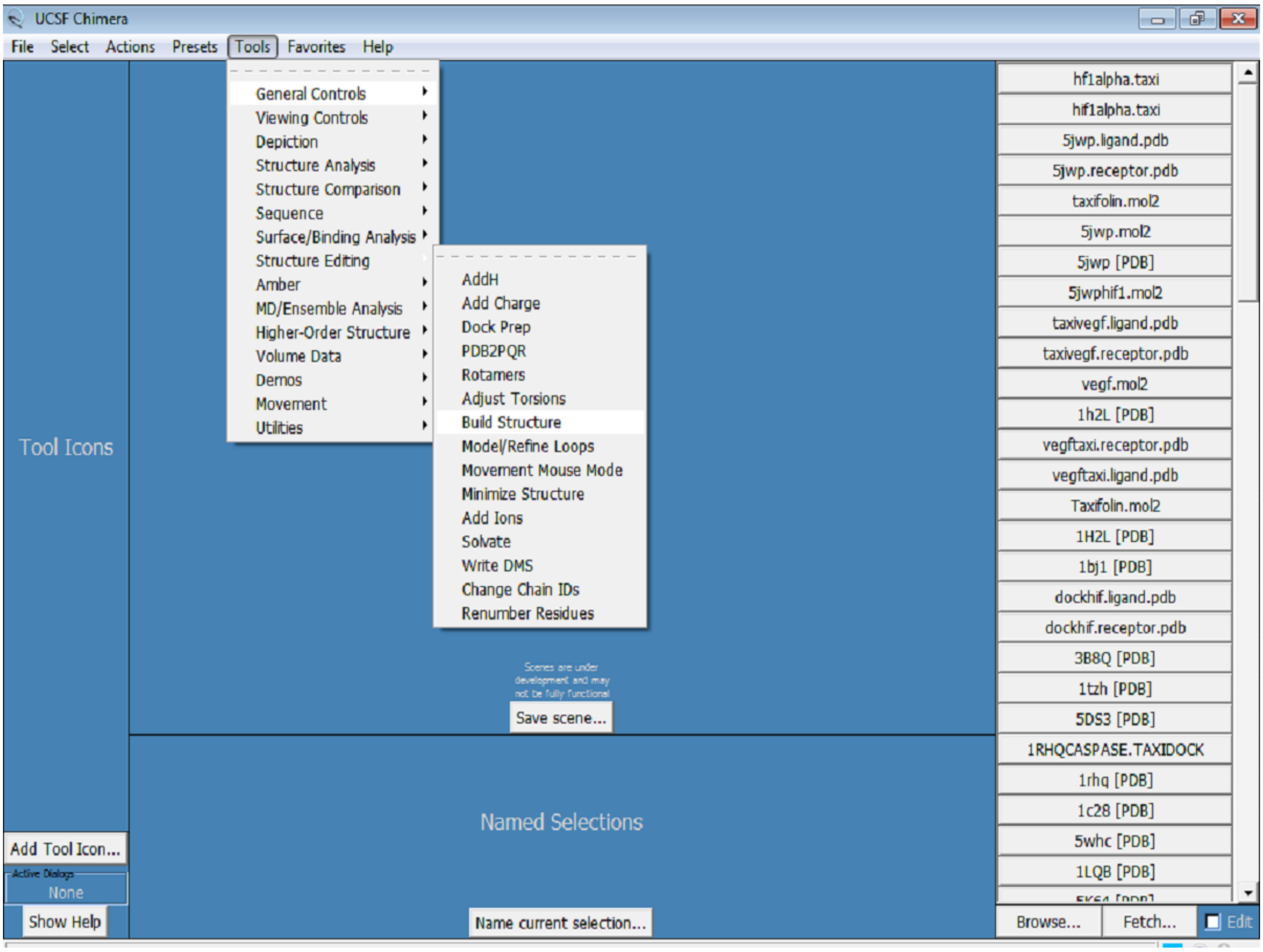
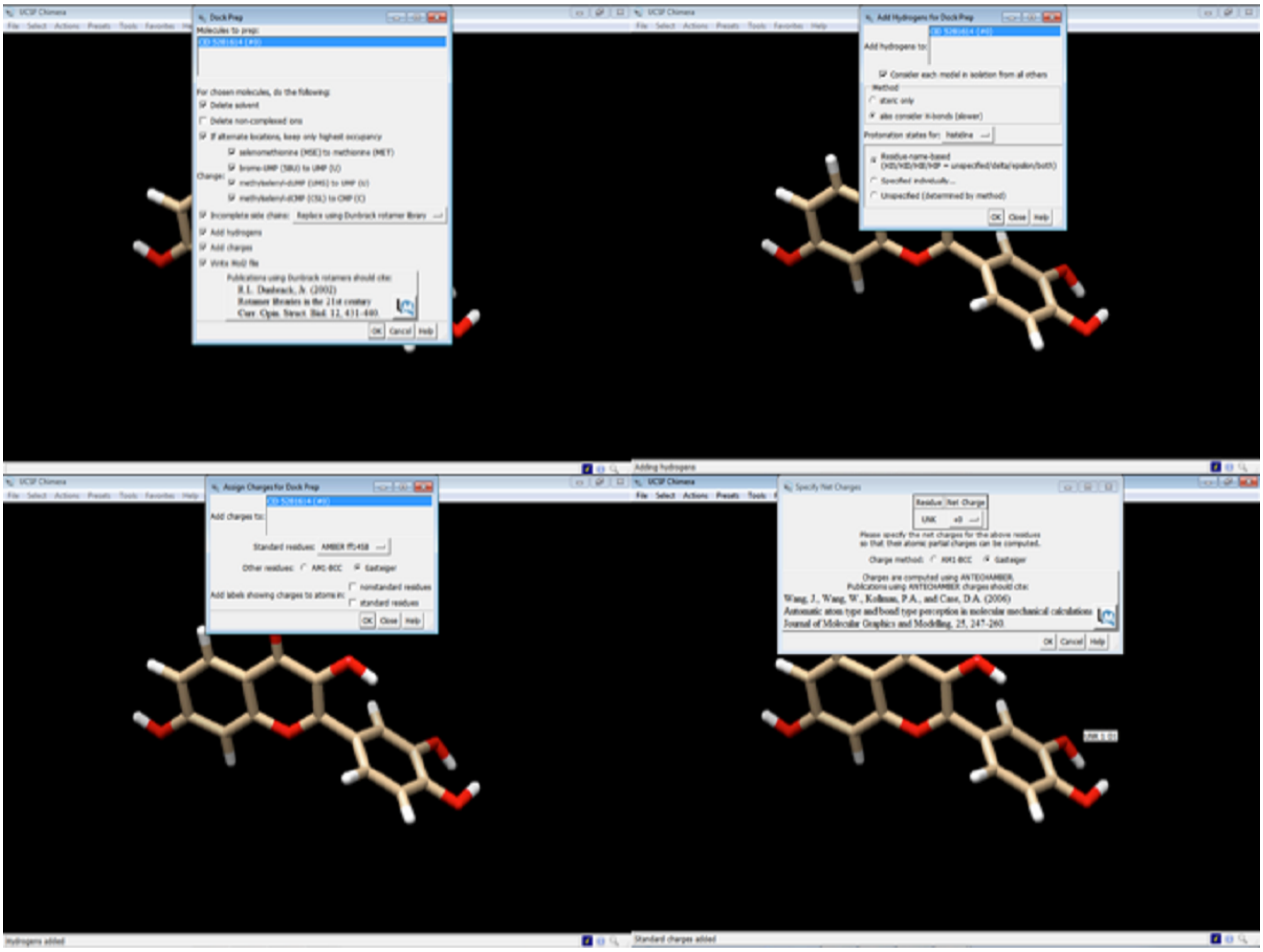


Figure 12. Screenshot of preparing the ligand for docking.



Dock Prep

Molecules to prep: CID 5281614 (#0)

For chosen molecules, do the following:

☑ Delete solvent

□ Delete non-complexed ions

☑ If alternate locations, keep only highest occupancy

Change:
☑ selenomethionine (MSE) to methionine (MET)
☑ bromo-UMP (5BU) to UMP (U)
☑ methylselenyl-dUMP (UMS) to UMP (U)
☑ methylselenyl-dCMP (CSL) to CMP (C)

☑ Incomplete side chains: Replace using Dunbrack rotamer library

☑ Add hydrogens

☑ Add charges

☑ Write Mol2 file

Publications using Dunbrack rotamers should cite:
R.L. Dunbrack, Jr. (2002)
Rotamer libraries in the 21st century
Curr. Opin. Struct. Biol. 12, 431-440

Add Hydrogens for Dock Prep

Add hydrogens to: CID 5281614 (#0)

☑ Consider each model in isolation from all others

Method
○ steric only
● also consider H-bonds (slower)

Protonation states for: histidine

● Residue-name-based
(HIS/HID/HIE/HIP = unspecified/delta/epsilon/both)
○ Specified individually...
○ Unspecified (determined by method)

Adding hydrogens

Assign Charges for Dock Prep

Add charges to: CID 5281614 (#0)

Standard residues: AMBER ff14SB

Other residues: ○ AM1-BCC ● Gasteiger

Add labels showing charges to atoms in: □ nonstandard residues □ standard residues

Hydrogens added

Specify Net Charges

Residue	Net Charge
UNK	+0

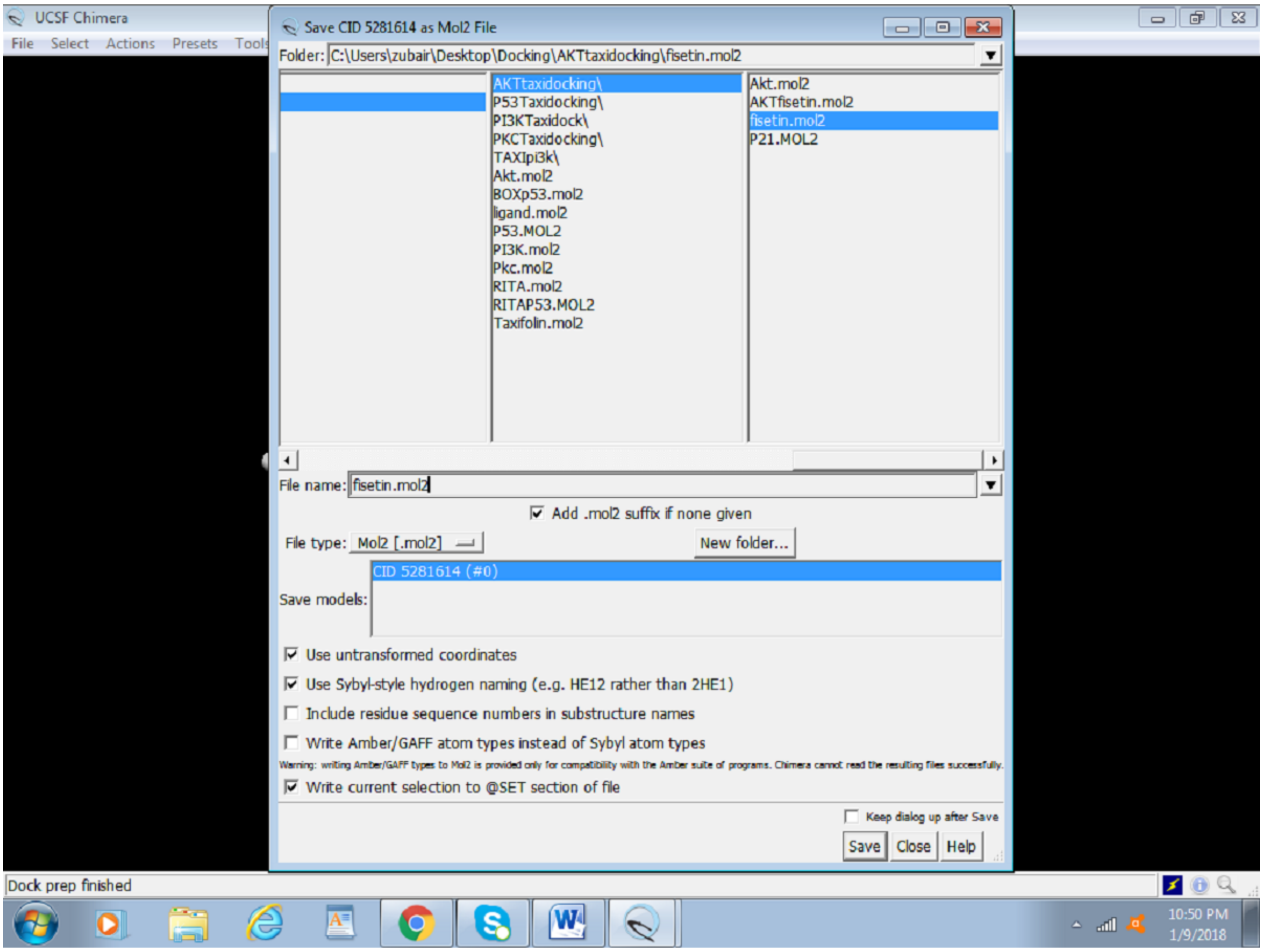
Please specify the net charges for the above residues
so that their atomic partial charges can be computed.

Charge method: ○ AM1-BCC ● Gasteiger

Charges are computed using ANTECHAMBER.
Publications using ANTECHAMBER charges should cite:
Wang, J., Wang, W., Kollman, P.A., and Case, D.A. (2006)
Automatic atom type and bond type perception in molecular mechanical calculations
Journal of Molecular Graphics and Modelling, 25, 247-260

Standard charges added

Figure 13. Screenshot illustrating the location of the ligand Mol2 file.



Save CID 5281614 as Mol2 File

Folder: C:\Users\zubair\Desktop\Docking\AKTtaxidocking\fisetin.mol2

AKTtaxidocking\
P53Taxidock\
PI3KTaxidock\
PKCTaxidocking\
TAXIpi3k\
Akt.mol2
BOXp53.mol2
ligand.mol2
P53.MOL2
PI3K.mol2
Pkc.mol2
RITA.mol2
RITAP53.MOL2
Taxifolin.mol2

Akt.mol2
AKTfisetin.mol2
fisetin.mol2
P21.MOL2

File name: fisetin.mol2

☑ Add .mol2 suffix if none given

File type: Mol2 [.mol2]

Save models: CID 5281614 (#0)

☑ Use untransformed coordinates

☑ Use Sybyl-style hydrogen naming (e.g. HE12 rather than 2HE1)

☐ Include residue sequence numbers in substructure names

☐ Write Amber/GAFF atom types instead of Sybyl atom types

Warning: writing Amber/GAFF types to Mol2 is provided only for compatibility with the Amber suite of programs. Chimera cannot read the resulting files successfully.

☑ Write current selection to @SET section of file

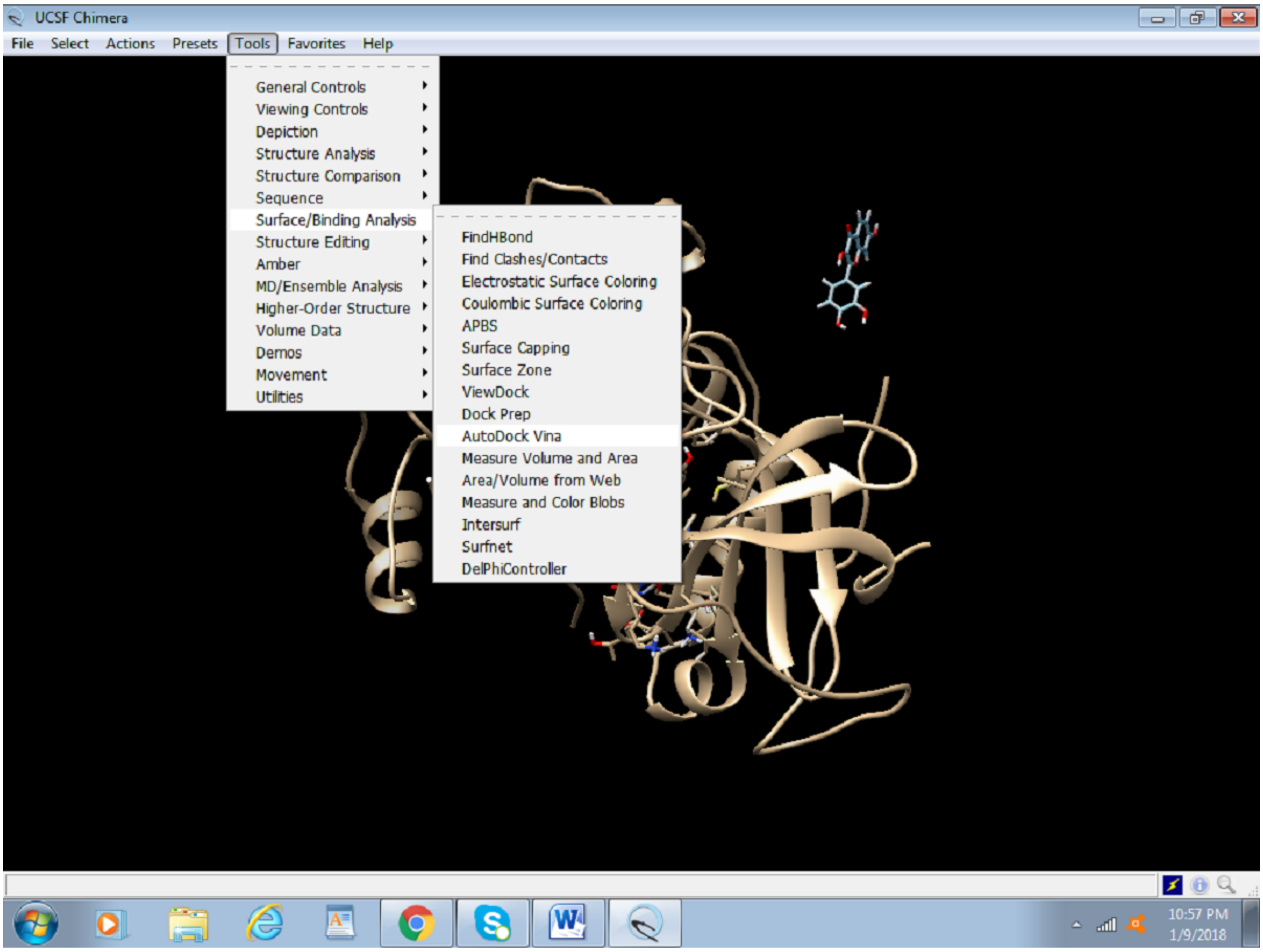
☐ Keep dialog up after Save

Docking

The following steps outline the process for docking:

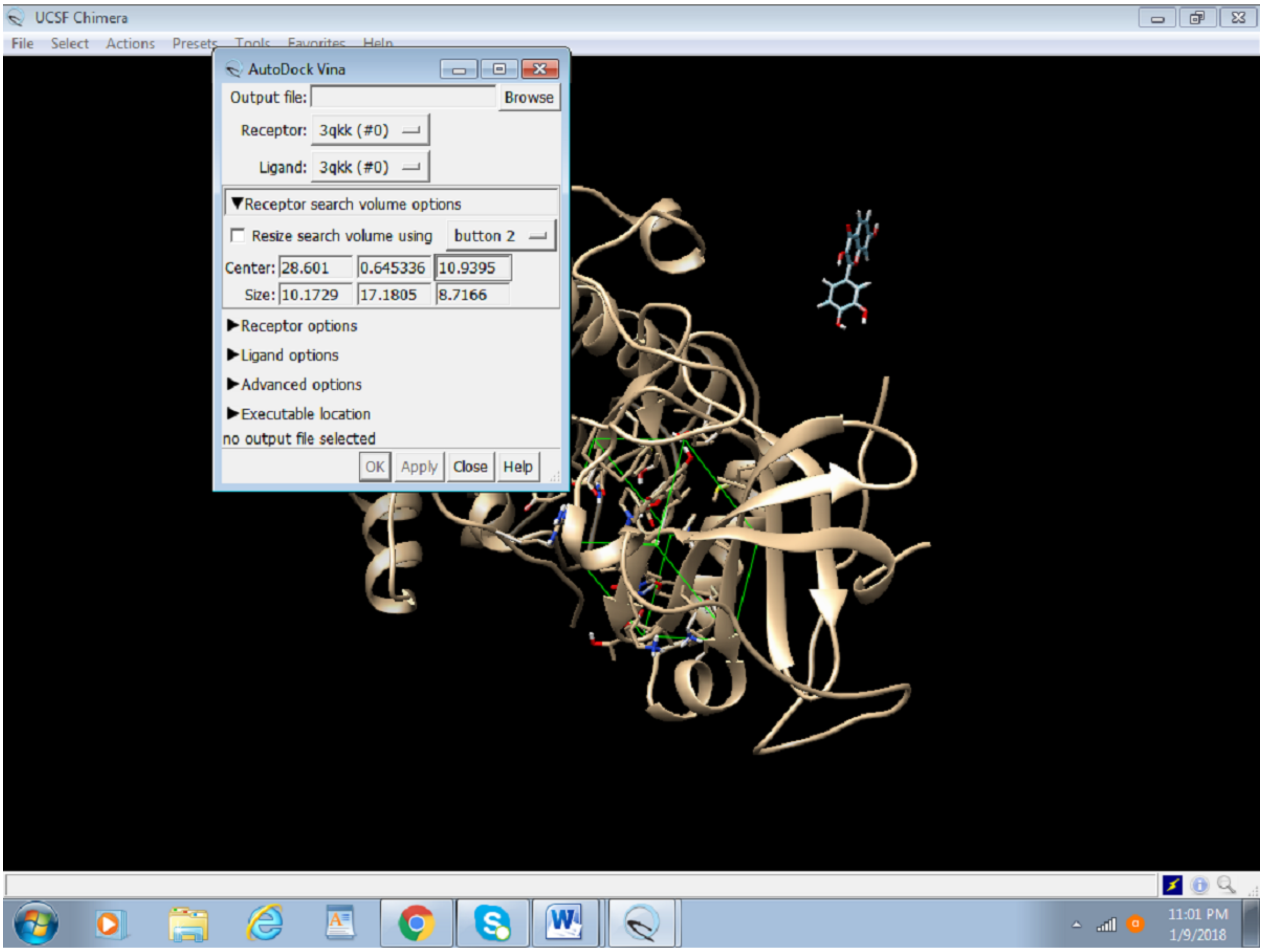
1. Click on Tools > Surface or Binding Analysis > Autodock Vina (Figure 14).

Figure 14. Screenshot of the process to access the Autodock Vina tool in Chimera.



2. We will set up the grid box values on the active site; this is usually where the previous inhibitor was present. In case an inhibitor is absent or the active site is relatively unknown, the size of the box and the location of the amino acids are determined by reading the literature (Figure 15). For the purpose of this protocol, we will use the active site that already had an inhibitor attached to it.

Figure 15. Screenshot of configuring the grid box values in Chimera.



