Accepted Manuscript

Identification of chikungunya virus nsP2 protease inhibitors using structure-base approaches

Phuong T. V. Nguyen, Haibo Yu*, Paul A. Keller*

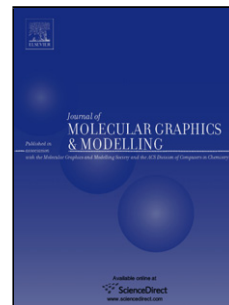School of Chemistry, University of Wollongong, 2522 Australia

Corresponding author email: hyu@uow.edu.au (HY) and keller@uow.edu.au (PAK)

Abstract

The nsP2 protease of Chikungunya virus (CHIKV) is one of the essential components of viral replication and it plays a crucial role in the cleavage of polyprotein precursors for the viral replication process. Therefore, it is gaining attention as a potential drug design target against CHIKV. Based on the recently determined crystal structure of the nsP2 protease of CHIKV, this study identified potential inhibitors of the virus using structure-based approaches with a combination of molecular docking, virtual screening and molecular dynamics (MD) simulations. The top hit compounds from database searching, using the NCI Diversity Set II, with targeting at five potential binding sites of the nsP2 protease, were identified by blind dockings and focused dockings. These complexes were then subjected to MD simulations to investigate the stability and flexibility of the complexes and to gain a more detailed insight into the interactions between the compounds and the enzyme. The hydrogen bonds and hydrophobic contacts were characterized for the complexes. Through structural alignment, the catalytic residues Cys1013 and His1083 were identified in the N-terminal region of the nsP2 protease. The absolute binding free energies were estimated by the linear interaction energy approach and compared with the binding affinities predicted with docking. The results provide valuable information for the development of inhibitors for CHIKV.

Keywords: nsP2 protease, chikungunya virus, structure-based approaches, molecular docking, virtual screening, molecular dynamics simulations.

1

1. Introduction

Chikungunya disease is caused by the chikungunya virus (CHIKV) which is one of the neglected tropical diseases that has emerged [1, 2] or reemerged [3-5] in recent years. This virus is transmitted from human to human by the bite of *Aedes aegpti* and *Aedes albopictus* mosquitoes [6]. The virus was first recorded in Tanganyiaka, Africa in 1952 [7] with the name chikungunya derived from the local dialect, meaning "that which bends up" [7] literally describing the stooped posture of patients who have to suffer joint pains for weeks or even years [7, 8]. The virus has been largely neglected due to its sporadic reemergence [3, 9], but it has recently gained attention with epidemics occurring since 2006 in nearly 40 different countries [10] from Africa to Asia, Europe, US and Australia spread by infected travelers. Even though the chikungunya disease is non-fatal, its characteristic symptoms include high fever, rash, headache, arthralgia, myalgia and polyarthralgia which affects patients, and results in significant health problems worldwide [9-13]. Unfortunately, there are no effective drugs or specific treatments available [10].

CHIKV belongs to the *Alphavirus* genus in the *Togaviridae* family [2]. Like other alphavirus, CHIKV attaches to the host through receptor-mediated endocytosis depending on endosome acidification [14]. The complete genome of CHIKV was determined in 2002 and has a length of 11805 nucleotides [15]. It consists of two open reading frames (ORFs) [12, 15] with the first ORF encoding four non-structural proteins (nsP), namely nsP1, nsP2, nsP3 and nsP4 and the second ORF encoding structural proteins including the capsid (C), envelope glycoproteins E1 and E2 and two small cleavage products (E3 and 6K). The envelope glycoproteins are involved with viral entry while the non-structural proteins are essential components and play a crucial role in viral replication and transcription [10, 16]. Therefore, they are considered as potential targets which several studies utilized to identify potential inhibitors against CHIKV [10, 17-22].

Among the non-structural proteins, the nsP2 protein is an attractive target. This protein is a multifunctional enzyme [10, 16]. In its free form, the protein induces cytotoxicity and is responsible for transcriptional shut-off [2]. The complete nucleotide sequence of the CHIKV genome revealed that the nsP2 protein is the largest non-structural protein, being 798 amino acids long and possessing a large net positive charge [15]. In the N-terminal domain, there are functions of nucleoside triphosphatase, helicase, and RNA-dependent 5′-triphosphatase while "activities of cysteine protease", also known as thiol protease, are present in the C-terminal domain (the nsP2 protease) [2, 16]. The proteolytic activity of the nsP2 protein plays an important role in the cleavage of non-structural polyproteins, critical for viral replication [16]. The mechanism is related to the

2

deprotonation of a thiol group of cysteine at the active site by assistance of an adjacent histidine residue [10]. This was confirmed by mutagenesis studies in the nsP2 protease of the Sindbis virus (a virus in the same family of CHIKV). The replacement of either cysteine or histidine at its active site can completely abolish the proteolytic function [23].