Output file:

Receptor: 3qkk (#0)

Ligand: 3qkk (#0)

□ Resize search volume using button 2

Center: 28.601 0.645336 10.9395

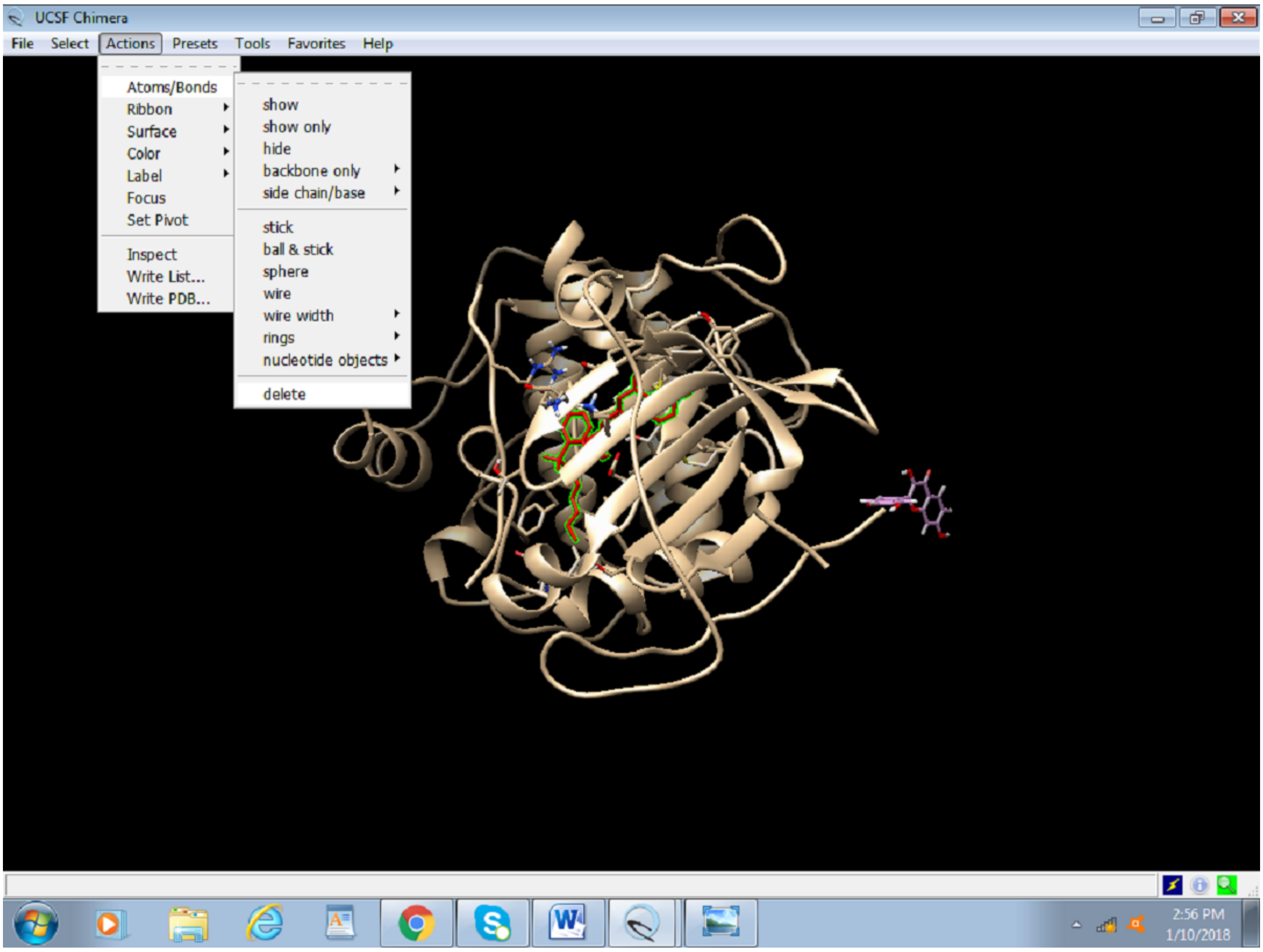
Size: 10.1729 17.1805 8.7166

no output file selected

3. Browse the output file and save as Akt Fisetin.pdbqt in the same directory.

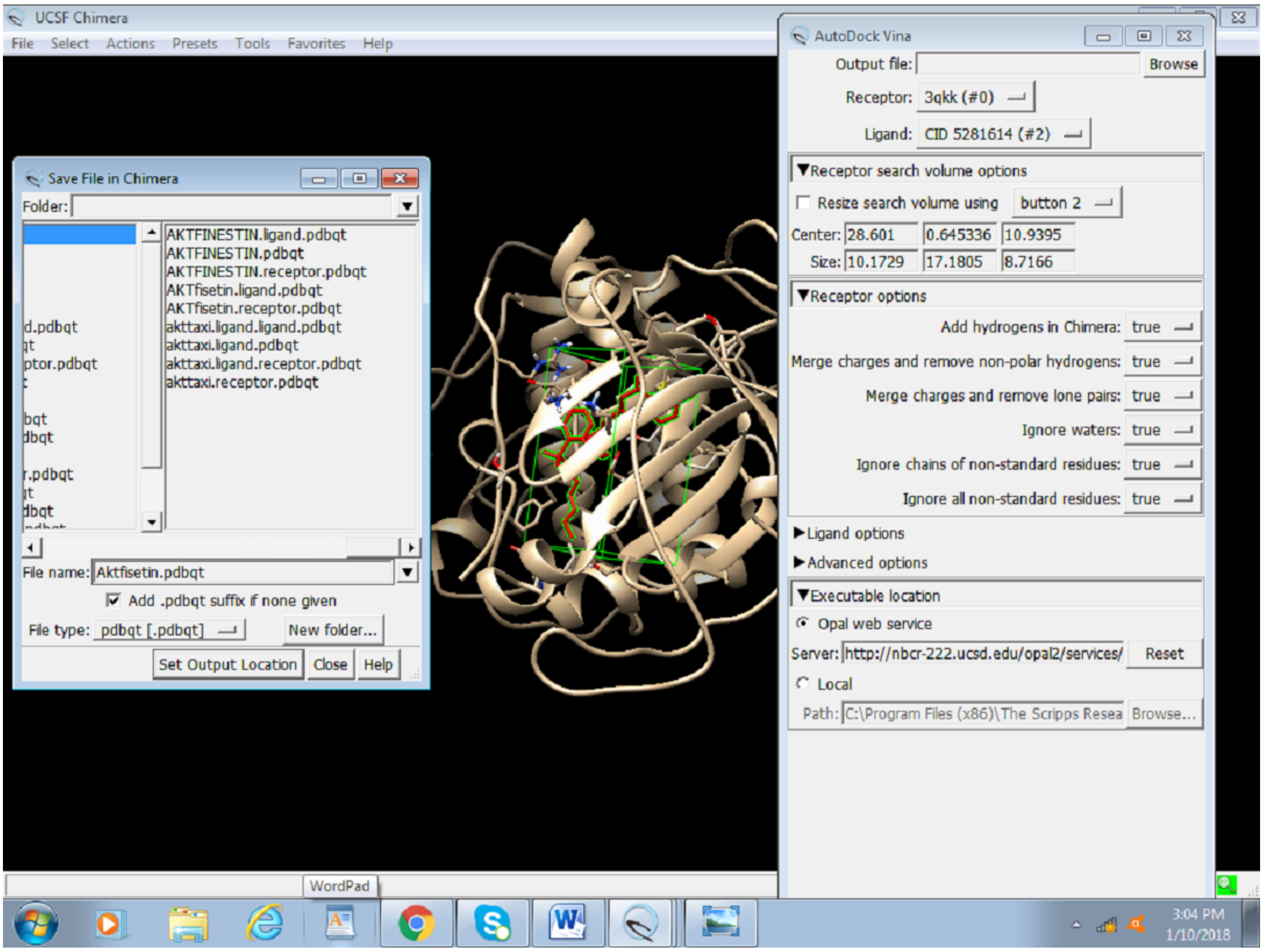
4. Delete the inhibitor molecule attached to the original 3D structure. Thereafter, select Actions > Atoms and Bonds > Delete (Figure 16). The removal of the inhibitor is important to easily visualize the docking results. The 3QKK PDB needs to be saved again as preped_Akt.PDB.

Figure 16. Screenshot of deletion of the inhibitor that is bound to the protein.



5. Choose the receptor as the protein (preped_Akt) from the drop-down menu and the ligand as prep_fisetin.mol2. It is important to set the right receptor and ligand. In the receptor and ligand options, change everything to TRUE (Figure 17).

Figure 17. Screenshot of the receptor and ligand options configuration in Autodock Vina.



Save File in Chimera

Folder:

AKTFINESTIN.ligand.pdbqt
AKTFINESTIN.pdbqt
AKTFINESTIN.receptor.pdbqt
AKTfisetin.ligand.pdbqt
AKTfisetin.receptor.pdbqt
akttaxi.ligand.ligand.pdbqt
akttaxi.ligand.pdbqt
akttaxi.ligand.receptor.pdbqt
akttaxi.receptor.pdbqt

File name: Aktfisetin.pdbqt

☑ Add .pdbqt suffix if none given

File type: pdbqt [.pdbqt]

AutoDock Vina

Output file:

Receptor: 3qkk (#0)

Ligand: CID 5281614 (#2)

▼Receptor search volume options

□ Resize search volume using button 2

Center: 28.601 0.645336 10.9395

Size: 10.1729 17.1805 8.7166

▼Receptor options

Add hydrogens in Chimera: true

Merge charges and remove non-polar hydrogens: true

Merge charges and remove lone pairs: true

Ignore waters: true

Ignore chains of non-standard residues: true

Ignore all non-standard residues: true

▼Executable location

● Opal web service

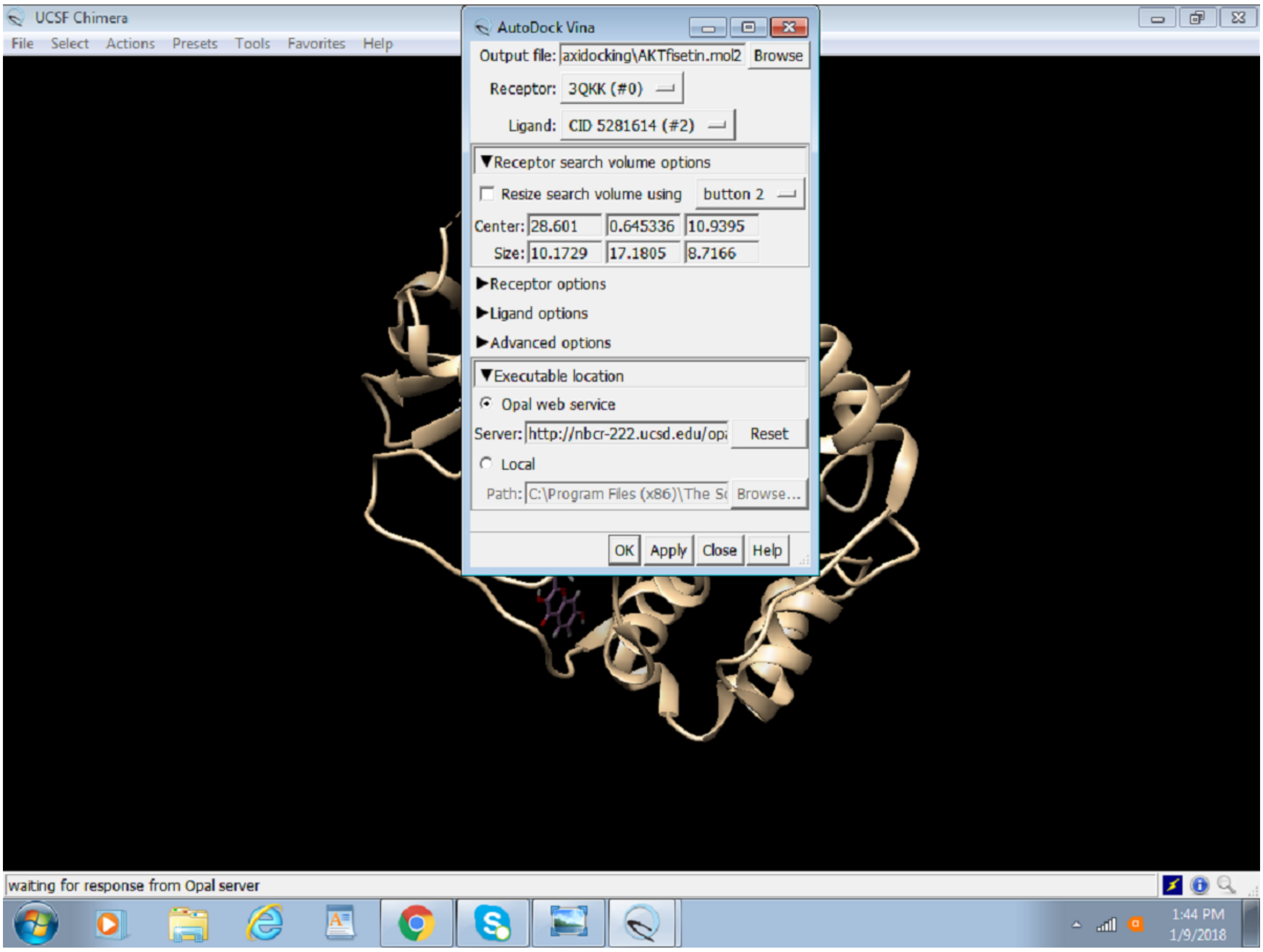
Server: http://nbcr-222.ucsd.edu/opal2/services/

○ Local

Path: C:\Program Files (x86)\The Scripps Resea

6. Select the Opal Web service or enter the local path where the installed version of Autodock Vina is placed and click on OK (Figure 18).

Figure 18. Selection of the Opal web service app in Chimera.



Output file: axidocking\AKTfisetin.mol2

Receptor: 3QKK (#0)

Ligand: CID 5281614 (#2)

□ Resize search volume using button 2

Center: 28.601 0.645336 10.9395

Size: 10.1729 17.1805 8.7166

● Opal web service

Server: http://nbcr-222.ucsd.edu/op:

○ Local

Path: C:\Program Files (x86)\The S(

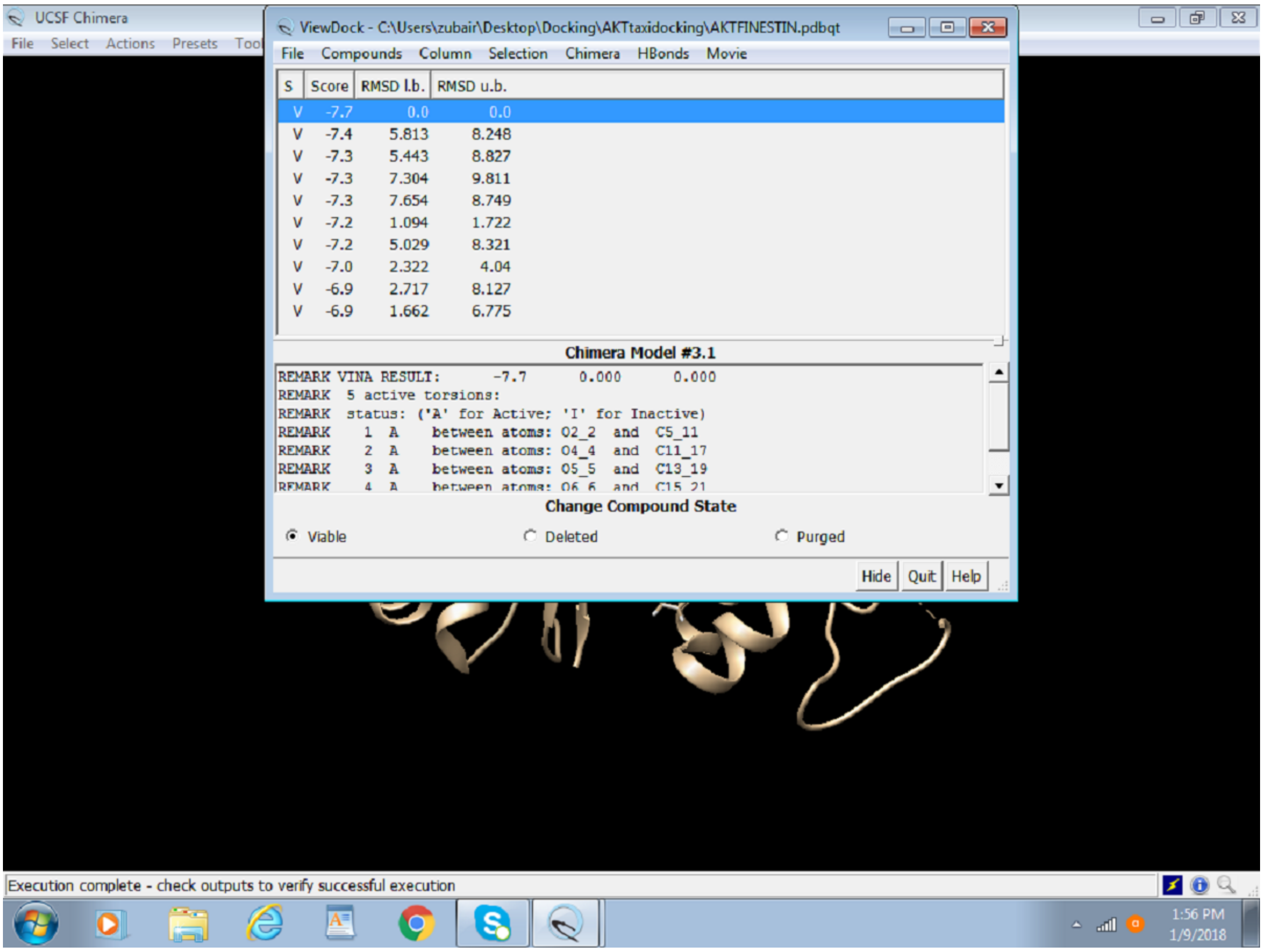
waiting for response from Opal server

Results

Outcome of Docking

After the successful run of Autodock Vina, the following dialogue box will appear with the solution. Figure 19 portrays the final step of Docking, that is, outcome/results of docking, which are score, root-mean-square deviation (RMSD) lower bound, and RMSD upper bound.

Figure 19. Screenshot of the Result Box after completion of docking.



UCSF Chimera
File Select Actions Presets Tool
ViewDock - C:\Users\zubair\Desktop\Docking\AKTtaxidocking\AKTFINESTIN.pdbqt
File Compounds Column Selection Chimera HBonds Movie
Hide Quit Help

S	Score	RMSD l.b.	RMSD u.b.
V	-7.7	0.0	0.0
V	-7.4	5.813	8.248
V	-7.3	5.443	8.827
V	-7.3	7.304	9.811
V	-7.3	7.654	8.749
V	-7.2	1.094	1.722
V	-7.2	5.029	8.321
V	-7.0	2.322	4.04
V	-6.9	2.717	8.127
V	-6.9	1.662	6.775

Chimera Model #3.1

```
REMARK VINA RESULT:      -7.7      0.000      0.000
REMARK  5 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: O2_2  and  C5_11
REMARK    2  A    between atoms: O4_4  and  C11_17
REMARK    3  A    between atoms: O5_5  and  C13_19
REMARK    4  A    between atoms: O6_6  and  C15_21
```

Change Compound State

● Viable ○ Deleted ○ Purged

Execution complete - check outputs to verify successful execution

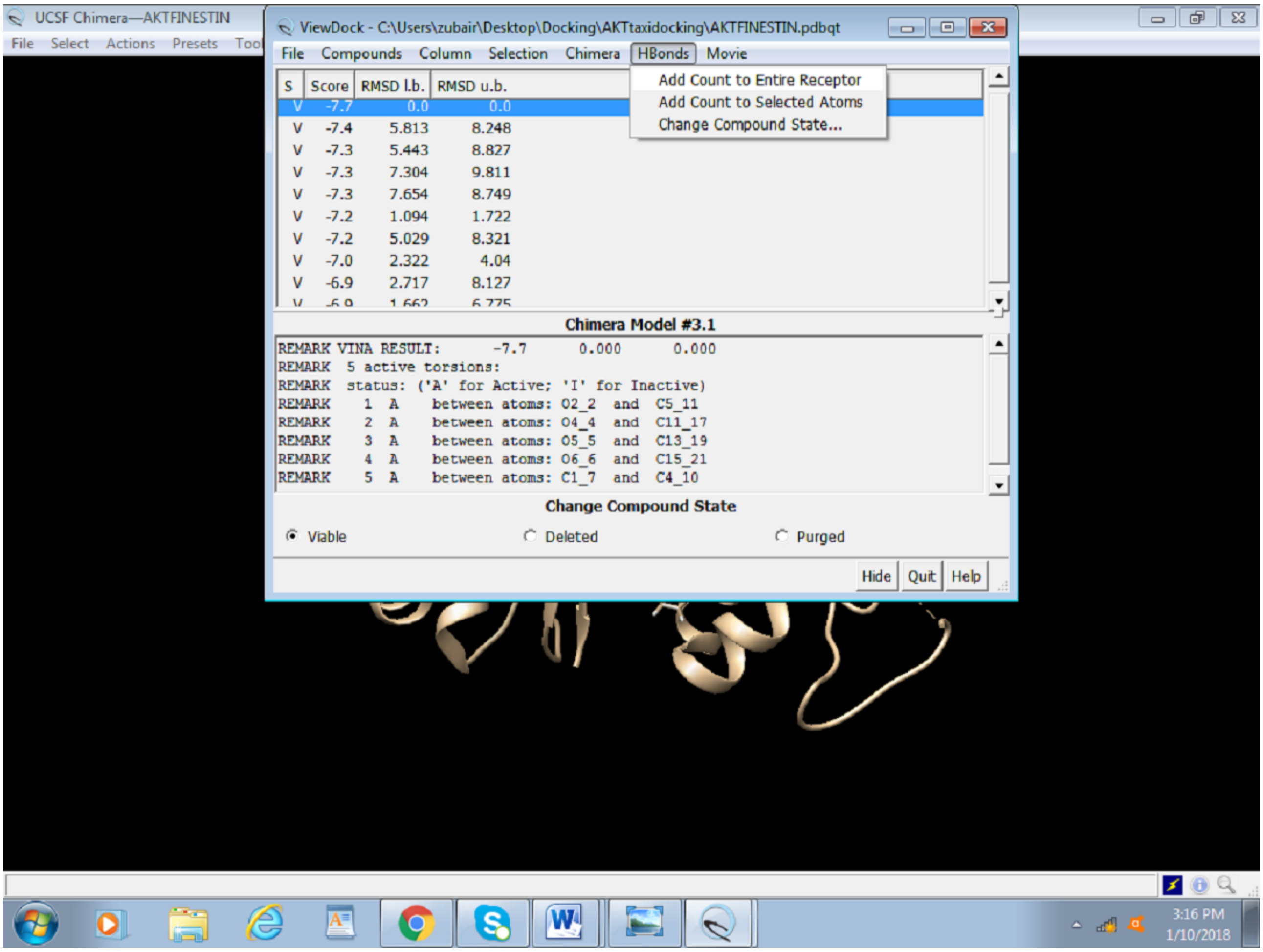
1:56 PM
1/9/2018

Visualizing the Docking

Visualization of docking can be done as follows:

1. To see the hydrogen bonding between the receptor and the ligands using the result dialogue box (Figure 20), select H Bonds > Add Count to the Entire Receptor.

Figure 20. Screenshot of visualizing hydrogen bonding between the receptor and ligand.


UCSF Chimera—AKTFINESTIN
File Select Actions Presets Tool
ViewDock - C:\Users\zubair\Desktop\Docking\AKTtaxidocking\AKTFINESTIN.pdbqt
File Compounds Column Selection Chimera HBonds Movie
Add Count to Entire Receptor
Add Count to Selected Atoms
Change Compound State...
3:16 PM
1/10/2018

S	Score	RMSD l.b.	RMSD u.b.
V	-7.7	0.0	0.0
V	-7.4	5.813	8.248
V	-7.3	5.443	8.827
V	-7.3	7.304	9.811
V	-7.3	7.654	8.749
V	-7.2	1.094	1.722
V	-7.2	5.029	8.321
V	-7.0	2.322	4.04
V	-6.9	2.717	8.127
V	-6.9	1.662	6.775

Chimera Model #3.1

```
REMARK VINA RESULT:      -7.7      0.000      0.000
REMARK  5 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: O2_2  and  C5_11
REMARK    2  A    between atoms: O4_4  and  C11_17
REMARK    3  A    between atoms: O5_5  and  C13_19
REMARK    4  A    between atoms: O6_6  and  C15_21
REMARK    5  A    between atoms: C1_7  and  C4_10
```

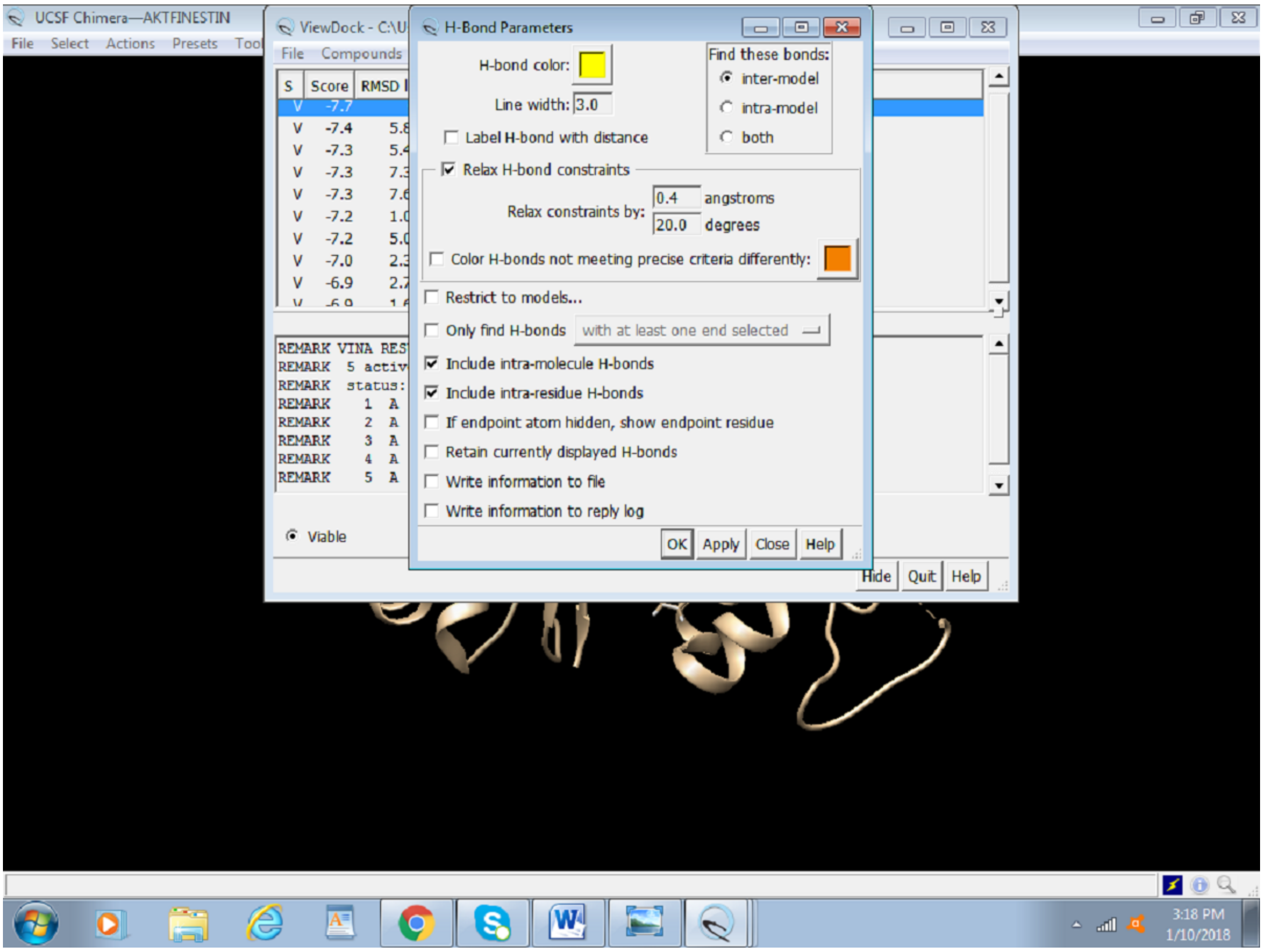
Change Compound State

● Viable ○ Deleted ○ Purged

Hide Quit Help

2. This opens an H-bond parameter dialogue box (Figure 21). Select Intermodel to visualize bonding between receptor and ligand. Different parameters can be adjusted to better picture the bonding. The table showing all the information on hydrogen bonds and RMSD is presented at the end of the docking session (Figure 22).

Figure 21. Screenshot of the H-bond Parameters dialogue box.



H-Bond Parameters

H-bond color: 🟨

Line width: 3.0

□ Label H-bond with distance

Find these bonds:
● inter-model
○ intra-model
○ both

☑ Relax H-bond constraints

Relax constraints by: 0.4 angstroms
20.0 degrees

□ Color H-bonds not meeting precise criteria differently: 🟧

□ Restrict to models...

□ Only find H-bonds with at least one end selected

☑ Include intra-molecule H-bonds

☑ Include intra-residue H-bonds

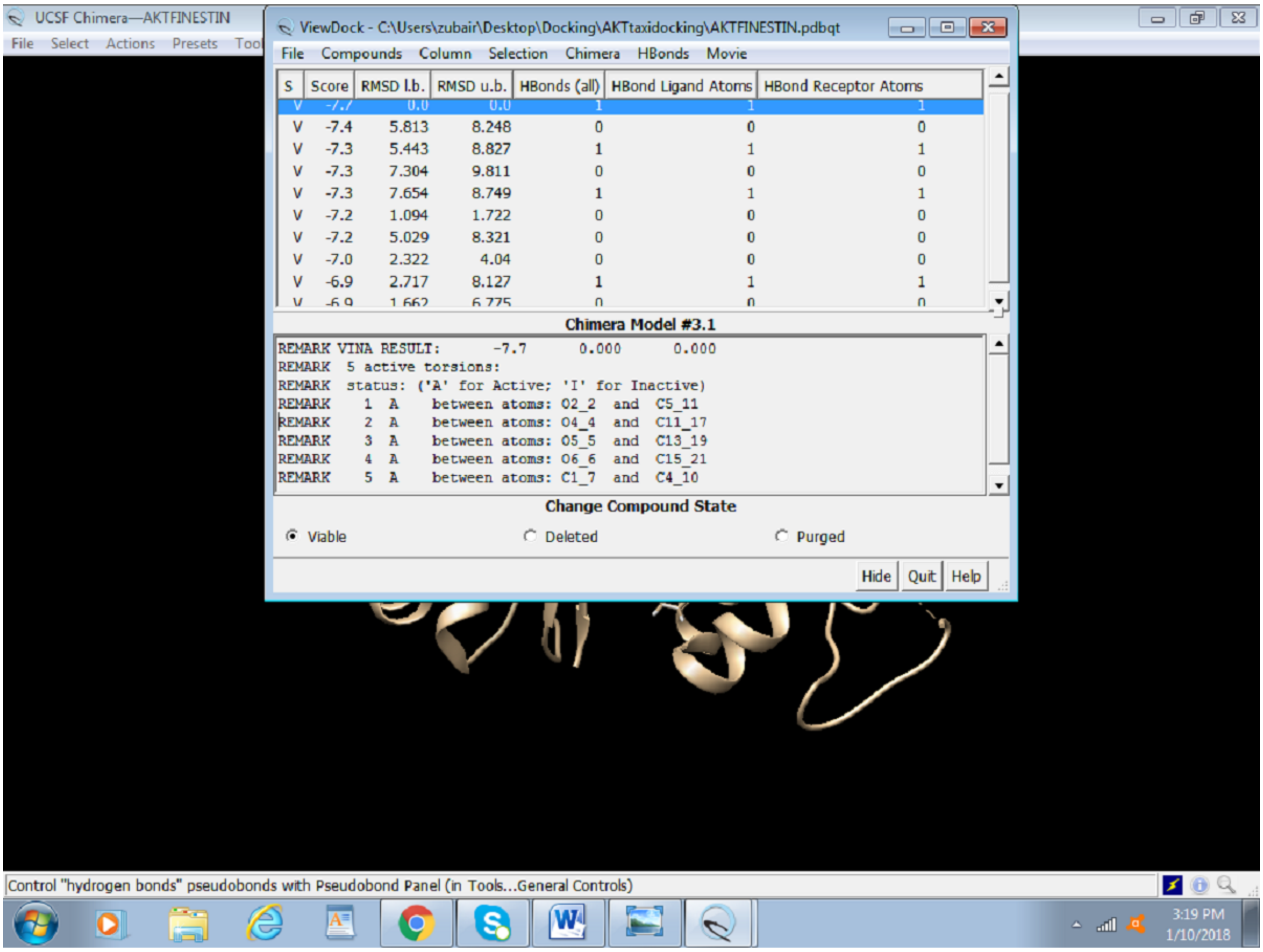
□ If endpoint atom hidden, show endpoint residue

□ Retain currently displayed H-bonds

□ Write information to file

□ Write information to reply log

Figure 22. Screenshot of the table showing the number of hydrogen bonds and root-mean-square deviation values.



UCSF Chimera—AKTFINESTIN
File Select Actions Presets Tool
ViewDock - C:\Users\zubair\Desktop\Docking\AKTtaxidocking\AKTFINESTIN.pdbqt
File Compounds Column Selection Chimera HBonds Movie
Control "hydrogen bonds" pseudobonds with Pseudobond Panel (in Tools...General Controls)
3:19 PM
1/10/2018

S	Score	RMSD l.b.	RMSD u.b.	HBonds (all)	HBond Ligand Atoms	HBond Receptor Atoms
V	-7.7	0.0	0.0	1	1	1
V	-7.4	5.813	8.248	0	0	0
V	-7.3	5.443	8.827	1	1	1
V	-7.3	7.304	9.811	0	0	0
V	-7.3	7.654	8.749	1	1	1
V	-7.2	1.094	1.722	0	0	0
V	-7.2	5.029	8.321	0	0	0
V	-7.0	2.322	4.04	0	0	0
V	-6.9	2.717	8.127	1	1	1
V	-6.9	1.662	6.775	0	0	0

Chimera Model #3.1

```
REMARK VINA RESULT:      -7.7      0.000      0.000
REMARK  5 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: O2_2  and  C5_11
REMARK    2  A    between atoms: O4_4  and  C11_17
REMARK    3  A    between atoms: O5_5  and  C13_19
REMARK    4  A    between atoms: O6_6  and  C15_21
REMARK    5  A    between atoms: C1_7  and  C4_10
```

Change Compound State

● Viable ○ Deleted ○ Purged

3. To be able to retrieve the docking session later at any stage, it can be saved by selecting File > Save Session as > An Akt Fisetin Docking (name the session).

Discussion

Computationally fast and accurate docking of a ligand with a target protein can be performed using AutoDock Vina in Chimera. This protocol will help researchers who are not able to use Autodock and Autodock Vina due to its command-line interface and do not have access to high-end software such as Gold Suite and Molecular Operating Environment to perform computational docking easily. The use of Chimera with Autodock Vina has not been demonstrated before, and due to the ease of the graphical user interface of Chimera, it can be a go-to tool for someone who is just starting to learn bioinformatics.

Conflicts of Interest

None declared.

References

1. Seeliger D, de Groot BL. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. J Comput Aided Mol Des 2010 May 17;24(5):417-422 [FREE Full text] [doi: 10.1007/s10822-010-9352-6] [Medline: 20401516]
2. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010 Jan 30;31(2):455-461 [FREE Full text] [doi: 10.1002/jcc.21334] [Medline: 19499576]
3. Goddard TD, Huang CC, Ferrin TE. Visualizing density maps with UCSF Chimera. J Struct Biol 2007 Jan;157(1):281-287. [doi: 10.1016/j.jsb.2006.06.010] [Medline: 16963278]
4. Sun X, Ma X, Li Q, Yang Y, Xu X, Sun J, et al. Anti-cancer effects of fisetin on mammary carcinoma cells via regulation of the PI3K/Akt/mTOR pathway: In vitro and in vivo studies. Int J Mol Med 2018 Aug 02;42(2):811-820 [FREE Full text] [doi: 10.3892/ijmm.2018.3654] [Medline: 29749427]
5. Zhang X, Tang N, Hadden TJ, Rishi AK. Akt, FoxO and regulation of apoptosis. Biochim Biophys Acta 2011 Nov;1813(11):1978-1986 [FREE Full text] [doi: 10.1016/j.bbamcr.2011.03.010] [Medline: 21440011]

6. Syed D, Adhami V, Khan M, Mukhtar H. Inhibition of Akt/mTOR signaling by the dietary flavonoid fisetin. Anticancer Agents Med Chem 2013 Sep 01;13(7):995-1001 [FREE Full text] [doi: 10.2174/1871520611313990129] [Medline: 23293889]
7. Liao Y, Shih Y, Chao C, Lee X, Chiang T. Involvement of the ERK signaling pathway in fisetin reduces invasion and migration in the human lung cancer cell line A549. J Agric Food Chem 2009 Oct 14;57(19):8933-8941. [doi: 10.1021/jf902630w] [Medline: 19725538]
8. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol 2003 Dec;10(12):980 [FREE Full text] [doi: 10.1038/nsb1203-980] [Medline: 14634627]
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000 Jan 01;28(1):235-242 [FREE Full text] [doi: 10.1093/nar/28.1.235] [Medline: 10592235]

Abbreviations

CID: compound ID
PDB: Protein Data Bank
RMSD: root-mean-square deviation
SMILES: simplified molecular-input line-entry system
UCSF: University of California, San Francisco

Edited by G Eysenbach; submitted 03.04.19; peer-reviewed by E Dilshad, H Liu; comments to author 19.07.19; revised version received 09.09.19; accepted 24.09.19; published 19.06.20.

Please cite as:
Butt SS, Badshah Y, Shabbir M, Rafiq M
Molecular Docking Using Chimera and Autodock Vina Software for Nonbioinformaticians
JMIR Bioinformatics Biotechnol 2020;1(1):e14232
URL: https://bioinform.jmir.org/2020/1/e14232
doi:10.2196/14232

Original Paper

Structural Basis for Designing Multiepitope Vaccines Against COVID-19 Infection: In Silico Vaccine Design and Validation

Sukrit Srivastava[1,2], PhD; Sonia Verma[3], MSc; Mohit Kamthania[4], PhD; Rupinder Kaur[5], PhD; Ruchi Kiran Badyal[6], MPhil; Ajay Kumar Saxena[2], PhD; Ho-Joon Shin[7], PhD; Michael Kolbe[8,9], PhD; Kailash C Pandey[3], PhD

[1]Infection Biology Group, Department of Biotechnology, Mangalayatan University, Aligarh, India

[2]Molecular Medicine Laboratory, School of Life Science, Jawaharlal Nehru University, New Delhi, India

[3]Parasite-Host Biology Group, Protein Biochemistry and Engineering Lab, ICMR-National Institute of Malaria Research, New Delhi, India

[4]Department of Biotechnology, Institute of Applied Medicines and Research, Ghaziabad, India

[5]Department of Chemistry, Guru Nanak Dev University, Amritsar, India

[6]Department of Economics, Mangalayata University, Aligarh, India

[7]Department of Microbiology, School of Medicine, Ajou University, Suwon, Gyeonggi-do, Republic of Korea

[8]Centre for Structural Systems Biology, Department for Structural Infection Biology, Helmholtz-Centre for Infection Research, Hamburg, Germany

[9]Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg, Hamburg, Germany

Corresponding Author:
Sukrit Srivastava, PhD
Infection Biology Group
Department of Biotechnology
Mangalayatan University
Aligarh-Mathura Highway
Aligarh, 202145
India
Phone: 91 8178718421
Email: srivastav.sukrit@gmail.com

Abstract

Background: The novel coronavirus disease (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to the ongoing 2019-2020 pandemic. SARS-CoV-2 is a positive-sense single-stranded RNA coronavirus. Effective countermeasures against SARS-CoV-2 infection require the design and development of specific and effective vaccine candidates.

Objective: To address the urgent need for a SARS-CoV-2 vaccine, in the present study, we designed and validated one cytotoxic T lymphocyte (CTL) and one helper T lymphocyte (HTL) multi-epitope vaccine (MEV) against SARS-CoV-2 using various in silico methods.

Methods: Both designed MEVs are composed of CTL and HTL epitopes screened from 11 Open Reading Frame (ORF), structural and nonstructural proteins of the SARS-CoV-2 proteome. Both MEVs also carry potential B-cell linear and discontinuous epitopes as well as interferon gamma–inducing epitopes. To enhance the immune response of our vaccine design, truncated (residues 10-153) *Onchocerca volvulus* activation-associated secreted protein-1 was used as an adjuvant at the N termini of both MEVs. The tertiary models for both the designed MEVs were generated, refined, and further analyzed for stable molecular interaction with toll-like receptor 3. Codon-biased complementary DNA (cDNA) was generated for both MEVs and analyzed in silico for high level expression in a mammalian (human) host cell line.

Results: In the present study, we screened and shortlisted 38 CTL, 33 HTL, and 12 B cell epitopes from the 11 ORF protein sequences of the SARS-CoV-2 proteome. Moreover, the molecular interactions of the screened epitopes with their respective human leukocyte antigen allele binders and the transporter associated with antigen processing (TAP) complex were positively validated. The shortlisted screened epitopes were utilized to design two novel MEVs against SARS-CoV-2. Further molecular models of both MEVs were prepared, and their stable molecular interactions with toll-like receptor 3 were positively validated. The codon-optimized cDNAs of both MEVs were also positively analyzed for high levels of overexpression in a human cell line.

Conclusions: The present study is highly significant in terms of the molecular design of prospective CTL and HTL vaccines against SARS-CoV-2 infection with potential to elicit cellular and humoral immune responses. The epitopes of the designed

MEVs are predicted to cover the large human population worldwide (96.10%). Hence, both designed MEVs could be tried in vivo as potential vaccine candidates against SARS-CoV-2.

(JMIR Bioinformatics Biotechnol 2020;1(1):e19371) doi:10.2196/19371

KEYWORDS

COVID-19; severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); coronavirus; human transporter associated with antigen processing (TAP); toll-like receptor (TLR); epitope; immunoinformatics; molecular docking, molecular dynamics simulation; multiepitope vaccine

Introduction

The novel coronavirus disease (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in the ongoing outbreak of a severe form of respiratory disease leading to death with a mortality rate of 3.4% [1]. SARS-CoV-2 is a novel coronavirus associated with a respiratory disease that initiated in the city of Wuhan in Hubei province, China. The disease is highly contagious; as of March 21, 2020, it had spread to 182 countries and territories since its outbreak in China in December 2019. Worldwide, as of March 21, 2020, the total number of confirmed cases was reported to be 266,073, and the total death count was reported to be 11,184 [2]. Overall, SARS-CoV-2 infection has created a global emergency. The economic impact of COVID-19 is even harsher and has placed the world at economic risk. As of March 9, 2020, the worst case scenario was a US $2 trillion shortfall in global income, with a $220 billion impact on developing countries. The COVID-19 shock will cause a recession in several countries and depress global annual growth this year to below 2.5%, which is the recessionary threshold for the world economy [3].

The infection mechanism and pathogenesis of SARS-CoV-2 are currently largely unknown. According to the National Center for Biology Information (NCBI) protein sequence database [4], the proteome of SARS-CoV-2 is composed of 11 Open Reading Frame (ORF), structural and non-structural proteins. These include a polyprotein (ORF1ab), surface protein (S protein), ORF3, envelope protein (E protein), membrane protein (M protein), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid protein (N protein), and ORF10. The actual functions and pathogenic or proliferative roles of these SARS-CoV-2 coronavirus proteins are currently largely unknown.

The SARS-CoV-2 polyprotein (ORF1ab), with a length of 7096 amino acids (AAs), is composed of 16 different expressed proteins, namely leader protein (nsp1, location: 1-180 AA); nsp2 (location: 181-818 AA); nsp3 (former nsp1, carries conserved domains: N-terminal acidic, predicted phosphoesterase, papain-like proteinase, Y-domain, transmembrane domain 1 and adenosine diphosphate-ribose 1"-phosphatase, location: 819-2763 AA); nsp4 (contains transmembrane domain 2, location: 2764-3263 AA); 3C-like proteinase (nsp5, main proteinase, mediates cleavage downstream of nsp4, location: 3264-3569 AA); nsp6 (putative transmembrane domain, location: 3570-3859 AA); nsp7 (location: 3860-3942 AA); nsp8 (location: 3943-4140 AA); nsp9 (ssRNA-binding protein, location: 4141-4253 AA); nsp10 (formerly known as growth-factor-like protein, location: 4254-4392 AA); nsp11 (location: 4393-4405 AA); RNA-dependent RNA polymerase (nsp12, location: 4393-5324 AA); helicase (nsp13; zinc-binding domain, NTPase/helicase domain, RNA 5'-triphosphatase, location: 5325-5925 AA); 3'-to-5' exonuclease (nsp14, location: 5926-6452 AA); endo RNAse (nsp15, location: 6453-6798 AA); and 2'-O-ribose methyltransferase (nsp16; location: 6799-7096 AA).

The SARS-CoV-2 coronavirus S protein is a structural protein that acts as a spike protein; its location is 21563-25384 AA, and its length is 1273 AA. The ORF3a protein is located at 25393-26220 AA, and its length is 275 AA. The E protein (ORF4) is a structural protein; its location is 26245-26472 AA, and its length is 75 AA. The M protein (ORF5) is a structural protein; its location is 26523-27191 AA, and its length is 222 AA. The ORF6 protein is located at 27202-27387 AA, and its length is 61 AA. The ORF7a protein is located at 27394-27759 AA, and its length is 121 AA. The ORF7b protein is located at 27756-27887 AA, and its length is 43 AA. The SARS-CoV-2 coronavirus ORF8 protein is located at 27894-28259 AA, and its length is 121 AA. The N protein) (ORF9) is a structural protein; its location is 28274-29533 AA, and its length is 419 AA. The ORF10 protein is located at 29558-29674 AA and has a length of 38 AA [4].

Although the exact mechanisms and roles of the abovementioned proteins of the SARS-CoV-2 coronavirus proteome are not well known, these proteins are potential candidates for use in vaccines against SARS-CoV-2 coronavirus infection. In this study, we screened high-potential epitopes from all the abovementioned proteins; further, we designed and proposed cytotoxic T lymphocyte (CTL) and helper T lymphocyte (HTL) multiepitope-based vaccine candidates against SARS-CoV-2 coronavirus infection.

Methods

Background

In this study on SARS-CoV-2 coronavirus, we screened potential epitopes and designed and proposed two multiepitope vaccines (MEVs) composed of screened CTL and HTL epitopes with overlapping regions of B cell epitopes. Hence, the proposed MEVs have the potential to elicit both humoral and cellular immune response. To enhance immune response, truncated (residues 10-153) *Onchocerca volvulus* activation-associated secreted protein-1 (Ov-ASP-1) was utilized as an adjuvant at the N-termini of both MEVs. The truncated Ov-ASP-1 was chosen due to its potential to activate antigen-processing cells (APCs) [5-7]. All the SARS-CoV-2 proteins mentioned in the introduction were utilized to screen potential CTL, HTL, and

B cell epitopes. The screened epitopes were further studied to identify overlapping consensus regions among them. The epitopes showing regions of partial or complete overlap were chosen for further detailed studies.

The chosen CTL and HTL epitopes were analyzed for their molecular interactions with their respective human leukocyte antigen (HLA) allele binders. Moreover, the molecular interactions of the chosen CTL epitopes were analyzed for with the transporter associated with antigen processing (TAP) cavity to observe their smooth passage from the cytoplasm to the endoplasmic reticulum (ER) lumen [8,9]. Tertiary models of both MEVs were generated and refined. Both MEV models were further utilized to screen B cell linear and discontinuous epitopes as well as interferon gamma (IFNγ)-inducing epitopes.

Molecular signaling by multiple toll-like receptors is an essential component of the innate immune response against SARS-CoV-2. Because Ov-ASP-1 primarily binds APCs among human peripheral blood mononuclear cells and triggers proinflammatory cytokine production via toll-like receptor 3 (TLR3), the molecular interactions of both the CTL and HTL MEV models with TLR3 were further analyzed by molecular docking studies [10-13]. Furthermore, the codon-optimized cDNAs of both MEVs were analyzed and were found to have high levels of expression in a mammalian (human) cell line, which would facilitate in vivo expression, experimentation, and trials (see Supplementary Figure S1 in Multimedia Appendix 1).

Screening of Potential Epitopes

T cell Epitope Prediction

Screening of CTL Epitopes

The CTL epitopes were screened using the Immune Epitope Database (IEDB) tools MHC (major histocompatibility complex)-I Binding Predictions and MHC-I Processing Predictions [14-16]. These two tools use six different methods (consensus, NN-align, SMM-align, combinatorial library, Sturniolo, and NetMHCIIpan), and they generate a percentile rank and a total score, respectively.

The screening is based on the total number of cleavage sites in the protein. The TAP score estimates an effective –log value of the half maximal inhibitory concentration (IC_{50}) for binding to the TAP of a peptide or its N-terminal prolonged precursors. The MHC binding prediction score is the –log(IC_{50}) value for binding to the MHC of a peptide [17]. The IC_{50} values (nanomolar) for each epitope and MHC allele binding pair were also obtained using the MHC-I Binding Predictions IEDB tool. Epitopes with high, intermediate, and low affinities of binding to their HLA allele binders have IC_{50} values of <50 nM, <500 nM, and <5000 nM, respectively.

The immunogenicities of all the screened CTL epitopes were also obtained using the MHC I Immunogenicity IEDB tool [17] with all parameters set to the default to analyze the first, second, and C-terminus amino acids of each screened epitope. The tool predicts the immunogenicity of a given peptide-MHC complex based on the physiochemical properties of its constituting amino acids and their positions within the peptide sequence.

Screening of HTL Epitopes

To screen out the HTL epitopes from the SARS-CoV-2 proteins, the IEDB tool MHC-II Binding Predictions was used. This tool generates a percentile rank for each potential peptide. The lower the percentile rank, the higher the affinity. This percentile rank is generated by the combination of three different methods, namely combinatorial library, SMM_align, and Sturniolo, and by comparing the score of the peptide against the scores of five million other random 15-mer peptides in the SWISS-PROT database [18-21]. The rank from the consensus of all three methods was generated by the median percentile rank of the three methods.

Population Coverage by CTL and HTL Epitopes

The IEDB Population Coverage tool was used to elucidate the world human population coverage by the shortlisted 38 CTL and 33 HTL epitopes derived from 9 SARS-CoV-2 proteins [22]. T cells recognize the complex between a specific major MHC molecule and a particular pathogen-derived epitope. The given epitope will only elicit a response in an individual who expresses an MHC molecule that is capable of binding that particular epitope. This denominated MHC restriction of T cell responses and the MHC polymorphism provides the basis for population coverage study. The MHC types are expressed at dramatically different frequencies in different ethnicities. Hence, a vaccine with larger population coverage could be of greater importance [21]. Clinical administration of multiple epitopes, including both CTL and HTL epitopes, is predicted here to have a higher probability of larger human population coverage worldwide.

B Cell Epitope Prediction

Sequence-Based B Cell Epitope Prediction

The protein sequence–based Bepipred Linear Epitope Prediction method [23] was utilized to screen linear B cell epitopes from 11 different SARS-CoV-2 protein ORFs. The B Cell Epitope Prediction Tools of the IEDB server were utilized. In this screening, parameters such as the hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity, and antigenic propensity of the polypeptides are correlated with their location in the protein. This enables a search for continuous epitopes predicted from a protein sequence. The prediction is based on the propensity scales for each of the 20 amino acids. For a window size n, i – (n – 1)/2 neighboring residues on each side of residue i are used to compute the score for residue i. The Bepipred Linear Epitope Prediction method used here is based on the propensity scale method as well as the physiochemical properties of the given antigenic sequence to screen potential epitopes [23].

Characterization of Potential Epitopes

Epitope Conservation Analysis

The shortlisted CTL, HTL, and B cell epitopes screened from eleven SARS-CoV-2 proteins were analyzed for the conservancy of their amino acid sequences using the IEDB Epitope Conservancy Analysis tool. The epitope conservancy is the number of protein sequences retrieved from the NCBI protein database that contains that particular epitope. The analysis was performed against the entire respective source protein sequences of SARS-CoV-2 proteins retrieved from the NCBI protein database [24].

Epitope Toxicity Prediction

The ToxinPred tool was used to analyze the toxicity of the shortlisted CTL, HTL, and B cell epitopes. The tool enables the identification of highly toxic or nontoxic short peptides. The toxicity check analysis was performed using the support vector machine-based ToxinPred method using a dataset of 1805 positive sequences and 3593 negative sequences from SWISS-PROT as well as an alternative dataset comprising the same 1805 positive sequences and 12,541 negative sequences from the Translated European Molecular Biology Laboratory (TrEMBL) database [25].

Overlapping Residue Analysis

The overlapping residue analysis for the shortlisted 38 CTL, 33 HTL, and 12 B cell linear epitopes was performed using multiple sequence alignment analysis with the European Bioinformatics Institute's Clustal Omega tool [26]. The Clustal Omega multiple sequence alignment tool virtually aligns any number of protein sequences and delivers an accurate alignment.

Selection of Epitopes for Molecular Interaction Studies With HLA Alleles and the TAP Transporter

Based on the overlapping residue analysis of the shortlisted CTL, HTL, and linear B cell epitopes, a few CTL and HTL epitopes were chosen for further analysis. The chosen epitopes are circled in Supplementary Figure S10 (Multimedia Appendix 1). These epitopes were chosen based on their partial or full overlapping sequence regions among all three types of epitopes (CTL, HTL, and B cell). The chosen epitopes were further analyzed for their interactions with their respective HLA allele binders and TAP cavity interactions.

Molecular Interaction Analysis of the Selected Epitopes with HLA Alleles and the TAP Transporter

Tertiary Structure Modeling of HLA Alleles and Selected T Cell Epitopes

SWISS-MODEL [27] was used for homology modeling of the HLA class I and II allele binders of the chosen epitopes. The amino acid sequences of the HLA allele binders were retrieved from the Immuno Polymorphism Database (IPD-IMGT/HLA). Templates for homology modeling were chosen based on the highest amino acid sequence similarity. All the generated HLA allele models had acceptable QMEAN values (cutoff -4.0) (Supplementary Table S1, Multimedia Appendix 1). The QMEAN value gives a composite quality estimate involving both global and local analysis of the model [28].

PEP-FOLD 2.0 [29], a de novo structure prediction tool available at RPBS Web Portal, was utilized to generate tertiary structures for the chosen CTL and HTL epitopes.