Although the function and role of the nsP2 protease is well-known, few studies have focused on exploring inhibitors for CHIKV targeting this enzyme. Several studies have constructed homology models [19, 24, 25] of the nsP2 protease based on the counterparts of the Venezuelan equine encephalitis virus (VEEV) to assist in drug design for CHIKV, before a crystal structure of the nsP2 protease became available in late 2011. With the homology models, inhibitors of the nsP2 enzyme have been screened, allowing further compound designs based on pharmacophore models of the inhibitors [19, 24, 25]. The active site was predicted in the major surface groove of the C-terminal-domain region, and likely binds to the substrate polyprotein sequences in the cleavage process [10]. The key residues responsible for interactions between the nsP2 protease and ligands were established and pharmacophore features of inhibitors were suggested [19]. The structure-activity relationships of hit compounds were evaluated with predictions of CHIKV replication inhibition at low μM concentrations [25]. Based on the recently reported crystal structure, this current study aimed to discover potential new inhibitors of CHIKV nsP2 protease through structure-based approaches by combining molecular docking and molecular dynamics simulations. *In silico* virtual screening based on docking was performed to explore inhibitors and potential binding sites of the nsP2 protease and predict their binding modes. Furthermore, to understand not only the rigid structures but also the dynamic behaviour, molecular dynamics simulations were subsequently carried out to provide details of the behaviour of the nsP2 protease and its complexes and their flexibilities upon binding of small molecules. The binding free energy for docked structures binding to the enzyme was estimated based on the linear interaction energy method.

2. Material and methods

Structure-based virtual screening

Preparation of protein target and ligands: The 3D crystal structure of the nsP2 protease of CHIKV was obtained from the Protein Data Bank (PDB id: 3TRK) and served as a protein target in docking. The chemical libraries of the National Cancer Institute (NCI) Diversity Set II, which contains diverse drug-like compounds, were selected for screening. Before docking, the protein was minimized with the steepest-descent algorithm for 3,000 steps using the Accelerys Discovery

3

Studio 2.0 software package with CHARMM force field to relax the structure and remove the steric overlap [26]. The minimized structure of protein was then prepared for docking in AutoDock Tools (version 1.5.4).

Virtual screening with AutoDock Vina: Virtual screening utilised docking with AutoDock Vina (version 1.1.2) [27]. The protein was kept rigid while the ligands were fully flexible. The Lamarckian Genetic Algorithm was used to search space for docking and the binding affinity of a complex was calculated on the basis of a set of weighted energies in empirical-scoring function. The key docking parameters were defined including the location of the docking site (centre x, y, z) and the size of a grid box. Initially, potential binding sites were detected, based on the blind docking in which the box was sufficiently large to cover the whole protein 60 Å × 70 Å × 60 Å centred at the centre of protein, or 66 Å × 86 Å × 70 Å centred at centre of a previous reported binding site, The next step was a focused docking with a smaller box (20 Å × 20 Å × 20 Å), centred on the potential binding site of interest. Additionally, the MetaPocket was used to identify potential pocket sites of the nsP2 protease [28]. In total, five potential binding pockets were identified including the active site. The top ten compounds were selected according to their binding affinities with the default scoring function in Vina. The docking results were analysed in Accelerys Discovery Studio 2.0. The drug-like properties of these compounds were also evaluated using the Lipinski Guidelines [29].

Molecular dynamics simulations

The program NAMD [30] (version 2.9) was used to run molecular dynamics simulations of the docked complexes to investigate the stability and flexibility of the nsP2 protease and its complexes and to study the detailed interactions between ligands and the nsP2 protease. The protein atoms were represented by the CHARMM22 force field [31] and the corresponding parameters for the ligands were generated with AmberTools (version 13) [32]. MD simulations were conducted for the apo-state (PDB id: 3TRK) and eight complexes of the nsP2 protease and hit compounds were obtained from virtual screening. The systems were solvated in a cubic box of TIP3P water molecules and neutralized by counterions to achieve the physiological ionic concentration of 0.15 M with NaCl. All simulations were run under the periodic boundary conditions with the initial water box size of 89.7 Å × 89.7 Å × 89.7 Å. The temperature was set at 298.15 K and the pressure at 1 atm. The Langevin algorithm was used to maintain the temperature and pressure coupling. The Particle Mesh Ewald (PME) algorithm was used to compute long-range electrostatic interactions [33]. The cutoff distance for van der Waals interactions was set at 12 Å and the pair-list distance

4

was 13.5 Å. The minimization process was applied first and followed by equilibrium simulations with weak harmonic restraints on the heavy atoms for 3 ns. The production runs were continued for 50 ns. The trajectories for analysis were saved every 10 ps. The resulting trajectories were analyzed by the CHARMM [34] and VMD (version 1.9.1) including hydrogen bond and hydrophobic contact interactions [35].