Molecular Interaction Analysis of Chosen CTL and HTL Epitopes With HLA Alleles

The PatchDock tool was utilized for in silico molecular docking studies of the selected CTL and HTL epitopes with their respective HLA class I and II allele binders. PatchDock utilizes an algorithm for unbound (real-life) docking of molecules for protein-protein complex formation. The algorithm carries out rigid docking, and the surface variability/flexibility is implicitly addressed through liberal intermolecular penetration. The algorithm focuses on the initial molecular surface fitting on localized, curvature-based surface patches, the use of geometric hashing and pose clustering for initial transformation detection, computation of shape complementarity utilizing Distance Transform, efficient steric clash detection and geometric fit scoring based on multiresolution shape representation, and utilization of biological information by focusing on hotspot-rich surface patches [30-32].

Molecular Interaction Analysis of Selected CTL Epitopes With the TAP Transporter

The TAP transporter plays an important role in the presentation of a CTL epitope. From the cytosol after proteasome processing, the fragmented peptide of the foreign protein is transported to the ER through the TAP transporter. From the ER, these short peptides reach the Golgi bodies and are then presented on the cell surface [9]. Molecular interaction studies of the chosen CTL epitopes within the TAP cavity were performed by molecular docking using the PatchDock tool. For accurate prediction, the cryo-EM structure of the TAP transporter (PDB ID: 5u1d) was used by removing the antigen from the TAP cavity of the original structure [8].

Design, Characterization, and Molecular Interaction Analysis of MEVs With Immune Receptors

Design of the MEVs

The screened and shortlisted high-scoring 38 CTL and 33 HTL epitopes were utilized to design CTL and HTL MEVs (Tables 1 and 2). Two short peptides, EAAAK and GGGGS, were used as rigid and flexible linkers, respectively (Supplementary Figure S2, Multimedia Appendix 1). The GGGGS linker provides proper conformational flexibility to the tertiary structure of the vaccine and hence facilitates stable conformation of the vaccine. The EAAAK linker facilitates domain formation and hence aids the vaccine to obtain its final stable structure. Truncated Ov-ASP-1 protein was utilized as an adjuvant at the N termini of both the CTL and HTL MEVs [5-7,33-37].

Characterization of the Designed MEVs

Physicochemical Property Analysis of the Designed MEVs

The ProtParam tool [38] was utilized to analyze the physiochemical properties of the amino acid sequences of the designed CTL and HTL MEVs. The ProtParam analysis performs an empirical investigation of the amino acid sequence in a given query. ProtParam computes various physicochemical properties derived from a given protein sequence.

IFNγ-Inducing Epitope Prediction

From the designed amino acid sequences of both MEVs, potential IFNγ epitopes were screened by the IFN epitope server using a hybrid motif and support vector machine approach; the motif-based method used was MERCI (Motif-EmeRging and with Classes-Identification). This tool predicts peptides from protein sequences that have the capacity to induce IFNγ release from $CD4^+$ T cells. This module generates overlapping peptides from the query sequence and predicts IFNγ-inducing peptides. For the screening, the IEDB database was used with 3705 IFNγ-inducing and 6728 non–IFNγ-inducing MHC class II binders [39,40].

MEV Allergenicity and Antigenicity Prediction

Both the designed MEVs were further analyzed for allergenicity and antigenicity prediction using the AlgPred [41] and VaxiJen [42] tools, respectively. The AlgPred prediction is based on the similarity of an already known epitope with any region of the submitted protein. To screen allergenicity, the SWISS-PROT data set consisting of 101,725 non-allergens and 323 allergens was used. VaxiJen utilizes an alignment-free approach that is based solely on the physicochemical properties of the query amino acid sequence. To predict the antigenicity, VaxiJen uses bacterial, viral, and tumor protein datasets to derive models for the prediction of the antigenicity of a whole protein. Every set consists of known 100 antigens and 100 nonantigens.

Tertiary Structure Modeling, Refinement, and Validation of the MEVs

The tertiary structures of the designed CTL and HTL MEVs were generated by homology modeling using the I-TASSER modeling tool [43]. I-TASSER is a protein structure prediction tool that is based on the sequence-to-structure-to-function paradigm. The tool generates 3D atomic models from multiple threading alignments and iterative structural assembly simulations for a submitted AA sequence. I-TASSER is based on the structure templates identified by LOMETS, a metaserver from the Protein Data Bank (PDB) library. I-TASSER only uses the template with the highest Z-score, which is the difference between the raw and average scores in the unit of standard deviation. For each target model, the I-TASSER simulations generate a large ensemble of structural conformations, called decoys. To select the final models, I-TASSER uses the SPICKER program to cluster all the decoys based on their pairwise structure similarity and reports up to 5 models. A normalized Z-score >1 indicates a good alignment and vice versa. The Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of the query protein. Ranking of template proteins is based on the TM-score of the structural alignment between the query structure model and known structures. The root mean square deviation (RMSD) is the RMSD between template residues and query residues that are structurally aligned by the TM-align algorithm.

Both the generated MEV models were refined using the ModRefiner [44] and GalaxyRefine [45] tools. The TM-score generated by ModRefiner indicates the structural similarity of the refined model to the original input model. The closer the TM-score to 1, the greater the similarity of the original and refined models. The RMSD of the refined model shows the conformational deviation from the initial input models.

The GalaxyRefine tool refines the query tertiary structure by repeated structure perturbation and by using the subsequent structural relaxation by the molecular dynamics simulation. The GalaxyRefine tool generates reliable core structures from multiple templates and then rebuilds unreliable loops or termini using an optimization-based refinement method [46,47]. To avoid any breaks in the 3D model, GalaxyRefine uses the triaxial loop closure method. The MolProbity score generated for a given refined model indicates the log-weighted combination of the clash score, the percentage of Ramachandran unfavored residues and the percentage of bad side chain rotamers.

Validation of the Refined Models of the CTL and HTL MEVs

The refined CTL and HTL MEV 3D models both were further validated by the RAMPAGE analysis tool [48,49]. The generated Ramachandran plots for the MEV models show the sterically allowed and disallowed residues along with their dihedral psi (ψ) and phi (ϕ) angles.

Linear and Discontinuous B-cell Epitope Prediction of the MEVs

The ElliPro antibody epitope prediction tool available at the IEDB was used to screen the linear and discontinuous B cell epitopes from the MEV models. The ElliPro method analyses are based on the location of a residue in the 3D structure of a protein. For example, the residues lying outside an ellipsoid covering 90% of the inner core protein residues score the highest protrusion index (PI) of 0.9. The discontinuous epitopes predicted by the ElliPro tool are clustered based on the distance R in angstroms between the centers of mass of two residues lying outside the largest possible ellipsoid. A larger value of R indicates that more distant residues (residue discontinuity) are screened in the epitopes [50,51].

Molecular Interaction Analysis of MEVs With an Immunological Receptor

Molecular Docking Studies of the MEVs and TLR3

Molecular interaction analysis of both designed MEVs with TLR3 was performed by molecular docking and molecular dynamics simulations. Molecular docking was performed using the PatchDock server [32]. PatchDock utilizes an algorithm for unbound docking of molecules (mimicking the real world environment) for protein-protein complex formation, as explained earlier [30,31]. For molecular docking, the 3D structure of the human TLR3 ectodomain was retrieved from the PDB (PDB ID: 2A0Z). The study provides the dynamical properties of the designed system with the MEV-TLR3 complexes and guesses at the interactions between the molecules; also, it gives exact predictions of bulk properties, including hydrogen bond formation and the conformation of the molecules forming the complex.

Molecular Dynamics Simulation Studies of the MEVs and the TLR3 Complex

The MEV-TLR3 molecular interactions were further evaluated using molecular dynamics simulations. The molecular dynamics simulations were performed for 10 nanoseconds using YASARA (Yet Another Scientifc Artifcial Reality Application) [52]. The simulations were performed in an explicit water environment in a dodecahedron simulation box at a constant temperature (298 kelvin) and pressure (1 atmosphere) at pH 7.4 with a periodic cell boundary condition. The solvated systems were neutralized with counterions (sodium chloride, concentration 0.9 molar). The AMBER14 force field was applied to the systems during the simulations [53,54]. Long-range electrostatic energies and forces were calculated using the particle mesh–based Ewald method [55]. The solvated structures were minimized by the steepest descent method at a temperature of 298 K and a constant pressure. Then, the complexes were equilibrated for a period of 1 nanosecond. After equilibration, a production molecular dynamics simulation was run for 10 ns at a constant temperature and pressure, and time frames were saved every 10 picoseconds for each simulation. The RMSD and root mean square fluctuation (RMSF) values for the alpha carbon (C_α) atoms, backbone atoms, and all the atoms of both MEV complexes were analyzed for each simulation conducted.

Generation and Analysis of cDNA of the MEVs

cDNAs of both MEVs, codon-optimized for expression in a mammalian (human) cell line, were generated using the Java Codon Adaptation Tool. The generated cDNAs of both the MEVs were further analyzed by the GenScript Rare Codon Analysis Tool. This tool analyzes the GC content, codon adaptation index (CAI) and tandem rare codon frequency for a given cDNA [56,57]. The CAI indicates the possibility of cDNA expression in a chosen expression system. The tandem rare codon frequency indicates the presence of low-frequency codons in a given cDNA.

Results

Screening of Potential Epitopes

T Cell Epitope Prediction

Screening of CTL Epitopes

CTL epitopes were screened using the MHC-I Binding Predictions and MHC-I Processing Predictions IEDB tools. These epitopes were shortlisted based on the total number of cleavage sites in the protein, low IC_{50} (nM) values for epitope-HLA class I allele pairs, and binding to the TAP cavity.

The 38 epitopes predicted by the MHC-I Binding Predictions tool with the highest percentile ranks were shortlisted for MEV design and are listed in Table 1. The remaining 101 epitope-HLA I allele pairs are listed in Supplementary Table S8 (Multimedia Appendix 1). The 67 epitope-HLA I allele pairs predicted by the MHC-I Processing Predictions tool with the highest total scores are listed in Supplementary Table S9 (Multimedia Appendix 1).

The immunogenicities of the shortlisted CTL epitopes were also determined and are noted in Table 1 and in Supplementary Tables S8 and S9 (Multimedia Appendix 1). A higher immunogenicity score indicates greater immunogenic potential of the given epitope.

Table 1. Characteristics of the shortlisted high-percentile-ranking SARS-CoV-2 CTL epitopes and their respective HLA allele binders.

SARS-CoV-2[a] protein	Peptide length, amino acids	Peptide sequence	Conservancy (%)	Immunogenicity	Toxicity	Allele	Methods used[b]	Percentile rank
E protein[c]	9	LLFLAFVVF	480/482 (99.59)	0.2341	Nontoxic	B15:01	Consensus (ann/smm/comblib_sidney 2008)	0.1
E protein	9	LTALRLCAY	478/482 (99.17)	0.01886	Nontoxic	A01:01	Consensus (ann/smm)	0.12
M protein[d]	11	YFIASFRLFAR	474/477 (99.37)	0.19709	Nontoxic	A33:01	ann	0.03
M protein	10	ATSRTLSYYK[e]	472/477 (98.95)	–0.13563	Nontoxic	A11:01	Consensus (ann/smm)	0.06
N protein[f]	9	MEVTPSGTW	485/498 (97.39)	–0.06279	Nontoxic	B44:02	Consensus (ann/smm)	0.06
N protein	9	KPRQKRTAT	487/498 (97.79)	–0.20542	Nontoxic	B07:02	Consensus (ann/smm/comblib_sidney 2008)	0.1
orf10	9	MGYINVFAF	477/480 (99.38)	–0.09452	Nontoxic	B35:01	Consensus (ann/smm/comblib_sidney 2008)	0.1
orf10	10	GYINVFAFPF[e]	232/236 (98.31)	0.20158	Nontoxic	A23:01	Consensus (ann/smm)	0.11
orf-1ab	11	SEMVMCGGSLY	452/456 (99.12)	0.32633	Nontoxic	B44:02	ann	0.03
orf-1ab	11	FYWFFSNYLKR	455/456 (99.78)	0.37766	Nontoxic	A33:01	ann	0.04
orf-1ab	8	ISNSWLMW	454/456 (99.56)	–0.24791	Nontoxic	B58:01	ann	0.05
orf-1ab	10	ETISLAGSYK	455/456 (99.78)	0.08174	Nontoxic	A68:01	Consensus (ann/smm)	0.06
orf-1ab	9	QEILGTVSW	455/456 (99.78)	0.27341	Nontoxic	B44:02	Consensus (ann/smm)	0.06
orf-1ab	9	STFNVPMEK	456/456 (100.00)	–0.32016	Nontoxic	A11:01	Consensus (ann/smm)	0.06
orf-1ab	10	RMYIFFASFY	456/456 (100.00)	0.21107	Nontoxic	A30:02	Consensus (ann/smm)	0.06
orf-1ab	10	FLFVAAIFYL	454/456 (99.56)	–0.19814	Nontoxic	A02:01	Consensus (ann/smm)	0.06
orf-1ab	10	RYFRLTLGVY	456/456 (100.00)	0.03976	Nontoxic	A30:02	Consensus (ann/smm)	0.06
orf-1ab	9	FLNGSCGSV	456/456 (100.00)	–0.20585	Nontoxic	A02:03	Consensus (ann/smm)	0.06
orf-1ab	9	CTDDNALAY	476/479 (99.37)	0.32004	Nontoxic	A01:01	Consensus (ann/smm)	0.06
orf-1ab	10	CTDDNALAYY[e]	476/479(99.37)	0.28694	Nontoxic	A01:01	Consensus (ann/smm)	0.06
orf-1ab	11	MYKGLPWNVVR	456/456 (100.00)	–0.11151	Nontoxic	A33:01	ann	0.06
orf-1ab	10	SIINNTVYYTK[e]	456/456 (100.00)	0.15936	Nontoxic	A11:01	Consensus (ann/smm)	0.06
orf-1ab	10	LPVNVAFELW	450/456 (98.68)	–0.00254	Nontoxic	B53:01	Consensus (ann/smm)	0.06
orf-1ab	9	DEWSMATYY[e]	455/456 (99.78)	0.07355	Nontoxic	B44:03	Consensus (ann/smm)	0.07
orf-1ab	10	YILFTRFFYV	454/456 (99.56)	–0.02845	Nontoxic	A02:06	Consensus (ann/smm)	0.07
orf-1ab	10	YIFFASFYYV	456/456 (100.00)	0.12661	Nontoxic	A02:06	Consensus (ann/smm)	0.07
ORF3a	9	YLYALVYFL[e]	456/456 (100.00)	0.40924	Nontoxic	A02:01	Consensus (ann/smm/comblib_sidney 2008)	0.1
ORF3a	10	IPYNSVTSSI	454/456 (99.56)	0.13772	Nontoxic	B51:01	Consensus (ann/smm)	0.11

SARS-CoV-2[a] protein	Peptide length, amino acids	Peptide sequence	Conservancy (%)	Immunogenicity	Toxicity	Allele	Methods used[b]	Percentile rank
Orf6	8	RTFKVSIW	466/481 (96.88)	0.13151	Nontoxic	B57:01	ann	0.05
Orf6	11	AEILLIIMRTF	471/481 (97.92)	–0.32835	Nontoxic	B44:02	ann	0.06
ORF7a	8	RARSVSPK	480/481 (99.79)	–0.18221	Nontoxic	A30:01	ann	0.11
ORF7a	10	QLRARSVSPK	479/481 (99.58)	0.1815	Nontoxic	A03:01	Consensus (ann/smm)	0.16
orf7b	9	FLAFLLFLV	472/480 (98.33)	–0.16177	Nontoxic	A02:03	Consensus (ann/smm)	0.07
orf8	9	HFYSKWYIR	472/480 (98.33)	–0.27456	Nontoxic	A31:01	Consensus (ann/smm)	0.11
S protein[g]	10	WTA-GAAAYYV	470/472 (99.58)	0.15455	Nontoxic	A68:02	Consensus (ann/smm)	0.06
S protein	10	FPNITNLCPF	472/472 (100.00)	0.1009	Nontoxic	B53:01	Consensus (ann/smm)	0.06
S protein	10	NYNYLYRLFR	465/472 (98.52)	0.08754	Nontoxic	A33:01	Consensus (ann/smm)	0.07
S protein	8	NYLYRLFR	465/472 (98.52)	0.13144	Nontoxic	A33:01	ann	0.07

[a] SARS-CoV-2: severe acute resipiratory syndrome coronavirus 2.

[b] Methods: ann: artificial neural network. Comblib_sidney2008: combinatorial peptide libraries [19]. smm: stabilized matrix method.

[c] E protein: envelope protein.

[d] M protein: membrane protein.

[e] Matches a recently published epitope, indicating consensus with results [58].

[f] N protein: nucleocapsid protein.

[g] S protein: surface protein.

Screening of HTL Epitopes

The screening of HTL epitopes from 11 different SARS-CoV-2 ORF proteins was performed based on percentile rank. The smaller the percentile rank, the higher the affinity of the peptide to its respective HLA allele binders. The 33 epitopes with high percentile ranking were shortlisted (Table 2). An additional 180 potential HTL cell epitope-HLA allele II pairs with high percentile ranks screened in our study are listed in Supplementary Table S10 (Multimedia Appendix 1).

Table 2. Characteristics of the shortlisted high-scoring SARS-CoV-2 HTL epitopes and their respective HLA allele binders.

SARS-CoV-2[a] protein	Peptide	Conservancy (%)	Toxicity	Alleles	Methods used[b]	Percentile rank
E protein[c]	LLFLAFVVFLLVTLA	480/482 (99.59)	Nontoxic	DPA1-03:01/DPB1-04:02	Consensus (comb.lib./smm/nn)	0.02
E protein	VLLFLAFVVFLLVTL	480/482 (99.59)	Nontoxic	DPA1-03:01/DPB1-04:02	Consensus (comb.lib./smm/nn)	0.02
M protein[d]	GLMWLSYFIASFRLF	465/477 (97.48)	Nontoxic	DPA1-01:03/DPB1-02:01	Consensus (comb.lib./smm/nn)	0.05
M Protein	LMWLSYFIASFRLFA	466/477 (97.69)	Nontoxic	DPA1-01:03/DPB1-02:01	Consensus (comb.lib./smm/nn)	0.05
M protein	LSYYKLGASQRVAGD[e]	472/477 (98.95)	Nontoxic	DRB1-09:01	Consensus (comb.lib./smm/nn)	0.06
N protein[f]	AQFAPSASAFFGMSR	486/498 (97.59)	Nontoxic	DRB1-09:01	Consensus (comb.lib./smm/nn)	0.01
N protein	IAQFAPSASAFFGMS	485/498 (97.39)	Nontoxic	DRB1-09:01	Consensus (comb.lib./smm/nn)	0.01
N protein	PQIAQFAPSASAFFG	485/498 (97.39)	Nontoxic	DRB1-09:01	Consensus (comb.lib./smm/nn)	0.01
ORF1ab	AIILASFSASTSAFV	456/456 (100.00)	Nontoxic	DRB1-09:01	Consensus (comb.lib./smm/nn)	0.01
ORF1ab	ESPFVMMSAPPAQYE[e]	456/456 (100.00)	Nontoxic	DRB1-01:01	Consensus (comb.lib./smm/nn)	0.01
ORF1ab	IILASFSASTSAFVE	456/456 (100.00)	Nontoxic	DRB1-09:01	Consensus (comb.lib./smm/nn)	0.01
ORF1ab	QESPFVMMSAPPAQY	456/456 (100.00)	Nontoxic	DRB1-01:01	Consensus (comb.lib./smm/nn)	0.01
ORF1ab	SPFVMMSAPPAQYEL	456/456 (100.00)	Nontoxic	DRB1-01:01	Consensus (comb.lib./smm/nn)	0.01
ORF3a	FVRATATIPIQASLP	478/481 (99.37)	Nontoxic	DPA1-02:01/DPB1-14:01	NetMHCIIpan	0.12
ORF3a	LLFVTVYSHLLLVAA	467/481 (97.08)	Nontoxic	DRB1-01:01	Consensus (comb.lib./smm/nn)	0.1
ORF6	FKVSIWNLDYIINLI	478/481 (99.38)	Nontoxic	DQA1-01:01/DQB1-05:01	Consensus (comb.lib./smm/nn)	0.02
ORF6	KVSIWNLDYIINLII	478/481 (99.38)	Nontoxic	DQA1-01:01/DQB1-05:01	Consensus (comb.lib./smm/nn)	0.02
ORF6	TFKVSIWNLDYIINL[e]	478/481 (99.38)	Nontoxic	DQA1-01:01/DQB1-05:01	Consensus (comb.lib./smm/nn)	0.02
ORF7a	IILFLALITLATCEL	479/480 (99.79)	Nontoxic	DRB1-01:01	Consensus (comb.lib./smm/nn)	0.16
ORF7a	ILFLALITLATCELY	479/480 (99.79)	Nontoxic	DRB1-01:01	Consensus (comb.lib./smm/nn)	0.16
ORF7b	CFLAFLLFLVLIMLI	231/236 (97.88)	Nontoxic	DPA1-03:01/DPB1-04:02	Consensus (comb.lib./smm/nn)	0.03
ORF7b	LCFLAFLLFLVLIML	231/236 (97.88)	Nontoxic	DPA1-03:01/DPB1-04:02	Consensus (comb.lib./smm/nn)	0.02
ORF7b	YLCFLAFLLFLVLIM	231/236 (97.88)	Nontoxic	DPA1-03:01/DPB1-04:02	Consensus (comb.lib./smm/nn)	0.02
ORF8	CTQHQPYVVDDPCPI	476/480 (99.17)	Nontoxic	DRB3-01:01	Consensus (comb.lib./smm/nn)	0.08
ORF8	HQPYVVDDPCPIHFY	476/480 (99.17)	Nontoxic	DRB3-01:01	Consensus (comb.lib./smm/nn)	0.08

SARS-CoV-2[a] protein	Peptide	Conservancy (%)	Toxicity	Alleles	Methods used[b]	Percentile rank
ORF8	QPYVVDDPCPIHFYS	476/480 (99.17)	Nontoxic	DRB3-01:01	Consensus (comb.lib./smm/nn)	0.07
ORF10	INVFAFPFTIYSLLL	476/480 (99.17)	Nontoxic	HLA-DPA1-01:03/DPB1-02:01	Consensus (comb.lib./smm/nn)	0.29
ORF10	YINVFAFPFTIYSLL	476/479 (99.37)	Nontoxic	DPA1-01:03/DPB1-02:01	Consensus (comb.lib./smm/nn)	0.29
S protein[g]	KTQSLLIVNNATNVV	472/472 (100.00)	Nontoxic	DRB1-13:02	Consensus (smm/nn/sturniolo)	0.01
S protein	LLIVNNATNVVIKVC	469/472 (99.36)	Nontoxic	DRB1-13:02	Consensus (smm/nn/sturniolo)	0.01
S protein	QSLLIVNNATNVVIK	471/472 (99.79)	Nontoxic	DRB1-13:02	Consensus (smm/nn/sturniolo)	0.01
S protein	SLLIVNNATNVVIKV[e]	471/472 (99.79)	Nontoxic	DRB1-13:02	Consensus (smm/nn/sturniolo)	0.01
S protein	TQSLLIVNNATNVVI	471/472 (99.79)	Nontoxic	DRB1-13:02	Consensus (smm/nn/sturniolo)	0.01

[a] SARS-CoV-2: severe acute resipiratory syndrome coronavirus 2.

[b] Methods: comb.lib.: combinatorial library. nn: neural network. smm: stabilized matrix method.

[c] E protein: envelope protein.

[d] M protein: membrane protein.

[e] Matches a recently published epitope, indicating consensus with results [58].

[f] N protein: nucleocapsid protein

[g] S protein: surface protein.

Population Coverage by CTL and HTL Epitopes

The population coverage by the shortlisted epitopes was also studied, particularly in China, France, Italy, the United States, South Asia, East Asia, Northeast Asia, and the Middle East. From this study, we can conclude that the combined use of all the shortlisted CTL and HTL epitopes would have an average worldwide population coverage as high as 96.10% (SD 23.74) (Supplementary Table S12, Multimedia Appendix 1).

B Cell Epitope Prediction

Sequence-Based B Cell Epitope Prediction

To screen B cell epitopes, we used the Bepipred Linear Epitope Prediction method. In our study, we screened 12 B cell epitopes from 11 SARS-CoV-2 ORF proteins which show partial or complete overlap with the shortlisted CTL and HTL epitopes (Table 3). An additional 206 B cell epitopes with epitope lengths of at least four AAs and a maximum of 20 AAs were screened and are listed in Supplementary Table S11, Multimedia Appendix 1.

Table 3. Characteristics of the shortlisted SARS-CoV-2 linear B cell epitopes obtained by the BepiPred method.

SARS-CoV-2[a] protein	Peptide length, amino acids	Conservancy (%)	Overlapping B cell epitope	Toxicity
M protein[b]	12	471/477 (98.74)	KLGASQRVAGDS	Nontoxic
N protein[c]	42	483/498 (96.99)	RLNQLESKMSGKGQQQQGQTVTKKSAAEASK KPRQKRTATKA	Nontoxic
ORF1ab	20	455/456 (99.78)	GTTQTACTDDNALAYYNTTK	Nontoxic
ORF3a	12	478/481 (99.37)	QGEIKDATPSDF	Nontoxic
ORF3a	6	471/481 (97.92)	PYNSVT	Nontoxic
ORF7a	9	479/480 (99.79)	LYHYQECVR	Nontoxic
ORF7a	26	470/480 (97.92)	VKHVYQLRARSVSPKLFIRQEEVQEL	Nontoxic
ORF8	23	460/480 (95.83)	QSCTQHQPYVVDDPCPIHFYSKW	Nontoxic
ORF8	9	476/480 (99.17)	RVGARKSAP	Nontoxic
S protein[d]	11	470/472 (99.58)	TPGDSSSGWTA	Nontoxic
S protein	35	470/472 (99.58)	FPNITNLCPFGEVFNATRFASVYAWNRKRISNCVA	Nontoxic
S protein	62	454/472 (96.19)	NLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIY QAGSTPCNGVEGFNCYFPLQSYGFQPTN	Nontoxic

[a] SARS-CoV-2: severe acute resipiratory syndrome coronavirus 2.