Additional simulations of ligands in water were conducted to calculate the binding free energies with the linear interaction energy (LIE, described in the following section). The other parameters for MD simulations were the same with simulations for protein as aforementioned. The average values of van der Waals and electrostatic contribution were calculated for the equilibrated part of the simulations.

Binding free energy calculation from MD simulations

There have been different approaches developed over the years, such as free energy perturbation, thermodynamic integration and linear interaction energy (LIE), to calculate the ligand binding free energies [36, 37]. In this study, the LIE approach was applied to calculate the absolute binding free energies for ligands when complexed with the nsP2 protease. The main reason for this approach was that this method effects compromise between speed and accuracy and so it is practically more feasible for a large number of ligands. Derived from the linear response theory, the binding free energies were approximated as

$$\Delta G_{bind} = \alpha \times <V_{bound}^{vdw} - V_{unbound}^{vdw}> + \beta \times <V_{bound}^{elec} - V_{unbound}^{elec}> + \gamma \qquad (1)$$

The van der Waals and electrostatic components were obtained from two simulations: one with the ligand in an aqueous solution and the other with the nsP2 protease-ligand complex in an aqueous solution. The NAMDEnergy plugin in VMD was used to compute the energy components over the frames obtained from the MD simulations [30, 35]. In Eq. (1), α, β and γ are empirical parameters and can be fitted to the target data. The value α is often set to 0.18 for a wide variety of protein-ligand systems [37]. The value β represents a function of the chemical nature of the ligand, so in principle it can be parameterized from explicit solvent free energy calculations of different chemical entities [37]. The value of γ depends on the hydrophobic nature of the binding site [37].