[b] M protein: membrane protein.

[c] N protein: nucleocapsid protein

[d] S protein: surface protein.

Characterization of Potential Epitopes

Epitope Conservation Analysis

Sequence conservation analysis of the screened CTL, HTL, and B cell epitopes showed the highly conserved nature of the shortlisted epitopes. The amino acid sequences of both the CTL epitopes and the HTL epitopes were found to be significantly conserved among the NCBI-retrieved protein sequences of SARS-CoV-2 (the CTL epitopes were 96.88%-100% conserved and the HTL epitopes were 97.08%-100% conserved; see Tables 1, 2, and 4 and Supplementary Tables S8, S9, S10, and S11, Multimedia Appendix 1).

Epitope Toxicity Prediction

Toxicity analyses of all the screened CTL, HTL, and B cell epitopes were also performed. The ToxinPred study of all the shortlisted epitopes showed that they all are nontoxic (Tables 1, 2, and 4; Supplementary Tables S8, S9, S10, and S11, Multimedia Appendix 1).

Overlapping Residue Analysis

The AA sequence overlap among the shortlisted CTL, HTL, and B cell epitopes from 11 SARS-CoV-2 ORF proteins was analyzed using the Clustal Omega multiple sequence alignment analysis tool. The analysis showed that several CTL, HTL, and B cell epitopes had overlapping AA sequences. The CTL, HTL, and B cell epitopes with two or more overlapping AA residues are shown in Supplementary Figure S3 (Multimedia Appendix 1).

Selection of Epitopes for Molecular Interaction Studies with HLA Alleles and the TAP Transporter

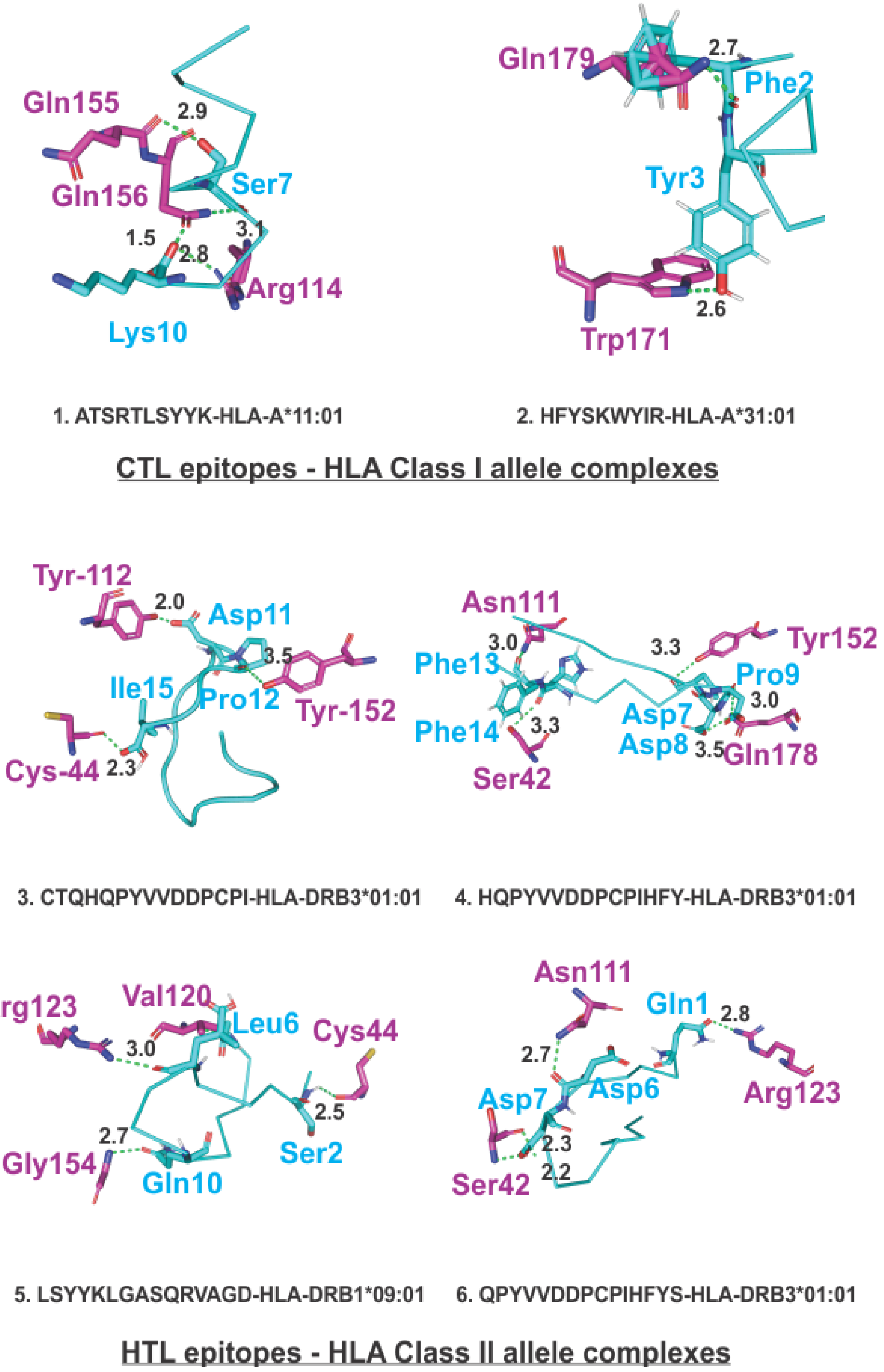
The epitopes showing overlap among the CTL, HTL, and B cell epitopes are circled in Supplementary Figure S10 (Multimedia Appendix 1) and were chosen for further study of their interactions with HLA alleles and the TAP transporter.

Molecular Interaction Analysis of Selected Epitopes With HLA Alleles and the TAP Transporter

Molecular Interaction Analysis of the Chosen CTL and HTL Epitopes With HLA Alleles

Molecular docking studies of the chosen CTL and HTL epitopes with their respective HLA class I and II allele binders were performed using the PatchDock tool. Images were generated by PyMOL [59]. The study revealed significant molecular interactions between all the chosen epitopes and their HLA allele binders, showing the formation of multiple hydrogen bonds (Figure 1). Furthermore, B-factor analysis of all the epitope–HLA allele complexes showed that the epitope ligand had a stable (blue) binding conformation in complex with the HLA allele molecule (Supplementary Figure S4, Multimedia Appendix 1). The violet-indigo-blue-green-yellow-orange-red (VIBGYOR) color presentation was used, where blue is very stable.

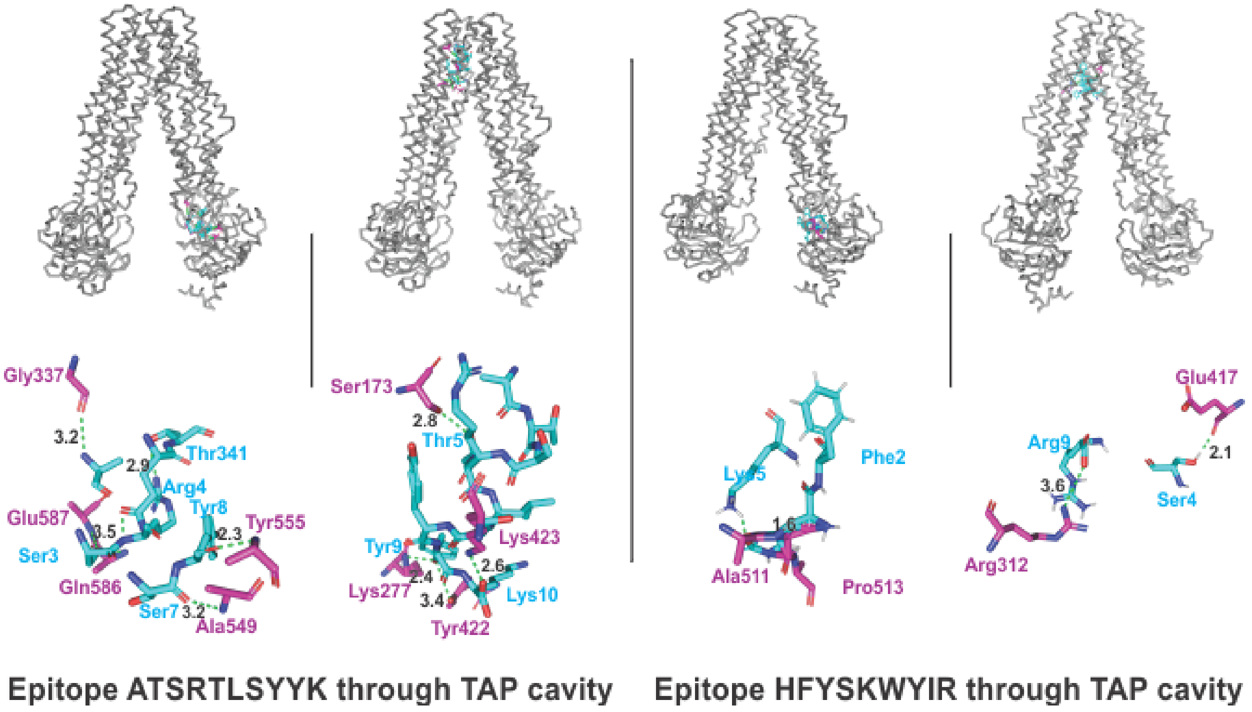
Figure 1. Molecular docking analysis of SARS-CoV-2 CTL epitopes and HLA alleles. Molecular docking of the chosen CTL and HTL epitopes (cyan sticks) binding the amino acid residues of their respective HLA class I and class II allele binders (magenta sticks). The study shows that the docked complexes are stable, with the formation of multiple hydrogen bonds (green dots, lengths in angstroms). CTL: cytotoxic T lymphocyte. HLA: human leukocyte antigen.



Molecular Interaction Analysis of Selected CTL Epitopes With the TAP Cavity

The molecular docking interaction analysis of the chosen CTL epitopes with the TAP cavity showed significantly strong molecular interactions with the formation of several hydrogen bonds at different sites of the TAP cavity. Two sites of interaction were of particular interest: one closer to the cytoplasmic end and another closer to the ER lumen (Figure 2). This study confirms the feasibility of transportation of the chosen CTL epitopes from the cytoplasm to the ER lumen, which is an essential event for the representation of an epitope by HLA allele molecules on the surface of antigen-presenting cells.

Figure 2. Molecular docking analysis of two CTL epitopes within the TAP transporter cavity. The molecular interactions of the CTL epitopes (cyan sticks) within the TAP cavity (gray ribbons/sticks) are shown. Detailed interactions between the residues of the epitopes and the TAP transporter residues are shown, with hydrogen bond formation indicated with green dots. H bonds are shown in green dots with lengths in angstroms. TAP: transporter associated with antigen processing.



Characterization and Molecular Interaction Analysis of the Designed MEVs with Immune Receptors

Characterization of the Designed MEVs

Physicochemical Property Analysis of the Designed MEVs

ProtParam analysis of both the CTL and HTL MEVs was performed to analyze their physiochemical properties. The empirical physiochemical properties of the CTL and HTL MEVs are given in Table 4. The aliphatic indices and grand averages of hydropathicity of both MEVs indicate their globular and hydrophilic natures. The instability index scores of both MEVs indicates the stable nature of the protein molecules.

Table 4. Physicochemical property analysis based on the amino acid sequences of the designed CTL and HTL MEVs.

Property	Cytotoxic T lymphocyte multiepitope vaccine	Helper T lymphocyte multiepitope vaccine
Length (amino acids)	704	810
Molecular weight (kilodaltons)	72.62	82.80
Theoretical protrusion index	9.70	8.64
Expected half-life (hours)		
Escherichia coli	10	10
Yeast	30	30
Mammalian cell	20	20
Aliphatic index	61.09	96.43
Grand average of hydropathicity	–0.090	0.501
Instability index	44.31	40.28

IFNγ-Inducing Epitope Prediction

IFNγ-inducing epitopes are involved in both the adaptive and the innate immune response. IFNγ-inducing 15mer peptide epitopes were screened from the amino acid sequences of the CTL and HTL MEVs using the IFNepitope server. A total of 20 CTL MEV and 20 HTL MEV INFγ-inducing positive epitopes with a score ≥1 were shortlisted (Supplementary Table S2, Multimedia Appendix 1).

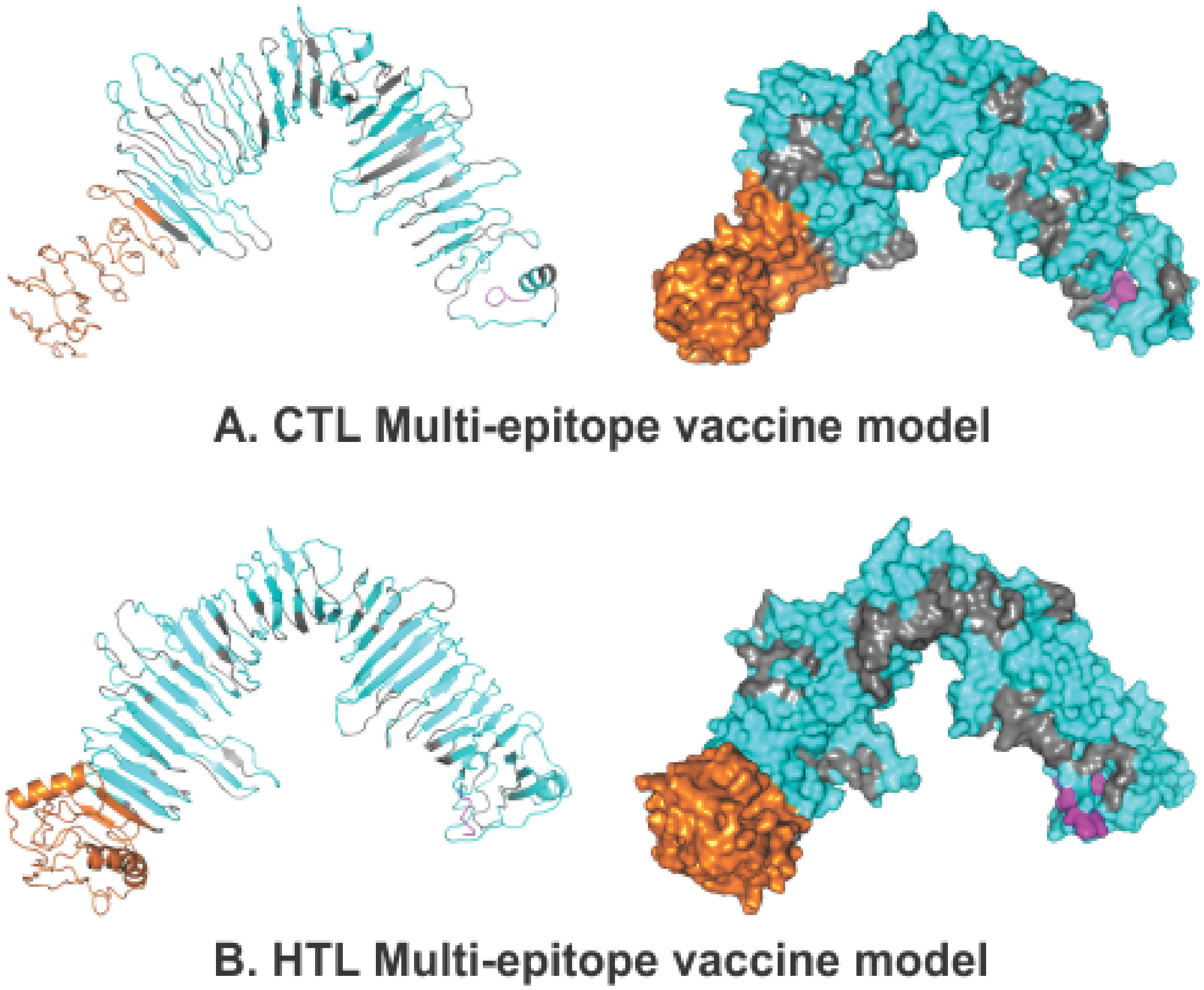
Allergenicity and Antigenicity Prediction of the MEVs

Both the CTL and HTL MEVs were found to be nonallergenic by the AlgPred analysis (scores of –0.95185601 and –1.1293352, respectively; the threshold was –0.4). The CTL and HTL MEVs were also indicated by VaxiJen analysis to be probable antigens (prediction scores of 0.4485 and 0.4215, respectively; the default threshold is 0.4). Hence, with the mentioned analysis tools, both the CTL and HTL MEVs are predicted to be nonallergic and antigenic in nature.

Tertiary Structure Modeling, Refinement, and Validation of the MEVs

3D homology models were generated for both the CTL and HTL MEvs using the I-TASSER modeling tool (Figure 3). The models were generated for the CTL MEV (PDB ID: 5n8pA, normal Z-score of 1.49, Cov of 0.92, TM-score of 0.916, and RMSD of 1.04 Å) and the HTL MEC (PDB ID: 5n8pA, normal Z-score of 1.52, Cov of 0.97, TM-score of 0.916, and RMSD of 1.04 Å).

Figure 3. Tertiary structure modelling of the CTL and HTL multiepitope vaccines. The epitopes are shown in cyan. The adjuvant (Ov-ASP-1) is shown in orange. The linkers are shown in gray, and the 6xHis tag is shown in magenta. Cartoon and surface presentations of both the MEVs are shown. CTL: cytotoxic T lymphocyte. HTL: helper T lymphocyte.


A. CTL Multi-epitope vaccine model
B. HTL Multi-epitope vaccine model

The generated CTL and HTL 3D models were both further refined by ModRefiner to repair any gaps, followed by GalaxyRefine refinement. The refinement by ModRefiner showed TM-scores of 0.9189 and 0.9498 for the CTL and HTL models, respectively; because these values are close to 1, the initial and refined models were structurally similar. After refinement, the RMSDs for the CTL and HTL models with respect to the initial model were 3.367 Å and 2.318 Å, respectively. Further, both the CTL and HTL MEV models were refined with GalaxyRefine, and model 1 was chosen based on the best scoring parameters. The CTL MEV model refinement output model (Ramachandran favored 83.6%, GDT-HA 0.9371, RMSD 0.459, MolProbity 2.539, clash score 23.2, and poor rotamers 1.8) and the HTL MEV model refinement output model (Ramachandran favored 87.7%, GDT-HA 0.9552, RMSD 0.402, MolProbity 2.537, clash score 27.9, and poor rotamers 1.6) show that well-refined and acceptable models were generated for both the MEVs. After refinement, all the mentioned parameters were found to be significantly improved in comparison to the initial CTL and HTL MEV models (Supplementary Table S3, Multimedia Appendix 1).

Validation of the Refined Models of the CTL and HTL MEVs

Both the CTL and HTL models were analyzed with the RAMPAGE analysis tool after refinement. The refined CTL MEV model was found to have 85.8% residues in the favored region, 11.3% residues in the allowed region, and only 3.0% residues in the outlier region; meanwhile, the refined HTL MEV model was found to have 88.9% residues in the favored region, 8.9% residues in the allowed region, and only 2.2% residues in the outlier region (Supplementary Figure S5, Multimedia Appendix 1).

Linear and Discontinuous B-cell Epitope Prediction From the MEVs

Linear and discontinuous B-cell epitope prediction was performed to identify potential linear and discontinuous epitopes in the refined 3D models of the CTL and HTL MEVs utilizing the ElliPro tool available on the IEDB server. The screening revealed that the CTL MEV carries 17 linear and 2 potential discontinuous B cell epitopes and the HTL MEV carries 17 linear and 4 potential discontinuous epitopes. The wide range of the PI scores of the linear and discontinuous epitopes in the CTL and HTL MEVs show the high potential of the epitopes to cause humoral immune response (PI scores: CTL MEV linear and discontinuous B cell epitopes: 0.511-0.828 and 0.664-0.767, respectively; HTL MEV linear and discontinuous B cell epitopes: 0.518-0.831 and 0.53-0.776, respectively) (Supplementary Tables S4, S5, S6, and S7, Multimedia Appendix 1).

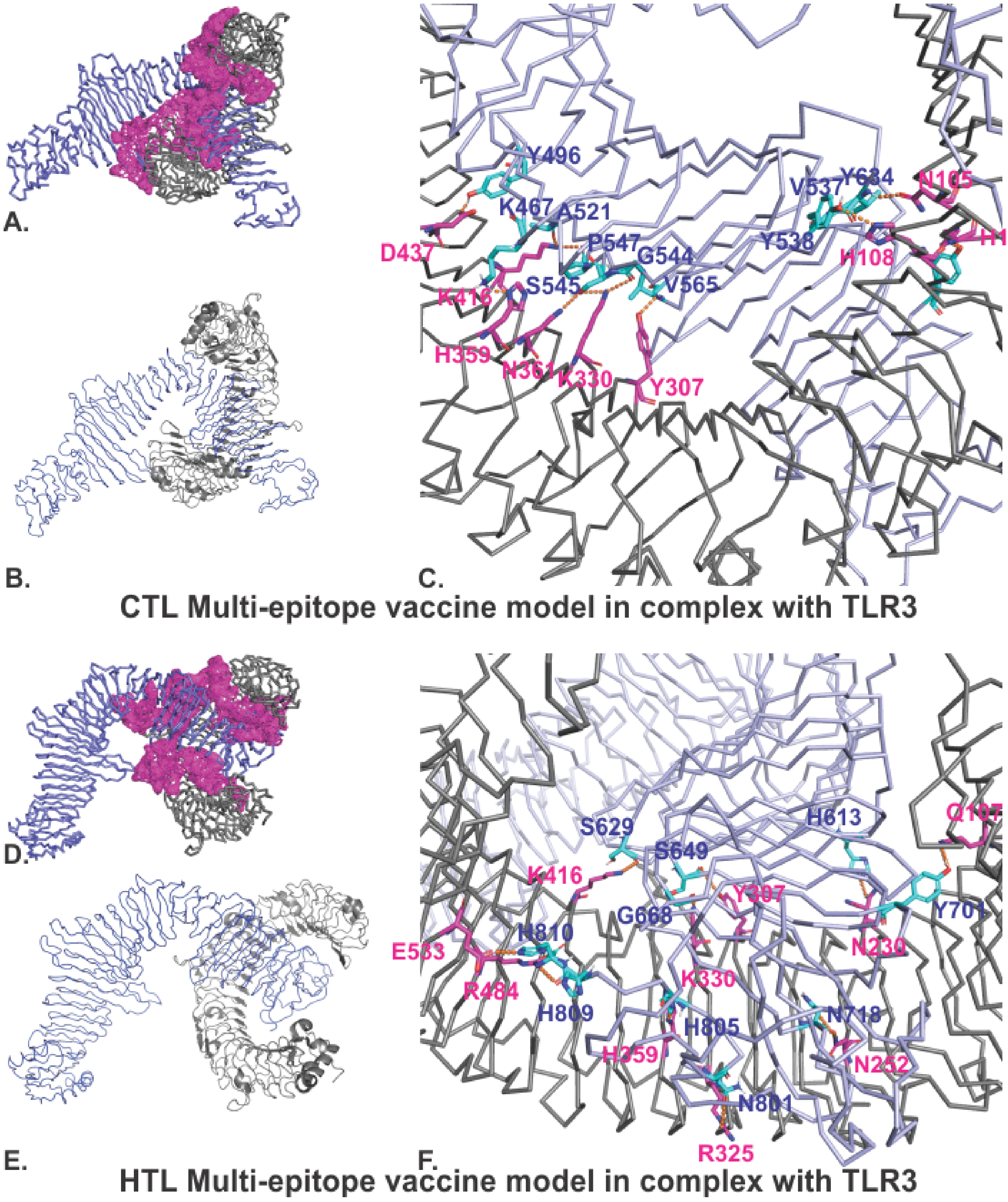
Molecular Interaction Analysis of the MEVs With Immunological Receptors

Molecular Docking Studies of the MEVs With TLR3

The refined models of both the CTL and HTL MEVs were further studied for their molecular interactions with the ectodomain of human TLR3. Therefore, molecular docking of the CTL and HTL MEV models with the TLR3 crystal structure model (PDB ID: 2A0Z) was performed utilizing the PatchDock tool. The generated docking conformations with the highest scores of 20776 and 20350 for the CTL and HTL MEVs, respectively, were chosen for further study. The highest docking score indicates the best geometric shape complementarity fitting conformation of the MEV and the TLR3 receptor as predicted by the PatchDock tool. Both the CTL and HTL MEVs fit into the ectodomain region of TLR3 after docking, involving numerous molecular interactions with active site residues of the TLR3 cavity region (Figure 4A, C, D, and F). As shown in Figure 4A and 4D, an entire patch of the TLR3 cavity surface is involved in the molecular interactions with the MEVs, favoring the formation of molecular complexes between the MEVs and the TLR3 ectodomain cavity. Paticular residues involved in this interaction are shown in Fig 4C and 4F (CTL:TLR3: Y496:D437, K467:H359, A521:K416, P547:K416, S545:N361, G544:K330, V565:Y307, Y538:H129, V537:N105, Y634:H108. HTL:TLR3: S629:K416, S649:Y307, G668:K330, H810:E533, H809:R484, H805:H359, N801:R325, H613:N230, N252:N718, Y701:Q107). The CTL and HTL MEVs showed the formation of multiple hydrogen bonds within the ectodomain cavity region of TLR3.

B-factor analysis of the MEV-TLR3 complexes was also performed. The B-factor indicates the displacement of the atomic positions from an average (mean) value, as in, the more flexible the atom, the larger its displacement from the mean position (mean-squares displacement) (Figure 4B, 4D). PDBsum [60] was used to calculate patches on the TLR3 receptor indicating the region of binding sites. The B-factor analysis of the CTL and HTL MEVs bound to the TLR3 receptor shows that most of the regions of the MEVs bound to TLR3 are stable. The B-factor analysis is represented by a VIBGYOR color presentation, where blue represents a low B-factor and red represents a high B-factor (Figure 4B, 4D). These results suggest tendencies toward stable complex formation for both the CTL and HTL MEVs with the ectodomain of the human TLR3 receptor.

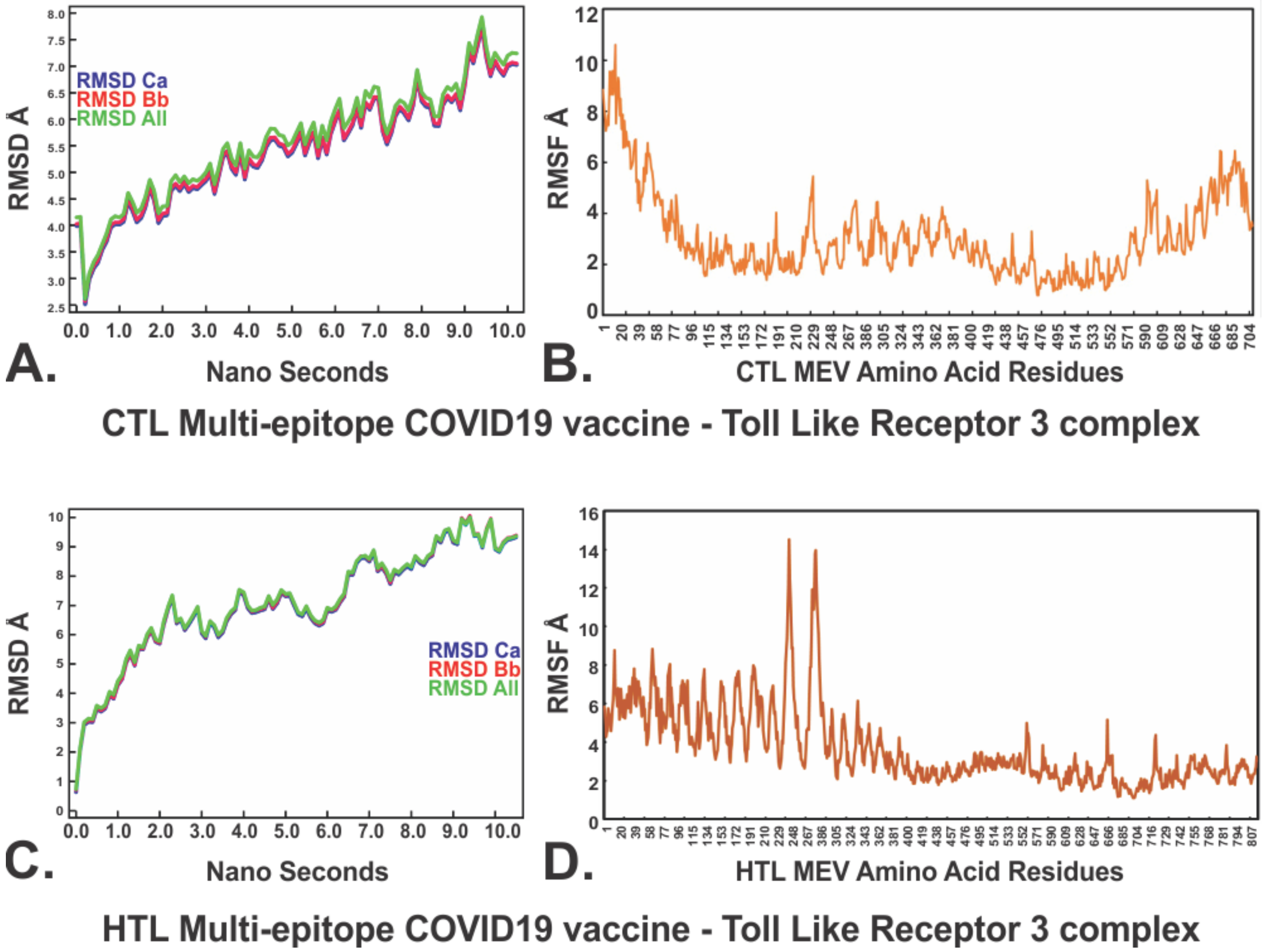
Figure 4. Molecular docking studies of the CTL and HTL MEVs with TLR3. (A), (D): The docking complexes of CTL-TLR3 and HTL-TLR3 with patches on the TLR3 receptor indicating the region of binding sites calculated by PDBsum [60]. (C), (F): Detailed molecular interactions between the binding site residues of the CTL and HTL MEVs and TRL3 (CTL, HTL: cyan; TLR3: magenta). Hydrogen bond formation is shown by orange dotted lines. (B), (E): B-factors of the docked MEVs to the TLR3 receptor. The presentation is in VIBGYOR color, with blue showing a low B-factor and red showing a high B-factor. Most of the MEV regions are blue, showing low B-factors; this indicates the formation of stable complexes with the TLR3 receptor. CTL, cytotoxic T lymphocyte. HTL, helper T lymphocyte. TLR3, toll-like receptor 3.



Molecular Dynamics Simulation Study of the Complexes of the MEVs with TLR3

Both the complexes CTL-TLR3 and HTL-TLR3 were further subjected to molecular dynamics simulation analysis to investigate the stability of the molecular interactions involved. Both the MEV-TLR3 complexes showed very convincing and reasonably stable RMSD values for the C_{α}, backbone, and all atoms (CTL-TLR3 complex: approximately 4-7.5 Å; HTL-TLR3 complex: approximately 3.0-9.8 Å) which stabilized toward the end (Figure 5A and 5C). The RMSDs of both complexes remained in the abovementioned RMSD range for a given time window of 10 ns at reasonably invariable temperature (approximately 278 K) and pressure (approximately 1 atm). The molecular docking and molecular dynamics simulation studies of all the MEV-TLR complexes indicate tendencies toward stable complex formation. Almost all the AA residues of the CTL and HTL MEVs complexed with TLR3 showed RMSFs in an acceptable range (approximately 2-6 Å) (Figure 5B and 5D). These results indicate that both the CTL-TLR3 and HTL-TLR3 complexes are stable, with acceptable molecular interaction tendencies.

Figure 5. Molecular dynamics simulations of the CTL and HTL MEVs with TLR3. (A), (C): Root mean square deviations for the Cα, backbone, and all atoms for the CTL MEV-TLR3 complex and the HTL MEV-TLR3 complex. (B), (D): Root mean square fluctuations of all the amino acid residues of the CTL MEV and the HTL MEV in complex with the TLR3 immune receptor. Å: angstroms. COVID-19: coronavirus disease. CTL: cytotoxic T lymphocyte. HTL: helper T lymphocyte. MEV: multiepitope vaccine. TL3: toll-like receptor 3. RMSD: root mean square deviation. RMSD Ca: root mean square deviation for the alpha carbon atoms. RMSD Bb: root mean square deviation for the backbone atoms. RMSD All: root mean square deviation for all atoms. RMSF: root mean square fluctuation.



In Silico Analysis of cDNA of the MEVs for Cloning and Expression Potency in a Mammalian Host Cell Line

cDNA optimized for CTL and HTL expression in a mammalian (human) host cell line was generated using the Java Codon Adaptation Tool. Further, the generated optimized cDNAs for both the MEVs were analyzed using the GenScript Rare Codon Analysis Tool. The analysis revealed that the codon-optimized cDNAs of both the CTL and HTL MEVs have crucial and favorable compositions for high-level expression in a mammalian cell line (CTL MEV: GC content 70.40%, CAI score 1.00, and 0% tandem rare codons; HTL MEV: GC content 69.26%, CAI score 1.00, and 0% tandem rare codons). Ideally, the GC content of cDNA should be 30%-70%; a CAI score that indicates the possibility of cDNA expression in a chosen expression system should be between 0.8 and 1.0; and the tandem rare codon frequency that indicates the presence of low-frequency codons in cDNA should be <30%. Tandem rare codons may hinder proper expression of the cDNA or even interrupt the translational machinery of the chosen expression system. Therefore, as per the GenScript Rare Codon analysis, the cDNAs of both the MEVs satisfy all the mentioned parameters and are predicted to have high expression in the mammalian (human) host cell line.

Discussion

Principal Findings

In the present study, we have reported the design of CTL and HTL multiepitope-based vaccine candidates against SARS-CoV-2 infection. These MEVs are composed of multiple CTL and HTL epitopes with truncated Ov-ASP-1 as an adjuvant at the N termini of both the MEVs. To design the abovementioned MEVs, we screened potential CTL and HTL epitopes from the entire proteome of the SARS-CoV-2 coronavirus. The screened epitopes showed potential due to their low IC_{50} values (nM) for HLA interaction, high immunogenicity, nontoxicity, favorable TAP cavity interaction, high conservancy, and high percentile rankings (determined using the IEDB MHC-I Binding Predictions and MHC-II Binding Predictions tools). Furthermore, the population coverage of the shortlisted 38 CTL and 33 HLT epitopes and their HLA allele binders was analyzed; the results were very satisfying, with a total world population coverage of 96.10%. Moreover, 12 B cell epitopes with lengths of 4-20 AAs were screened that showed full or partial overlap with the shortlisted CTL and HTL epitopes. All the shortlisted epitopes were highly conserved, with a conservancy range between 97.08% and 100%; at the same time, all the epitopes were nontoxic. All the shortlisted CTL, HTL, and B cell epitopes were also shown to overlap with each other, which further indicated their highly immunogenic nature. The overlapping epitopes of CTL and HTL were chosen for further analysis of their molecular interactions with HLA alleles and the TAP cavity. Molecular interaction analysis of the chosen overlapping epitopes with their respective HAL allele binders showed very favorable results. Similarly, the molecular interaction analysis of the CTL epitopes within the TAP cavity showed very favorable results for the smooth passage of the epitopes through the cavity from the cytoplasmic end (C terminal) to the ER lumen end (N terminal) of the transmembrane transporter. Further, the two MEVs were designed and modeled utilizing a flexible linker (GGGGS). The chosen adjuvant (truncated Ov-ASP-1) was linked at the N terminal of both the MEVs using a rigid linker (EAAAK). Modeling and further refinement of both the MEVs was performed, and highly sterically acceptable models were generated. The molecular weights of both the MEVs were also very acceptable for expression in suitable systems (CTL MEV: 72.62 kilodaltons, HTL MEV: 82.80 kDa). Further, both the MEVs were shown to contain 20 INFγ-inducing positive epitopes. Both the MEVs were also analyzed to contain numerous linear (CTL: 17, HTL: 17) and discontinuous (CTL: 2, HTL: 4) B cell epitopes. Both the MEVs were analyzed and found to be nonallergenic but antigenic in nature.

Furthermore, both the CTL and HTL MEVs were analyzed for their molecular interactions with the immune receptor TLR3. TLRs act as sentinels for the human immune system; therefore, favorable and stable interactions of both the MEVs with TLR3 are essential. In our study, we confirmed the stable interactions of both the CTL and HTL MEVs with the TLR3 receptor. Molecular docking studies revealed that numerous residues of both MEVs are involved in the formation of polar contacts with TLR3 receptor AA residues. Furthermore, the molecular dynamics studies confirmed stable molecular interactions between both MEVs and TLR3 based on the acceptable RMSDs for the backbones of both the CTL-MEV-TLR3 and HTL-MEV-TLR3 complexes.

Moreover, both MEVs were shown to have very favorable expression in vitro. We analyzed the codon-biased cDNAs for both the CTL and HTL MEVs for the mammalian (human) cell line expression system and found very acceptable CG contents and CAIs as well as 0% tandem rare codons. Therefore, both the designed MEVs can be expressed in the chosen expression system and further tested in vivo as potential vaccine candidates against SARS-CoV-2 infection.

Conclusion

We have designed and proposed two MEVs derived from multiple CTL and HTL epitopes against SARS-CoV-2 (COVID-19). The chosen CTL and HTL epitopes show significant sequence overlap with screened linear B cell epitopes. The shortlisted CTL and HTL epitopes were used to design CTL and HTL MEVs. Tertiary models of both the generated CTL and HTL MEVs were shown to contain potential linear and discontinuous B cell epitopes as well as potential INFγ epitopes. Therefore, the designed MEVs are predicted to be capable of eliciting humoral and cellular immune responses. Because Ov-ASP-1 binds to APCs and triggers pro-inflammatory cytokine production via TLR3, truncated Ov-ASP-1 was used as an adjuvant at the N termini of both the CTL and HTL MEV models. The molecular interactions of the chosen overlapping clustering epitopes with their respective HLA allele binders were validated by molecular docking studies. The molecular interactions of the chosen CTL epitopes with the TAP transporter cavity were also analyzed. Analysis of the average world population coverage by both the shortlisted CTL and HTL epitopes combined revealed coverage of 96.10% of the world population. The molecular interaction analysis of both the CTL and HTL MEVs with the immunoreceptor TLR3 showed very convincing structural fitting of the MEVs into the ectodomain of the TLR3 cavity. This result was further confirmed by molecular dynamics simulation studies of both the CTL-MEV-TLR3 and HTL-MEV-TLR3 complexes, indicating tendencies toward stable molecular complex formation of both MEVs with TLR3. cDNAs for both MEVs were generated considering codon-biasing for expression in a mammalian (human) host cell line. Both cDNAs were optimized with respect to their GC content and zero tandem rare codons to increase their possibility of high expression in the mammalian host cell line (human). Therefore, for further studies, both the designed CTL and HTL MEVs could be cloned, expressed, and tested for in vivo validation and animal trials as potential vaccine candidates against SARS-CoV-2 infection.

Acknowledgments

We are thankful to the Indian Council for Medical Research (ICMR) for providing a SRF fellowship to SV (ISRM/11(98)/2019). We are also thankful to the ICMR National Institute of Malaria Research for providing infrastructural and computational facilities for the experiments. We also thank Mangalayatan University for providing infrastructural facilities for the experiments.

Authors' Contributions

The protocol was designed by SS and MK. The methodology was performed by SS, SV, MK, and RK. The global economic risk analysis was performed by RKB. Data analysis, scientific writing, and revision of the article were performed by SS, SV, MK, RK, RKB, AKS, HJS, MK, and KCP.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[PDF File (Adobe PDF File), 1840 KB - bioinform_v1i1e19371_app1.pdf]

References

1. World Health Organization. 2020 Feb 20. WHO Director-General's opening remarks at the media briefing on COVID-19 URL: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---3-march-2020 [accessed 2020-05-29]
2. World Health Organization. 2020 Mar 20. Coronavirus disease 2019 (COVID-19) Situation Report – 61 URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200321-sitrep-61-covid-19.pdf?sfvrsn=6aa18912_2) [accessed 2020-03-21]
3. United Nations. 2020 Mar 09. The economic impact of COVID-19: Can policy makers avert a multi-trillion dollar crisis? URL: https://unctad.org/en/pages/PressRelease.aspx?OriginalVersionID=548) [accessed 2020-05-29]
4. National Center for Biotechnology Information. SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences URL: https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/ [accessed 2020-05-29]
5. MacDonald AJ, Cao L, He Y, Zhao Q, Jiang S, Lustigman S. rOv-ASP-1, a recombinant secreted protein of the helminth Onchocercavolvulus, is a potent adjuvant for inducing antibodies to ovalbumin, HIV-1 polypeptide and SARS-CoV peptide antigens. Vaccine 2005 May 16;23(26):3446-3452 [FREE Full text] [doi: 10.1016/j.vaccine.2005.01.098] [Medline: 15837368]
6. Guo J, Yang Y, Xiao W, Sun W, Yu H, Du L, et al. A truncated fragment of Ov-ASP-1 consisting of the core pathogenesis-related-1 (PR-1) domain maintains adjuvanticity as the full-length protein. Vaccine 2015 Apr 15;33(16):1974-1980 [FREE Full text] [doi: 10.1016/j.vaccine.2015.02.053] [Medline: 25736195]
7. He Y, Barker SJ, MacDonald AJ, Yu Y, Cao L, Li J, et al. Recombinant Ov-ASP-1, a Th1-biased protein adjuvant derived from the helminth Onchocerca volvulus, can directly bind and activate antigen-presenting cells. J Immunol 2009 Apr 01;182(7):4005-4016 [FREE Full text] [doi: 10.4049/jimmunol.0800531] [Medline: 19299698]
8. Oldham M, Grigorieff N, Chen J. Structure of the transporter associated with antigen processing trapped by herpes simplex virus. Elife 2016 Dec 09;5 [FREE Full text] [doi: 10.7554/eLife.21829] [Medline: 27935481]
9. Abele R, Tampé R. The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing. Physiology (Bethesda) 2004 Aug;19:216-224 [FREE Full text] [doi: 10.1152/physiol.00002.2004] [Medline: 15304636]
10. Antoniou AN, Powis SJ, Elliott T. Assembly and export of MHC class I peptide ligands. Curr Opin Immunol 2003 Feb;15(1):75-81. [doi: 10.1016/s0952-7915(02)00010-9]
11. Delneste Y, Beauvillain C, Jeannin P. Innate immunity: structure and function of TLRs. Article in French. Med Sci (Paris) 2007 Jan;23(1):67-73 [FREE Full text] [doi: 10.1051/medsci/200723167] [Medline: 17212934]
12. Totura AL, Whitmore A, Agnihothram S, Schäfer A, Katze MG, Heise MT, et al. Toll-Like Receptor 3 Signaling via TRIF Contributes to a Protective Innate Immune Response to Severe Acute Respiratory Syndrome Coronavirus Infection. mBio 2015 May 26;6(3). [doi: 10.1128/mbio.00638-15]
13. Farina C, Krumbholz M, Giese T, Hartmann G, Aloisi F, Meinl E. Preferential expression and function of Toll-like receptor 3 in human astrocytes. J Neuroimmunol 2005 Feb;159(1-2):12-19. [doi: 10.1016/j.jneuroim.2004.09.009] [Medline: 15652398]
14. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz M, et al. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. Cell Mol Life Sci 2005 May;62(9):1025-1037. [doi: 10.1007/s00018-005-4528-2] [Medline: 15868101]
15. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhütter HG. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 2003 Aug 15;171(4):1741-1749 [FREE Full text] [doi: 10.4049/jimmunol.171.4.1741] [Medline: 12902473]
16. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 2009 Jan;61(1):1-13 [FREE Full text] [doi: 10.1007/s00251-008-0341-z] [Medline: 19002680]

17. Calis J, Maybeno M, Greenbaum J, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC class I presented peptides that enhance immunogenicity. PLoS Comput Biol 2013 Oct;9(10):e1003266 [FREE Full text] [doi: 10.1371/journal.pcbi.1003266] [Medline: 24204222]
18. Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics 2010 Nov 22;11:568 [FREE Full text] [doi: 10.1186/1471-2105-11-568] [Medline: 21092157]
19. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. Immunome Res 2008 Jan 25;4:2 [FREE Full text] [doi: 10.1186/1745-7580-4-2] [Medline: 18221540]
20. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 2007 Jul 04;8:238 [FREE Full text] [doi: 10.1186/1471-2105-8-238] [Medline: 17608956]
21. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol 1999 Jun;17(6):555-561. [doi: 10.1038/9858] [Medline: 10385319]
22. Bui H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics 2006 Mar 17;7:153 [FREE Full text] [doi: 10.1186/1471-2105-7-153] [Medline: 16545123]
23. Larsen J, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res 2006 Apr 24;2:2 [FREE Full text] [doi: 10.1186/1745-7580-2-2] [Medline: 16635264]
24. Bui H, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. BMC Bioinformatics 2007 Sep 26;8:361. [doi: 10.1186/1471-2105-8-361] [Medline: 17897458]
25. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Open Source Drug Discovery Consortium, et al. In silico approach for predicting toxicity of peptides and proteins. PLoS One 2013;8(9):e73957 [FREE Full text] [doi: 10.1371/journal.pone.0073957] [Medline: 24058508]
26. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011 Oct 11;7:539 [FREE Full text] [doi: 10.1038/msb.2011.75] [Medline: 21988835]
27. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 2006 Jan 15;22(2):195-201. [doi: 10.1093/bioinformatics/bti770] [Medline: 16301204]
28. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008 Apr;71(1):261-277. [doi: 10.1002/prot.21715] [Medline: 17932912]
29. Shen Y, Maupetit J, Derreumaux P, Tufféry P. Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. J Chem Theory Comput 2014 Oct 14;10(10):4745-4758. [doi: 10.1021/ct500592m] [Medline: 26588162]
30. Bell JK, Botos I, Hall PR, Askins J, Shiloach J, Segal DM, et al. The molecular structure of the Toll-like receptor 3 ligand-binding domain. Proc Natl Acad Sci USA 2005 Aug 02;102(31):10976-10980 [FREE Full text] [doi: 10.1073/pnas.0505077102] [Medline: 16043704]
31. Duhovny D, Nussinov R, Wolfson H. Efficient Unbound Docking of Rigid Molecules. In: Algorithms in Bioinformatics. WABI 2002. Lecture Notes in Computer Science, vol 2452. Berlin: Springer; 2020 Oct 10 Presented at: International Workshop on Algorithms in Bioinformatics; 2002; Rome, Italy p. 185-200. [doi: 10.1007/3-540-45784-4_14]
32. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson H. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 2005 Jul 01;33(Web Server issue):W363-W367 [FREE Full text] [doi: 10.1093/nar/gki481] [Medline: 15980490]
33. Hu W, Li F, Yang X, Li Z, Xia H, Li G, et al. A flexible peptide linker enhances the immunoreactivity of two copies HBsAg preS1 (21-47) fusion protein. J Biotechnol 2004 Jan 08;107(1):83-90. [doi: 10.1016/j.jbiotec.2003.09.009] [Medline: 14687974]
34. Hajighahramani N, Nezafat N, Eslami M, Negahdaripour M, Rahmatabadi SS, Ghasemi Y. Immunoinformatics analysis and in silico designing of a novel multi-epitope peptide vaccine against Staphylococcus aureus. Infect Genet Evol 2017 Mar;48:83-94. [doi: 10.1016/j.meegid.2016.12.010] [Medline: 27989662]
35. Chen X, Zaro JL, Shen W. Fusion protein linkers: property, design and functionality. Adv Drug Deliv Rev 2013 Oct;65(10):1357-1369 [FREE Full text] [doi: 10.1016/j.addr.2012.09.039] [Medline: 23026637]
36. Srivastava S, Kamthania M, Kumar Pandey R, Kumar Saxena A, Saxena V, Kumar Singh S, et al. Design of novel multi-epitope vaccines against severe acute respiratory syndrome validated through multistage molecular interaction and dynamics. J Biomol Struct Dyn 2019 Oct;37(16):4345-4360. [doi: 10.1080/07391102.2018.1548977] [Medline: 30457455]
37. Srivastava S, Kamthania M, Singh S, Saxena A, Sharma N. Structural basis of development of multi-epitope vaccine against Middle East respiratory syndrome using in silico approach. IDR 2018 Nov;Volume 11:2377-2391. [doi: 10.2147/idr.s175114]

38. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005:571-607.
39. Nagpal G, Gupta S, Chaudhary K, Dhanda SK, Prakash S, Raghava GPS. VaccineDA: Prediction, design and genome-wide screening of oligodeoxynucleotide-based vaccine adjuvants. Sci Rep 2015 Jul 27;5:12478 [FREE Full text] [doi: 10.1038/srep12478] [Medline: 26212482]
40. Dhanda SK, Vir P, Raghava GP. Designing of interferon-gamma inducing MHC class-II binders. Biol Direct 2013 Dec 05;8:30 [FREE Full text] [doi: 10.1186/1745-6150-8-30] [Medline: 24304645]
41. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res 2006 Jul 01;34(Web Server issue):W202-W209 [FREE Full text] [doi: 10.1093/nar/gkl343] [Medline: 16844994]
42. Doytchinova I, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 2007 Jan 05;8:4 [FREE Full text] [doi: 10.1186/1471-2105-8-4] [Medline: 17207271]
43. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010 Apr;5(4):725-738 [FREE Full text] [doi: 10.1038/nprot.2010.5] [Medline: 20360767]
44. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 2011 Nov 16;101(10):2525-2534 [FREE Full text] [doi: 10.1016/j.bpj.2011.10.024] [Medline: 22098752]
45. Ko J, Park H, Heo L, Seok C. GalaxyWEB server for protein structure prediction and refinement. Nucleic Acids Res 2012 Jul;40(Web Server issue):W294-W297 [FREE Full text] [doi: 10.1093/nar/gks493] [Medline: 22649060]
46. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics 2013 Jul 01;29(13):i266-i273 [FREE Full text] [doi: 10.1093/bioinformatics/btt211] [Medline: 23812992]
47. Shin WH, Lee GR, Heo L, Lee H, Seok C. Prediction of protein structure and interaction by GALAXY protein modeling programs. Bio Design 2014 Mar 31;2(1):1-11.
48. Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003 Feb 15;50(3):437-450 [FREE Full text] [doi: 10.1002/prot.10286] [Medline: 12557186]
49. Ramakrishnan C, Ramachandran G. Stereochemical Criteria for Polypeptide and Protein Chain Conformations. Biophys J 1965 Nov;5(6):909-933. [doi: 10.1016/s0006-3495(65)86759-5]
50. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol 2012;8(12):e1002829 [FREE Full text] [doi: 10.1371/journal.pcbi.1002829] [Medline: 23300419]
51. Ponomarenko J, Bui H, Li W, Fusseder N, Bourne P, Sette A, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 2008 Dec 02;9:514 [FREE Full text] [doi: 10.1186/1471-2105-9-514] [Medline: 19055730]
52. Krieger E, Vriend G. New ways to boost molecular dynamics simulations. J Comput Chem 2015 May 15;36(13):996-1007 [FREE Full text] [doi: 10.1002/jcc.23899] [Medline: 25824339]
53. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 2015 Aug 11;11(8):3696-3713 [FREE Full text] [doi: 10.1021/acs.jctc.5b00255] [Medline: 26574453]
54. Case D, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, et al. Amber 14. 2014. The FF14SB force field URL: https://ambermd.org/doc12/Amber14.pdf [accessed 2020-05-29]
55. Toukmaji A, Sagui C, Board J, Darden T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. The Journal of Chemical Physics 2000 Dec 22;113(24):10913-10927. [doi: 10.1063/1.1324708]
56. Morla S, Makhija A, Kumar S. Synonymous codon usage pattern in glycoprotein gene of rabies virus. Gene 2016 Jun 10;584(1):1-6. [doi: 10.1016/j.gene.2016.02.047] [Medline: 26945626]
57. Wu X, Wu S, Li D, Zhang J, Hou L, Ma J, et al. Computational identification of rare codons of Escherichia coli based on codon pairs preference. BMC Bioinformatics 2010 Jan 28;11:61 [FREE Full text] [doi: 10.1186/1471-2105-11-61] [Medline: 20109184]
58. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe 2020 Apr 08;27(4):671-680.e2 [FREE Full text] [doi: 10.1016/j.chom.2020.03.002] [Medline: 32183941]
59. Schrödinger. The PyMOL Molecular Graphics System, Version 2.0 URL: https://www.schrodinger.com/pymol [accessed 2020-05-29]
60. Laskowski RA, Hutchinson E, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 1997 Dec;22(12):488-490. [doi: 10.1016/s0968-0004(97)01140-7]

Abbreviations

AA: amino acid
Cα: alpha carbon
cDNA: complementary DNA
COVID-19: coronavirus disease
CTL: cytotoxic T lymphocyte
E protein: envelope protein
ER: endoplasmic reticulum
HLA: human leukocyte antigen
HTL: helper T lymphocyte
IC50: half maximal inhibitory concentration
IEDB: Immune Epitope Database
IFNγ: interferon gamma
MERCI: Motif-EmeRging and with Classes-Identification
MEV: multiepitope vaccine
MHC: major histocompatibility complex
M protein: membrane protein
NCBI: National Center for Biology Information
N protein: nucleocapsid protein
ORF: Open Reading Frame
Ov-ASP-1: Onchocerca volvulus activation-associated secreted protein-1
PDB: Protein Data Bank
PI: protrusion index
RMSD: root mean square deviation
RMSF: root mean square fluctuation
SARS-CoV-2: severe acute respiratory syndrome coronavirus 2
S protein: surface protein
TAP: transporter associated with antigen processing
TLR: toll-like receptor
TLR3: toll-like receptor 3
TrEMBL: Translated European Molecular Biology Laboratory
VIBGYOR: violet-indigo-blue-green-yellow-orange-red
YASARA: Yet Another Scientific Artificial Reality Application

Edited by G Eysenbach; submitted 15.04.20; peer-reviewed by B Rathi, C Chen; comments to author 20.05.20; revised version received 26.05.20; accepted 27.05.20; published 19.06.20.

Please cite as:
Srivastava S, Verma S, Kamthania M, Kaur R, Badyal RK, Saxena AK, Shin HJ, Kolbe M, Pandey KC
Structural Basis for Designing Multiepitope Vaccines Against COVID-19 Infection: In Silico Vaccine Design and Validation
JMIR Bioinformatics Biotechnol 2020;1(1):e19371
URL: https://bioinform.jmir.org/2020/1/e19371
doi:10.2196/19371
PMID:32776022

Original Paper

The Novel Coronavirus Enigma: Phylogeny and Analyses of Coevolving Mutations Among the SARS-CoV-2 Viruses Circulating in India

Anindita Banerjee[1*], PhD; Rakesh Sarkar[1*], MSc; Suvrotoa Mitra[1], MSc; Mahadeb Lo[1], MSc; Shanta Dutta[1], PhD; Mamta Chawla-Sarkar[1], PhD

Indian Council of Medical Research-National Institute of Cholera and Enteric Diseases, Kolkata, India

[*]these authors contributed equally

Corresponding Author:
Mamta Chawla-Sarkar, PhD
Indian Council of Medical Research-National Institute of Cholera and Enteric Diseases
P-33, CIT Road, Scheme-XM, Beliaghata,
Kolkata, 700010
India
Phone: 91 9830660999
Email: chawlam70@gmail.com

Abstract

Background: The RNA genome of the emerging novel coronavirus is rapidly mutating, and its human-to-human transmission rate is increasing. Hence, temporal dissection of their evolutionary dynamics, the nature of variations among different strains, and understanding the single nucleotide polymorphisms in the endemic settings are crucial. Delineating the heterogeneous genomic constellations of this novel virus will help us understand its complex behavior in a particular geographical region.

Objective: This is a comprehensive analysis of 95 Indian SARS-CoV-2 genome sequences available from the Global Initiative on Sharing All Influenza Data (GISAID) repository during the first 6 months of 2020 (January through June). Evolutionary dynamics, gene-specific phylogeny, and the emergence of the novel coevolving mutations in 9 structural and nonstructural genes among circulating SARS-CoV-2 strains across 12 different Indian states were analyzed.

Methods: A total of 95 SARS-CoV-2 nucleotide sequences submitted from India were downloaded from the GISAID database. Molecular Evolutionary Genetics Analysis, version X software was used to construct the 9 phylogenetic dendrograms based on nucleotide sequences of the SARS-CoV-2 genes. Analyses of the coevolving mutations were done in comparison to the prototype SARS-CoV-2 from Wuhan, China. The secondary structure of the RNA-dependent RNA polymerase/nonstructural protein NSP12 was predicted with respect to the novel A97V mutation.

Results: Phylogenetic analyses revealed the evolution of “genome-type clusters” and adaptive selection of “L”-type SARS-CoV-2 strains with genetic closeness to the bat severe acute respiratory syndrome–like coronaviruses. These strains were distant to pangolin or Middle East respiratory syndrome–related coronavirus strains. With regard to the novel coevolving mutations, 2 groups have been seen circulating in India at present, the “major group” (66/95, 69.4%) and the “minor group” (21/95, 22.1%) , harboring 4 and 5 coexisting mutations, respectively. The “major group” mutations fall in the A2a clade. All the minor group mutations, except 11083G>T (L37F, NSP6 gene), were unique to the Indian isolates.

Conclusions: This study highlights the rapidly evolving SARS-CoV-2 virus and the cocirculation of multiple clades and subclades. This comprehensive study is a potential resource for monitoring the novel mutations in the viral genome, interpreting changes in viral pathogenesis, and designing vaccines or other therapeutics.

(JMIR Bioinformatics Biotechnol 2020;1(1):e20735) doi:10.2196/20735

KEYWORDS

SARS-CoV-2; Indian isolates; phylogeny; nucleotide homology; novel coevolving mutations; NSP12/RdRP secondary structure; genetics; virus; evolution; mutation; COVID-19; genome; epidemiology; infectious disease

Introduction

The COVID-19 pandemic caused by the novel SARS-CoV-2 was initially reported from Wuhan, China in December 2019, but it spread across the world within 3 months [1]. As of July 21, 2020, more than 14.9 million people have been found to be infected by SARS-CoV-2, with a death toll of approximately 615,939 in more than 210 countries. Phylogenetic analyses reveal that SARS-CoV-2 clusters within the subgenus *Sarbecovirus* under the genus *Betacoronavirus* and has probably undergone zoonotic transmission from the bats through the possible intermediate host Malayan pangolins, culminating among humans [2]. The positive sense, single-stranded RNA genome of SARS-CoV-2 is continuously mutating and generating multiple clades within a short time span (December 2019 to June 2020). Hence, there is a need to dissect the complex evolutionary characteristics of this novel coronavirus, identifying the single nucleotide polymorphisms (SNPs) and other mutations among strains circulating across different parts of the world. Previous reports on the genetic and evolutionary dynamics of the SARS-CoV-2 virus have tried to deduce the mode of transmission that this virus made its way into humans from bats during the early phase of the pandemic, but many questions remain unanswered even though more sequence data has been made available. Therefore, studying the heterogeneous genomic constellations within specific geographical settings will help to understand its complex epidemiology and formulate region specific strategies to curb its spread and severity.

The first 3 cases from India with travel history to Wuhan were reported in Kerala during January 2020 [3]; subsequently, 4,02,529 active cases, 724,577 recovered cases, and 28,084 deaths have been officially recorded in India as of July 21, 2020, 8 AM India Standard Time GMT +5:30 [4]. India ranks third worldwide according to the number of COVID-19 infections and is geographically vulnerable to this novel virus, as it accounts for almost 6% of global and 3.5% of COVID-19–attributable mortality. In spite of high population density, poor hygiene conditions, and an overburdened health care system, the proportion of the total infected population is much lower when compared to other western countries, that is, 0.05% in India versus 0.87% in the United States, 0.73% in Brazil, 0.46% in Russia, and 0.4% in Italy. Though the average death rate due to SARS-CoV-2 infection in India (2.46%) is comparable to that of world (eg, the United States 3.88%, Europe 6.6%), 50% of deaths in India are attributable to the age group 40-64 years [5]. Thus, to understand the phylodynamics of circulating strains in India, this study was initiated to analyze the complete viral genome sequences submitted in the Global Initiative on Sharing All Influenza Data (GISAID) [6] from 95 SARS-CoV-2 representative strains circulating across 12 differentially affected states within India. To elucidate the possible ancestry, gene-wise phylogeny of these Indian strains has been deciphered with respect to other isolates reported from Europe, the United States, and China along with coronavirus strains belonging to other genera infecting humans and other animal hosts. The novel coevolving mutations among the Indian SARS-CoV-2 strains have also been analyzed.

Through this genome analyses and phylogenetic approach, we have attempted to focus on the natural evolution of SARS-CoV-2 from its existing ancestors within the zoonotic reservoir. Furthermore, analyzing the novel mutations accumulated within the viral genome over the period with reference to the Wuhan strains (clade O) will underscore their impact on the structure and function of viral proteins.

Methods

Sequence Mining

A total of 95 SARS-CoV-2 nucleotide sequences submitted from India from January to June 2020 were downloaded from the GISAID database for phylogenetic analyses and screening of novel mutations. Several other reference gene sequences of SARS-CoV-2 as well as other types of coronaviruses were downloaded from the GenBank database submitted from several other countries for dendrogram construction and further lineage analyses.

Phylogenetic Analyses and Screening of Mutations

Nine phylogenetic dendrograms were constructed with respect to 2 structural genes (spike and nucleocapsid) and 7 nonstructural genes (nonstructural protein [NSP]2, NSP3, NSP4, NSP6, NSP7, NSP8, and NSP12). Multiple sequence alignment for all the respective set of gene sequences was done using MUSCLE v3.8.31 (drive5). Amino acid sequences were deduced through TRANSEQ (EMBL-EBI). Phylogenetic dendrograms were constructed by Molecular Evolutionary Genetics Analysis, version X (MEGAX), using the maximum-likelihood statistical method (at 1000 bootstrap replicates) and using the best fit nucleotide substitution models for each dendrogram. The best fit models were determined through model testing parameter of MEGAX. Different novel coexisting mutations in the Indian strains were identified and analyzed in comparison to the prototype SARS-CoV-2 strain from Wuhan (MN908947.3/SARS-CoV-2 Wuhan-Hu-1).

Secondary Structure Prediction of RNA-Dependent RNA Polymerase Having A97V Mutation

We used the Chou and Fasman Secondary Structure Prediction (CFSSP) online server to predict the secondary structure of RNA-dependent RNA polymerase (RdRP)/NSP12 with novel A97V mutation [7].

Results

Phylogenetic Analysis of the Structural and Nonstructural Genes

Spike Gene

Among the 95 Indian study isolates, 93 strains clustered among themselves within the same lineage of *Betacoronavirus* SARS-CoV-2 in 3 different subclusters (A=53, B=12, and C=28 strains), while 2 strains, 1 from Telangana (EPI_ISL_ 431101) and the other from Maharashtra (EPI_ISL_ 479550), extruded out separately, close to subcluster C containing the clade-specific strains A1, A3, B1, B2, B4-1, and B4-2 in the phylogenetic dendrogram for the spike (S) gene. The prototype SARS-CoV-2 strain belonging to the O clade (MN908947.3/SARS-CoV-2/HUMAN/CHN/Wuhan-Hu-1/2019) was present in the same lineage with the Indian strains within subcluster C (>99% identity). Subcluster A comprised of the clade-specific A2 and A2a strains along with a tiger strain from the New York zoo and another carnivorous mammal, mink SARS-CoV-2. No specific pattern of temporal distribution of strains was observed among the 3 subclusters. All the representative Indian strains had 99%-100% nucleotide sequence homology among themselves. The Indian strains had 92.8%-93% and 83.5% homology with Bat (EPI_ISL_402131 /COV /BAT /YUNNAN /RATG13 /2013) and Pangolin coronavirus (EPI_ISL_410540 /COV /PANGOLIN /GUANGXI / P5L/2017), respectively. Homology was much less (75.8%-76.7%) with other bat severe acute respiratory syndrome (SARS)–like coronavirus strains (eg, MG772933.1/ SARS-LIKE-COV/BAT/ BAT-SL-COVZC45 /2017 and MG772934.1/ SARS-LIKE-COV/ BAT/ BAT-SL-COVZXC21 /2015), while Middle East respiratory syndrome–related coronavirus (MERS-CoV; KJ713299.1 /MERS-COV /CAMEL /SAU /KSA-CAMEL-376 /2013 and KU308549.1/MERS-COV/ HUMAN/ KOR/ SEOUL-SNU1-035/ 2015) were distantly related to the Indian SARS-CoV-2 strains (52.5%-52.9% identity; Multimedia Appendix 1).

Nucleocapsid Gene

The phylogenetic dendrogram for the nucleocapsid (N) gene revealed that out of 95 Indian study isolates, 92 strains clustered within the same lineage of *Betacoronavirus* SARS-CoV-2 in 3 different subclusters (A=33 strains, B=47 strains, and C=12 strains), while 3 strains from Tamil Nadu (EPI_ISL_ 458040), Gujarat (EPI_ISL_ 458107), and Delhi (EPI_ISL_ 435111) extruded out separately, close to subcluster C strains. Subcluster A comprised of strains from the earlier 3 months (January, February, and March), while subcluster B contained strains from the later 3 months. Subcluster C had mixed strains. The clade-specific strains (A1, A1a, A2, A2a, A5, B1, and B4-1) as well as the prototype SARS-CoV-2 strain O clade (MN908947.3/SARS-CoV-2/HUMAN/CHN/Wuhan-Hu-1/2019) clustered near subcluster B, while B4-2, A3, and B2 clade strains were close to subcluster C. All the representative Indian strains had >99.8% nucleotide identity among themselves as well as with the different clade-specific strains. The Indian strains had 91%-97% sequence identity with bat coronaviruses (EPI_ISL_402131 /COV/BAT /YUNNAN /RATG13 /2013, MG772933.1 /SARS-LIKE-COV /BAT /BAT-SL-COVZC45 /2017, and MG772934.1 /SARS-LIKE-COV /BAT/ BAT-SL-COVZXC21/ 2015) and 91% similarity with pangolin strains (EPI_ISL_410540 /COV /PANGOLIN/GUANGXI/ P5L /2017). In contrast to bat strains, the MERS-CoV strains (KJ713299.1/MERS-COV/CAMEL /SAU /KSA-CAMEL-376 /2013 and KU308549.1/MERS-COV/HUMAN /KOR /SEOUL-SNU1-035/2015) were genetically distant to the Indian SARS-CoV-2 strains (56.9%-57.3% identity; Multimedia Appendix 2).

RNA-Dependent RNA Polymerase Gene (RdRP/NSP12)

The phylogenetic dendrogram for the RdRP/NSP12 gene depicted that, among the 95 Indian study isolates, 93 strains clustered within the same lineage of *Betacoronavirus* SARS-CoV-2 into 3 subclusters (A=64 strains, B=22 strains, and C=7 strains). Two strains, one from Kerala (EPI_ISL_ 413523) and the other from Delhi (EPI_ISL_ 435111), were placed distant to these 3 subclusters in the dendrogram and were close to A1a, A2, A3, A5, B1, B2, B4-1, B4-2, and the prototype O clade strains. Subcluster A strains clustered with the A2a clade-specific strain while subcluster C clustered with A1. No temporal specificity was observed among the 3 subcluster strain distributions. All the Indian strains had >99.8% nucleotide identity among themselves as well as the different clade-specific strains. The prototype SARS-CoV-2 strain O clade (MN908947.3 /SARS-COV-2 /HUMAN /CHN /Wuhan-Hu-1 /2019) was distant to all the 3 subclusters. The Indian strains had 97.8% sequence homology with bat coronavirus (EPI_ISL_402131 /COV /BAT /YUNNAN /RATG13/2013) and 86.7%-88.6% similarity with both pangolin (EPI_ISL_410540 /COV /PANGOLIN /GUANGXI/ P5L/2017) and other bat SARS-like coronavirus strains (MG772933.1 /SARS-LIKE-COV /BAT /BAT-SL-COVZC45 /2017 and MG772934.1 /SARS-LIKE-COV /BAT /BAT-SL-COVZXC21/2015). MERS-CoV strains (KJ713299.1 /MERS-COV /CAMEL /SAU/ KSA-CAMEL-376/2013 and KU308549.1 /MERS-COV /HUMAN /KOR /SEOUL-SNU1-035 /2015) were distantly related to the Indian SARS-CoV-2 strains (68.1% identity; Multimedia Appendix 2).

NSP2, NSP3, NSP4, NSP6, NSP7, and NSP8 Genes

The dendrograms of all these 6 genes showed a similar pattern. All the 95 Indian strains clustered in 2 subclusters (A=39 and B=56 strains) within the *Betacoronavirus* lineage of SARS-CoV-2. Principally, subcluster A strains were from the first 3 months, whereas B contained strains from the next 3 months of 2020. Strains of subcluster A and B had 99.9%-100% DNA homology among themselves. All the clade-specific strains (A1, A1a, A2, A2a, A3, A5, B1, B2, B4-1, and B4-2) along with the prototype SARS-CoV-2 strain clade O (MN908947.3/SARS-CoV-2 /HUMAN /CHN /Wuhan-Hu-1 /2019) clustered close to subcluster A (99.9% identity), except NSP7 and NSP8 where the prototype clade O strain was present within subcluster B strains. SARS-CoV-2 strains isolated from carnivorous mammals like mink and tiger also grouped close to the subcluster A strains in all the dendrograms (99.9% identity). Subcluster A and B strains revealed 95.4%-98.1% nucleotide sequence similarity with bat coronavirus EPI_ISL_402131 /COV /BAT /YUNNAN /RATG13 /2013, while the pangolin-derived strain EPI_ISL_410540/COV/PANGOLIN/GUANGXI/P5L/2017 showed less identity (83%-87.5%). MERS-CoV strains NC_019843.3/MERS-COV/HUMAN/NLD/HCOV-EMC/2012 and KU740200.1 /MERS-COV /CAMEL /EGYPT /NRCE-NC163/2014 exhibited a significant phylogenetic distance (only 49.6%-60.8% homology) from the Indian isolates (Multimedia Appendicies 4-9).

L- and S-Type of SARS-CoV-2

SNPs at positions 8782 (NSP4 gene) and 28,144 (open reading frame [ORF]8) showed complete linkage among the Indian isolates under study. At these two sites, 93 strains showed a “CT” haplotype (designated as “L” type as T28,144 falls in the codon position which encodes amino acid leucine in the 84th position of ORF8 protein), while only 2 strains (1 from Kerala, EPI_ISL_413523, and 1 from Delhi, EPI_ISL_435111) revealed a “TC” haplotype (called as “S” type as C28,144 falls in the codon encoding serine at the 84th position of the ORF8 protein).

Analyses of Synonymous and Nonsynonymous Mutations

The Common Mutations in SARS-CoV-2 Indian Isolates

To explore the mutations among the 95 SARS-CoV-2 strains, we performed in-depth sequence analyses both at the genome level and at the corresponding amino acid level in different proteins, especially S glycoprotein, N protein, NSP2, NSP3, NSP4, NSP6, NSP7, NSP8, and RdRP/NSP12 with reference to the prototype SARS-CoV-2 strain (MN908947.3/SARS-CoV-2/HUMAN/CHN/Wuhan-Hu-1/2019). Out of 95 samples, 2 (2.1%) were found to have no significant “L”-type mutations (EPI_ISL_435111 and EPI_ISL_413523). Out of 93 “L”-type samples, 6 (6.3%; EPI_ISL_481156, EPI_ISL_476840, EPI_ISL_476023, EPI_ISL_458080, EPI_ISL_431101, and EPI_ISL_413522) harbored none of the mutations and were wild-type–like. Mutational analysis of the remaining 87 strains revealed circulation of two predominant “groups,” namely, the “major group” and the “minor group,” across India. The “major group,” which, of 95 isolates, was comprised of 66 (69.4%), revealed 4 coexisting SNPs: 241C>T in the five prime untranslated region (5′ UTR), 3037C>T (F106F) in the NSP3 gene, 14408C>T (P323L) in the NSP12 gene, and 23403A>G (D614G) in the S gene (Table 1). This “major group” of SARS-CoV-2 was predominantly found to circulate in regions like Delhi, Maharashtra, West Bengal, Odisha, Telangana, and Gujarat. The other 21 (22.1%) samples, which represent the “minor group,” harbored 5 coexisting mutations: 23929C>T (Y789Y) in the S gene, 28311C>T (P13L) in the N gene, 6312C>A (T1198K) in the NSP3 gene, 11083G>T (L37F) in the NSP6 gene, and 13730C>T (A97V) in the NSP12/RdRP gene (Table 2). Needless to say, the 5 coexisting mutations of the “minor group” and the 4 coexisting mutations of the “major group” did not overlap among the same SARS-CoV-2 strains. The “minor group” of SARS-CoV-2 predominated across Tamil Nadu (South) and Uttar Pradesh (Central/North).

Table 1. Single nucleotide polymorphisms associated with the major group SARS-CoV-2 strains (n=66) across India from January to June 2020.[a]

State and accession number	Spike glycoprotein (21,563-25,384 nts/1273 amino acids)				RdRP[b] protein (13,442-16,236 nts/932 amino acids)	NSP[c]3 protein (2720-8554 nts/1945 amino acids)			5'-UTR[d] (1-265 nts)/non-coding	N[e] protein (28,274-29,533 nts/419 amino acids)				NSP2 protein (806-2719 nts/638 amino acids)
	Q271R[f]	D614G[g]	G1124V[h]	D294D[i]	P323L[j]	F106F[k]	A994D[l]	K1249K[m]	241C>T	S194L[n]	RG203KR[o]	R41R[p]	T393I[q]	T85I[r]
Delhi (n=13)														
EPI_ISL_435061, EPI_ISL_435062, EPI_ISL_482665 (n=3)		✓			✓	✓			✓					
EPI_ISL_482498 (n=1)		✓			✓	✓			✓	✓				
EPI_ISL_435065-EPI_ISL_435069 (n=5)		✓		✓	✓	✓			✓	✓				
EPI_ISL_435070, EPI_ISL_435071 (n=2)		✓			✓	✓			✓		✓			
EPI_ISL_435063, EPI_ISL_435064 (n=2)		✓			✓	✓			✓					✓
Tamil Nadu (n=4)														
EPI_ISL_458032, EPI_ISL_458033, EPI_ISL_458044, EPI_ISL_458040 (n=4)		✓			✓	✓			✓					
Maharashtra (n=13)														

State and accession number	Spike glycoprotein (21,563-25,384 nts/1273 amino acids)				RdRP[b] protein (13,442-16,236 nts/932 amino acids)	NSP[c]3 protein (2720-8554 nts/1945 amino acids)			5’-UTR[d] (1-265 nts)/non-coding	N[e] protein (28,274-29,533 nts/419 amino acids)				NSP2 protein (806-2719 nts/638 amino acids)
	Q271R[f]	D614G[g]	G1124V[h]	D294D[i]	P323L[j]	F106F[k]	A994D[l]	K1249K[m]	241C>T	S194L[n]	RG203KR[o]	R41R[p]	T393I[q]	T85I[r]
EPI_ISL_479493, EPI_ISL_479510, EPI_ISL_479553, EPI_ISL_479533, EPI_ISL_479538, EPI_ISL_479497, EPI_ISL_479550, EPI_ISL_479554, EPI_ISL_479557, EPI_ISL_479560, EPI_ISL_479562, EPI_ISL_479571, EPI_ISL_479564 (N=13)		✓			✓	✓	✓		✓					
West Bengal (n=5)														
EPI_ISL_430466, EPI_ISL_430467 (n=2)		✓			✓	✓			✓					
EPI_ISL_430465 (n=1)		✓			✓	✓			✓		✓			
EPI_ISL_430468, EPI_ISL_430464 (n=2)		✓	✓		✓	✓			✓		✓			
Gujarat (n=11)														
EPI_ISL_458107, EPI_ISL_483878, EPI_ISL_476869, EPI_ISL_469036, MT576031 (n=5)		✓			✓	✓			✓					

State and accession number	Spike glycoprotein (21,563-25,384 nts/1273 amino acids)				RdRP[b] protein (13,442-16,236 nts/932 amino acids)	NSP[c]3 protein (2720-8554 nts/1945 amino acids)			5'-UTR[d] (1-265 nts)/non-coding	N[e] protein (28,274-29,533 nts/419 amino acids)				NSP2 protein (806-2719 nts/638 amino acids)
	Q271R[f]	D614G[g]	G1124V[h]	D294D[i]	P323L[j]	F106F[k]	A994D[l]	K1249K[m]	241C>T	S194L[n]	RG203KR[o]	R41R[p]	T393I[q]	T85I[r]
EPI_ISL_461484, EPI_ISL_476864 (n=2)		✓		✓	✓	✓			✓	✓				
EPI_ISL_471637 (n=1)		✓			✓	✓			✓		✓			
EPI_ISL_475058 (n=1)		✓			✓	✓	✓		✓		✓			
EPI_ISL_426414, EPI_ISL_426415 (n=2)	✓	✓			✓	✓		✓	✓			✓	✓	
Odisha (n=7)														
EPI_ISL_481154, EPI_ISL_481157 (n=2)		✓			✓	✓			✓					
EPI_ISL_481115, EPI_ISL_463078, EPI_ISL_481177, EPI_ISL_481180, EPI_ISL_481186 (n=5)		✓			✓	✓			✓		✓			
Madhya Pradesh (n=2)														
EPI_ISL_476884, EPI_ISL_476842 (n=2)		✓			✓	✓			✓					
Telangana (n=5)														
EPI_ISL_458080, EPI_ISL_431101 (n=2)		✓			✓	✓			✓					
EPI_ISL_431117 (n=1)		✓			✓	✓			✓	✓				
EPI_ISL_471588 (n=1)		✓		✓	✓	✓			✓	✓				

State and accession number	Spike glycoprotein (21,563-25,384 nts/1273 amino acids)				RdRP[b] protein (13,442-16,236 nts/932 amino acids)	NSP[c]3 protein (2720-8554 nts/1945 amino acids)			5’-UTR[d] (1-265 nts)/non-coding	N[e] protein (28,274-29,533 nts/419 amino acids)				NSP2 protein (806-2719 nts/638 amino acids)
	Q271R[f]	D614G[g]	G1124V[h]	D294D[i]	P323L[j]	F106F[k]	A994D[l]	K1249K[m]	241C>T	S194L[n]	RG203KR[o]	R41R[p]	T393I[q]	T85I[r]
EPI_ISL_ 471629 (n=1)		✓			✓	✓			✓		✓			
Karnataka (n=5)														
EPI_ISL_ 477207, EPI_ISL_ 477250 (n=2)		✓			✓	✓			✓					
EPI_ISL_ 477255, EPI_ISL_ 477237, 477239 (n=3)		✓		✓	✓	✓			✓	✓				
Uttar Pradesh (n=1)														
EPI_ISL_ 435060 (n=1)		✓			✓	✓			✓					

[a] Mutations were analyzed with compared to Wuhan-Hu-1 (MN908947.3).

[b] RdRP: RNA-dependent RNA polymerase.

[c] NSP: nonstructural protein.

[d] 5′ UTR: five prime untranslated region.

[e] N: nucleocapsid.

[f] 22374A>G.

[g] 23403A>G.

[h] 24933G>T.

[i] 22444C>T.

[j] 14408C>T.

[k] 3037C>T.

[l] 5700C>A.

[m] 6466A>G.

[n] 28854C>T.

[o] 28881-28883 GGG>AAC.

[p] 28396G>A.

[q] 29451C>T.

[r] 1059C>T

Table 2. Single nucleotide polymorphisms associated with the minor group SARS-CoV-2 strains (n=21) across India during January to June 2020.[a]

State and accession number	Spike glycoprotein (21,563-25,384 nts/1273 amino acids)	RdRP[b] protein (13,442-16,236 nts/932 amino acids)	N[c] protein (28,274-29,533 nts/419 amino acids)	NSP[d]3 protein (2720-8554 nts/1945 amino acids)		NSP6 protein (10,973-11,842 nts/290 amino acids)
	Y789Y (23929C>T)	A97V (13730C>T)	P13L (28311C>T)	S1197R (6310C>A)	T1198K (6312C>A)	L37F (11083G>T)
Tamil Nadu (n=8)						
EPI_ISL_435093-EPI_ISL_435096, EPI_ISL_435084, EPI_ISL_435087 (n=6)	✓	✓	✓		✓	✓
EPI_ISL_435091, EPI_ISL_435092 (n=2)	✓	✓	✓	✓	✓	✓
Maharashtra (n=1)						
EPI_ISL_435077 (n=1)	✓	✓	✓		✓	✓
Odisha (n=2)						
EPI_ISL_463017 (n=1)	✓	✓	✓		✓	✓
EPI_ISL_463010 (n=1)	✓	✓	✓	✓	✓	✓
Madhya Pradesh (n=1)						
EPI_ISL_476848 (n=1)	✓	✓	✓		✓	✓
Telangana (n=1)						
EPI_ISL_431103 (n=1)	✓	✓	✓		✓	✓
Karnataka (n=4)						
EPI_ISL_486399, EPI_ISL_486394, EPI_ISL_486408, EPI_ISL_MT396248 (n=4)	✓	✓	✓		✓	✓
Uttar Pradesh (n=3)						
EPI_ISL_435100, EPI_ISL_435099, EPI_ISL_435082 (n=3)	✓	✓	✓		✓	✓
Bihar (n=1)						
EPI_ISL_435112 (n=1)	✓	✓	✓		✓	✓

[a] Mutations were analyzed with compared to Wuhan-Hu-1 (MN908947.3).

[b] RdRP: RNA-dependent RNA polymerase.

[c] N: nucleocapsid.

[d] NSP: nonstructural protein.

The Unique Mutations in SARS-CoV-2 Indian Isolates

In addition to 23403A>G (D614G), 3 uncommon mutations, 23374A>G (Q271R), 24933G>T (G1124V), and 22444C>T (D294D), were also observed in the S gene of the “major group” (Table 1). Out of the 67 isolates of the major group, 28 revealed 4 novel mutations: 28854C>T (S194L; n=13), 28881-28883GGG>AAC (R203K and G204R; n=13), and coevolving mutation 29451C>T (T393I) and 28395G>A (R41R; n=2) in the N gene (Table 1). Intriguingly, 28854C>T (S194L) in the N gene was found to coevolve with the 22444C>T (D294D) mutation in the S gene of 11 samples in the major group (Table 1). We also observed 1059T>A (T85I) change within the NSP2 gene (n=2) and 6466A>G (K1249K) change in the NSP3 gene (n=2). With the 3 samples of the minor group, 6310C>A (S1197R) was found to be associated. No mutations were found within the NSP7 and NSP8 genes.

Effect of Missense Mutation A97V on the Secondary Structure of NSP12/RdRP

RdRP is the crucial enzyme for both viral RNA replication and maintenance of genomic fidelity. Thus, any significant change in RdRP structure could affect its functions, leading to an increase in the rate of mutagenesis in the genome. We have identified 2 missense mutations in the RdRP protein: P323L associated with the “major group” isolates and A97V associated with the “minor group” isolates. The effect of P323L on the secondary structure of RdRP has already been described [8].

Therefore, we analyzed the effect of novel mutation A97V on the secondary structure of RdRP by using the CFSSP server. The A97V mutation resulted in substitution of α-helixes at positions 94, 95, and 96 within the β-sheets in the RdRP secondary structure, which may alter its tertiary conformation and affect functionality (Multimedia Appendix 10).

Discussion

Principal Findings

The molecular and genetic characterization of SARS-CoV-2 pandemic strains worldwide has been studied by several scientific groups based on whole-genome sequencing [9,10]. Through this comprehensive analysis, we aimed to closely investigate the ancestry, evolutionary dynamics, accumulation of rapid mutations, and cross-genetic translation among the emerging SARS-CoV-2 strains across India. Rapid accumulation of several point mutations across the genome of SARS-CoV-2 since its origin is a prime driving force behind the evolution of different monophyletic clades. As depicted through the different phylogenetic dendrograms in our study, a monophyletic clade of all SARS-CoV-2 strains was seen with the prototype strain (Wuhan IME-WH01/2019). Clustering of all the Indian isolates with other SARS-CoV-2 strains reported worldwide (99.8%-100% nucleotide sequence identity) suggests the introduction of this virus in India was from several countries. The clustering pattern of the prototype strain from Wuhan in the phylogenetic dendrogram underscores the fact that China might have served as the origin of this zoonotic virus, which was eventually transmitted worldwide [11-13].

The origin of SARS-CoV-2 is still undetermined, but identification of its intermediate host is much needed to prevent further dissemination and interspecies transmission in the near future. Hence, we initiated this study as one of the first in India to decipher the gene-wise phylogenetics of SARS-CoV-2 strains circulating in this endemic setting. The results depicted genome-type clusters of the 95 Indian isolates, for the structural genes S and N, and the nonstructural gene RdRP/NSP12. Clustering of the study isolates with different clade-specific strains for different genes established the development of genome-type clustering. Though variations in DNA homology exists with respect to each gene, a recent bifurcation of these SARS-CoV-2 strains from the bat- and Malayan pangolin–derived SARS-like coronaviruses is supposed to have occurred, with a subsequent zoonotic transmission to humans, as depicted through all 9 dendrograms. Moreover, the SARS-CoV-2 strains were distant to MERS-CoV and other human coronaviruses. This conclusion goes at par with other phylogenetic studies establishing bat and pangolins as the proximal origin of SARS-CoV-2 [14-16].

Our study highlighted the low sequence similarity of the S gene of the Indian study isolates with some bat-derived strains like bat-SL-CoVZC45 and bat-SL-CoVZXC21, while maximum homology was noticed with bat SARS-like coronavirus (SARSr-CoV/RaTG13). This observation was consistent with a report where the S gene of SARS-CoV-2 strains circulating within China revealed the lowest sequence homology (nearly 70%) with bat strains (like SL-CoVZC45 and SL-CoVZXC21), in comparison to 96.2% identity to bat SARS-related coronavirus (SARSr-CoV/RaTG13). The RNA-binding domain within the S1 subunit of the S gene of all Indian SARS-CoV-2 and pangolin-derived strains were found to be evolutionarily conserved and phylogenetically much closer than bat RaTG13, underscoring the familiar mode of pathogenesis between the two. The Indian SARS-CoV-2 isolates too possess a polybasic cleavage site (RRAR; amino acid position 682-685) at the junction of S1 and S2 subunits of the S protein as reported by Andersen et al [14]. SARS-CoV-2 strains have been categorized into two major groups or types characterized by two SNPs at positions 8782 (NSP4 gene) and 28,144 (ORF8) that reveal complete linkage [17]. Among our Indian study isolates, frequency of the L-type (CT haplotype) was much higher (93/95, 97.9%) to the S-type (TC haplotype; 2/95, 2.1%), indicating the predominance of L-type over S-type in this geographical region.

Convoluted mutational analysis also revealed cocirculation of 2 groups of mutated SARS-CoV-2 strains in India. The “major group” of SARS-CoV-2 strains (66/95, 69.4%) represents the A2a clade reported previously from Africa, South America, Oceania, and South and West Asia, comprising of strains with coevolving mutations like 241 C>T (5′ UTR), 3037 C>T (F106F, NSP3), 14403 C>T (P323L, RdRP/NSP12), and 23403 A>G (D614G, S glycoprotein) [18-20]. Certain strains in the “major group” displayed 22374A>G (Q271R), 24933G>T (G1124V), and 22444C>T (D294D) changes in the S gene, which were unique to India. Missense mutations, Q271R and G1124V in the S protein, were found to reside around the N-linked glycosylation sites 282 and 1134, respectively, and these might affect the protein function [21]. It was not surprising to observe the triple site mutation 28881-28883 GGG>AAC (R203K and G204R) in the N gene of 13 SARS-CoV-2 strains of the “major group.” This has previously been reported from Mexico, South America, Australia, New Zealand, and a few Asian countries [22]. The 203/204 region is part of the SR dipeptide domain of the N protein (*S*RNS*S*RNSTPGSS*R*GTSPARMA) and changes in arginine at position 203 to lysine; and glycine at position 204 to arginine resulted in the insertion of a lysine residue between serine and arginine (*S*RNS*S*RNSTPGSS*KR*TSPARMA), which might interfere with the phosphorylation at serine residue required for normal functioning of the N protein [23]. This mutation demands particular attention as reduced pathogenicity has been observed previously in SARS-related coronavirus on deletion of the SR domain [24]. Mutations observed in the NSP3 gene at positions 6310 C>A (S1197R), 7392 C>T (P1558L), and 6466 A>G (K1249K) were completely unique to Indian strains. Few infrequent mutations at position 1059 T>A (T85I) in NSP2 and 8782 C>T (S76S) in NSP4 observed here have also been reported to be prevalent in other countries [19,22,25].

The “minor group” of Indian SARS-CoV-2 (21/95, 22.1%) was comprised of strains with 5 coevolving mutations: 13730C>T (A97V, RdRP/NSP12), 23929C>T (Y789Y, S), 28311C>T (P13L, N), 6312C>A (T1198K, NSP3), and 11083G>T (L37F, NSP6). All the “minor group” mutations were novel among the Indian isolates, except 11083G>T (L37F, NSP6), which was previously reported as an infrequent mutation from Australia, Japan, Netherlands, and some other European countries [18,26]. The L37F mutation strongly implies positive selection toward evolution of *Betacoronaviruses*, indicating a possible origin of the "minor group" out of this positive selection, with subsequent acquisition of mutations among the strains already harboring the 11083G>T change [25,26]. The interaction of NSP6 with NSP3 and NSP4 has been described to be essential for the formation of double membrane vesicles [25,26]. Hence, it is interesting to note the presence of a coexisting mutation 6312 C>A (T1198K) in NSP3 of the "minor group" strains, though the significance of this coexistence (L37F and T1198K) in context to the NSP6-NSP3 interaction can only be confirmed through association studies. The functional accuracy of RdRP is challenged due to the presence of the 13730 C>T (A97V) change, which was predicted to have significant effect on the secondary structure of RdRP [8]. In addition, A97V was found to be located in the nidovirus RdRP-associated nucleotidyl transferase domain whose function remains unknown [27]. The P13L mutation is located in the intrinsically disordered region of the N protein and might affect RND-binding activity of the N-terminal domain and C-terminal domain of the N protein [28,29].

Any significant mutation in the RdRP/NSP12 protein might alter replication machinery, thereby compromising the fidelity of viral RNA replication and subsequent accumulation of plausible novel mutations. The missense mutation 14408C>T (P323L) in RdRP was first observed in Italy (Lombardy) in February 2020. Few strains from Europe and North America since February 2020 have shown the emergence of mutations like 3037C>T (F106F, NSP3), 23403A>G (D614G), and 28881-28883GGG>AAC (R203K and G204R, N) in the SARS-CoV-2 genome harboring the 14408C>T (P323L) mutation within the RdRP gene, suggesting a probable association or coexistence of 14408C>T (P323L) and the emerging higher number of novel point mutations compared to viral genomes from Asia [30]. Therefore, we can assume that two mutations, 14408C>T (P323L) and 13730C>T (A97V), which were found to have significant influence on the secondary structure of RdRP, could play key roles in the simultaneous establishment of "two groups" of SARS-CoV-2 with several characteristic "co-evolving mutations" in India (Asia). However, this needs to be validated experimentally. A recent study reported that the frequency of mutations within the SARS-CoV-2 genome varies in different geographical areas, as SARS-CoV-2 gene sequences from Europe and North America present an overwhelming mutation frequency compared to that of Asia. Their study identified few recurrent mutations among isolates from Europe that were not detected among the viruses circulating within Asian countries, such as 3037C>T (F106F/NSP3 gene), 14408C>T (P323L/RdRP gene), 28881-28883GGG>AAC (R203K and G204R/N gene), and 23403A>G (D614G/S gene) [30]. In contrast, our analyses revealed that all these mutations accumulated over time beyond Europe and were profoundly seen among the "major group" of SARS-CoV-2 strains circulating across India (Asia).

The free availability of genome sequences in the publicly available servers like National Center for Biotechnology Information and GISAID has revolutionized the genome studies, resulting in continuous monitoring of mutations, recombination events, development of molecular diagnostics, identification of vaccine strains, etc. The ongoing deadly pandemic requires recording the complete patient metadata along with full genome sequences of the SARS-CoV-2 strains for better understanding of the epidemiology and virulence of this virus. Exploiting newer technologies that could help in recording additional information such as specific disease traits (comorbidity, respiratory scores, essential blood parameters), treatment, requirement of hospitalization or outpatient treatment, treatment outcome, life-threatening complication, or mortality in addition to the full viral genome sequences. This would also help in geographical region-based decisions regarding treatment modalities as well as inclusion of highly virulent subtypes of strains in vaccine formulations.

Conclusion

India harbors a greater risk of community transmission of COVID-19 due to high population density, a large population below the poverty line, and overburdened health care facilities. Hence, stringent surveillance and monitoring of the viral epidemiology and genetic diversity of a novel virus can pave way for better health care strategies and vaccine designing. This study provides comprehensive analysis of the ancestry, evolutionary dynamics, clade-specific genetic variations, as well as development of unique coevolving mutations among SARS-CoV-2 strains circulating across different regions in India. Owing to the lack of patient metadata, the impact of novel mutations on the clinical outcome or the difference in virulence of the two distinct groups of circulating strains in India could not be determined.

Acknowledgments

This study was supported by an intramural grant from the Indian Council of Medical Research, New Delhi, India. The authors acknowledge the hard work and dedication of scientists and laboratory staff in COVID-19 testing and next generation genome sequencing labs.

RS and ML were supported by fellowships from the University Grants Commission and Council of Scientific and Industrial Research, India, respectively.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Molecular phylogenetic analysis by the maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of S gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. The scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution having invariant sites (GTR+G+I). S: spike.

[PNG File , 4745 KB - bioinform_v1i1e20735_app1.png]

Multimedia Appendix 2

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of N gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution having invariant sites (GTR+G+I). N: nucleocapsid.

[PNG File , 4999 KB - bioinform_v1i1e20735_app2.png]

Multimedia Appendix 3

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of RdRP/NSP12 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution (GTR+G). NSP12: nonstructural protein 12; RdRP: RNA-dependent RNA polymerase.

[PNG File , 4475 KB - bioinform_v1i1e20735_app3.png]

Multimedia Appendix 4

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of the NSP2 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution (GTR+G). NSP2: nonstructural protein 2.

[PNG File , 4835 KB - bioinform_v1i1e20735_app4.png]

Multimedia Appendix 5

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of the NSP3 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the General Time Reversible model (GTR). NSP3: nonstructural protein 3.

[PNG File , 4291 KB - bioinform_v1i1e20735_app5.png]

Multimedia Appendix 6

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of NSP4 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model having invariant sites (GTR+I). NSP4: nonstructural protein 4.

[PNG File , 3841 KB - bioinform_v1i1e20735_app6.png]

Multimedia Appendix 7

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of the NSP6 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale ‐ bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the Tamura-3 model having invariant sites (T92+I). NSP6: nonstructural protein 6.

[PNG File , 4466 KB - bioinform_v1i1e20735_app7.png]

Multimedia Appendix 8

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of NSP7 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the Tamura-3 model with gamma distribution (T92+G). NSP7: nonstructural protein 8.

[PNG File , 4812 KB - bioinform_v1i1e20735_app8.png]

Multimedia Appendix 9

Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of NSP8 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution having invariant sites (GTR+G+I). NSP8: nonstructural protein 8.

[PNG File , 4996 KB - bioinform_v1i1e20735_app9.png]

Multimedia Appendix 10

Effect of A97V mutation on the secondary structure of RdRP/NSP12 protein. (A) Secondary structure of RdRP around 97th A (Alanine) residue of Wuhan isolate of SARS-CoV-2. (B) Secondary structure of RdRP around 97th V (Valine) residue of Indian isolate of SARS-CoV-2. NSP12: nonstructural protein 12; RdRP: RNA-dependent RNA polymerase.

[PNG File , 2119 KB - bioinform_v1i1e20735_app10.png]

References

1. Callaway E. Coronavirus vaccines: five key questions as trials begin. Nature 2020 Mar;579(7800):481. [doi: 10.1038/d41586-020-00798-8] [Medline: 32203367]
2. Zhou P, Yang X, Wang X, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020 Mar;579(7798):270-273 [FREE Full text] [doi: 10.1038/s41586-020-2012-7] [Medline: 32015507]
3. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, et al. Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res 2020;151(2 & 3):200-209 [FREE Full text] [doi: 10.4103/ijmr.IJMR_663_20] [Medline: 32242873]
4. #IndiaFightsCorona COVID-19. Government of India. 2020 May 23. URL: https://www.mygov.in/covid-19/?cbps=1
5. Mohanty SK, Sahoo U, Mishra US, Dubey M. Age pattern of premature mortality under varying scenarios of COVID-19 infection in India. medRxiv 2020 Jun 12. [doi: 10.1101/2020.06.11.20128587]
6. Global Initiative on Sharing all Influenza Data. GISAID EpiFlu database. GISAID 2020 Feb 03. [doi: 10.17616/R3Q59F]
7. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974 Jan 15;13(2):222-245. [doi: 10.1021/bi00699a002] [Medline: 4358940]
8. Chand GB, Banerjee A, Azad GK. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. PeerJ 2020;8:e9492. [doi: 10.7717/peerj.9492] [Medline: 32685291]
9. Adhikari SP, Meng S, Wu Y, Mao Y, Ye R, Wang Q, et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Infect Dis Poverty 2020 Mar 17;9(1):29 [FREE Full text] [doi: 10.1186/s40249-020-00646-x] [Medline: 32183901]
10. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020 Feb 22;395(10224):565-574 [FREE Full text] [doi: 10.1016/S0140-6736(20)30251-8] [Medline: 32007145]
11. Fisher D, Wilder-Smith A. The global community needs to swiftly ramp up the response to contain COVID-19. Lancet 2020 Apr 04;395(10230):1109-1110 [FREE Full text] [doi: 10.1016/S0140-6736(20)30679-6] [Medline: 32199470]
12. Chen G, Wu D, Guo W, Cao Y, Huang D, Wang H, et al. Clinical and immunological features of severe and moderate coronavirus disease 2019. J Clin Invest 2020 May 01;130(5):2620-2629. [doi: 10.1172/JCI137244] [Medline: 32217835]
13. Fan J, Liu X, Pan W, Douglas MW, Bao S. Epidemiology of coronavirus disease in Gansu Province, China, 2020. Emerg Infect Dis 2020 Jun;26(6):1257-1265. [doi: 10.3201/eid2606.200251] [Medline: 32168465]
14. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020 Apr;26(4):450-452 [FREE Full text] [doi: 10.1038/s41591-020-0820-9] [Medline: 32284615]
15. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 2020 May 01;368(6490):489-493 [FREE Full text] [doi: 10.1126/science.abb3221] [Medline: 32179701]

16. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. J Adv Res 2020 Jul;24:91-98 [FREE Full text] [doi: 10.1016/j.jare.2020.03.005] [Medline: 32257431]
17. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020 Mar 02;7(6):1012-1023. [doi: 10.1093/nsr/nwaa036]
18. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. Preprints 2020 Apr 30.
19. Wang JT, Lin YY, Chang SY, Yeh SH, Hu BH, Chen PJ, et al. The role of phylogenetic analysis in clarifying the infection source of a COVID-19 patient. J Infect 2020 Jul;81(1):147-178 [FREE Full text] [doi: 10.1016/j.jinf.2020.03.031] [Medline: 32277969]
20. Guan Q, Sadykov M, Nugmanova R, Carr MJ, Arold ST, Pain A. The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme. bioRxiv 2020 Apr 23. [doi: 10.1101/2020.04.21.054221]
21. Walls AC, Park Y, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020 Apr 16;181(2):281-292.e6 [FREE Full text] [doi: 10.1016/j.cell.2020.02.058] [Medline: 32155444]
22. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, et al. Large-scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geodistribution and a rich genetic variations of hotspots mutations. bioRxiv 2020 May 21. [doi: 10.1101/2020.05.03.074567]
23. Ibn Ayub M. Reporting two SARS-CoV-2 strains based on a unique trinucleotide-bloc mutation and their potential pathogenic difference. Preprints 2020 Apr 19. [doi: 10.20944/preprints202004.0337.v1]
24. Tylor S, Andonov A, Cutts T, Cao J, Grudesky E, Van Domselaar G, et al. The SR-rich motif in SARS-CoV nucleocapsid protein is important for virus replication. Can J Microbiol 2009 Mar;55(3):254-260. [doi: 10.1139/w08-139] [Medline: 19370068]
25. Phelan J, Deelder W, Ward D, Campino S, Hibberd ML, Clark TG. Controlling the SARS-CoV-2 outbreak, insights from large scale whole genome sequences generated across the world. bioRxiv 2020 May 26. [doi: 10.1101/2020.04.28.066977]
26. Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, Cauda R, et al. Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. J Infect 2020 Jul;81(1):e24-e27 [FREE Full text] [doi: 10.1016/j.jinf.2020.03.058] [Medline: 32283146]
27. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 2020 May 15;368(6492):779-782 [FREE Full text] [doi: 10.1126/science.abb7498] [Medline: 32277040]
28. Chang C, Hsu Y, Chang Y, Chao F, Wu M, Huang Y, et al. Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. J Virol 2009 Mar;83(5):2255-2264 [FREE Full text] [doi: 10.1128/JVI.02001-08] [Medline: 19052082]
29. Chang C, Hou M, Chang C, Hsiao C, Huang T. The SARS coronavirus nucleocapsid protein--forms and functions. Antiviral Res 2014 Mar;103:39-50 [FREE Full text] [doi: 10.1016/j.antiviral.2013.12.009] [Medline: 24418573]
30. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020 Apr 22;18(1):179 [FREE Full text] [doi: 10.1186/s12967-020-02344-6] [Medline: 32321524]

Abbreviations

CFSSP: Chou and Fasman Secondary Structure Prediction
GISAID: Global Initiative on Sharing All Influenza Data
MEGAX: Molecular Evolutionary Genetics Analysis, version X
MERS-CoV: Middle East respiratory syndrome–related coronavirus
N: nucleocapsid
NSP: nonstructural protein
ORF: open reading frame
RdRP: RNA-dependent RNA polymerase
S: spike
SARS: severe acute respiratory syndrome
SNP: single nucleotide polymorphism
5′ UTR: five prime untranslated region

Edited by G Eysenbach; submitted 27.05.20; peer-reviewed by A Antoniades, M Pradhan; comments to author 24.07.20; revised version received 25.07.20; accepted 24.08.20; published 07.09.20.

Please cite as:
Banerjee A, Sarkar R, Mitra S, Lo M, Dutta S, Chawla-Sarkar M
The Novel Coronavirus Enigma: Phylogeny and Analyses of Coevolving Mutations Among the SARS-CoV-2 Viruses Circulating in India
JMIR Bioinformatics Biotechnol 2020;1(1):e20735
URL: http://bioinform.jmir.org/2020/1/e20735/
doi:10.2196/20735
PMID:33496683

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org