5

3. Results and discussion

Identification of the active site of the nsP2 protease

It is known that the nsP2 protease carries out its function through a conserved catalytic dyad with a cysteine and a histidine, however they have not been conclusively identified in the CHIKV nsP2 protease. For instance, Bassetto et al. reported that Cys1013 and His1083 (the numbering is according to the CHIKV nsP2 sequence) in the homology model of the nsP2 protease were the catalytic residues compared to Cys477 and His546 in the VEEV nsP2 protease through a sequence alignment between the CHIKV nsP2 protease and the VEEV nsP2 protease. In contrast, similarly based the sequence alignment, Singh et al. predicted that the active site residues interacting with the peptide substrate include Lys1045, Gly1176, His1222 and Lys1239 without explicitly identifying the catalytic residues [19]. In this work, structural alignment of the nsP2 proteases from three alphaviruses, CHIKV (PDB id: 3TRK), Venezuelan equine encephalitis virus (VEEV, PDB id: 2HWK) and Sindbis virus (SINV, PDB id: 4GUA) were carried out with MUSTANG [38] (Figure S1). The sequence identity between the CHIKV nsP2 protease sequence and the counterparts in VEEV and SINV sequences are 40% and 44%, respectively. This structural alignment revealed that Cys1013 and His1083 are the two catalytic residues in the nsP2 protease of CHIKV (Figure 1). Their catalytic roles are supported by their close proximity in the crystal structure. Cys1233 and His1222 in the C-terminal domain, identified in the homology model by Singh et al., are separated by 15 Å in the solved crystal structure (between the S atom in Cys and the N atom in His); thus, they are not likely to be the catalytic residues unless there is a structural transition upon substrate binding. Conversely, Cys1013 and His1083 in the N-terminal domain are sufficiently close enough to carry out proton transfer (with the distance of 4.8 Å between the S atom in Cys1013 and the N atom in His1083); thus, they are properly positioned to be a catalytic dyad. In addition, the presence of Trp1084 close to His1083 was proposed to be necessary for a functional protease [39].
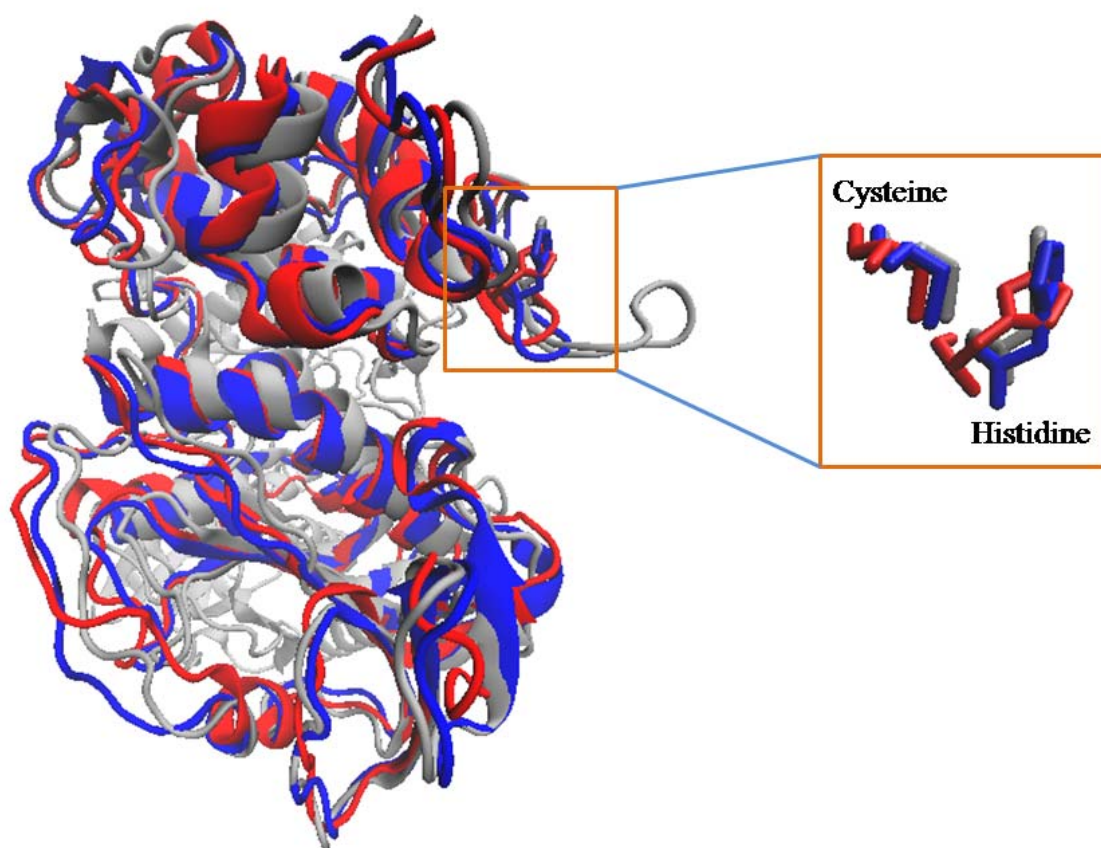
6

Figure 1. Superimposition of three crystal structures, namely the CHIKV nsP2 protease (PDB id: 3TRK, in blue), the VEEV nsP2 protease (PDB id: 2HWK, in red) and the structure of SINV (PDB id: 4GUA, in gray). The conserved catalytic residues, cysteine and histidine (in licorice, Cys1013 and His1083 in the CHIKV nsP2 protease (in blue), Cys477 and His546 in the VEEV nsP2 protease (in red) and Cys1021 and His1098 in the SINV structure (in gray)) are also shown.

Blind dockings and focused dockings to identify potential binding sites and hit compounds for the nsP2 protease

In order to identify potential inhibitors targeting the nsP2 protease, blind dockings were initially carried out to discover potential binding sites of the nsP2 protease and potent inhibitors. The compounds from library NCI Diversity Set II were docked to the entire protein (the nsP2 protease) with a grid box placed at either the centre at the binding site 1 (VST1) or at the centre of the protein (VST2). The binding site 1 (Pocket 1), including the key residues for interactions (Lys1045, Gly1176, His1222 and Lys1239), was taken from the published data by Singh et al [19]. The binding modes of the compounds were ranked by their predicted binding affinities. The top ten compounds (hit compounds) for each virtual screening (VST), are listed in Table 1 (their chemical structures was listed in Table S1) with their binding affinities and the locations.

7

Table 1. Results of the top ten hit compounds from the blind dockings. The binding affinities ΔG are in kcal/mol.

VST1[a]			VST2[b]		
Hits	ΔG	Location	Hits	ΔG	Location
1. NCI_61610	-10.6	Pocket 1	1. NCI_293778	-10.3	Pocket 4
2. NCI_293778	-10.3	Pocket 4	2. NCI_61610	-10.1	Pocket 1
3. NCI_116702	-9.2	Pocket 3	3. NCI_670283	-9.7	Pocket 4
4. NCI_37553	-9.2	Pocket 4	4. NCI_116702	-9.2	Pocket 3
5. NCI_84100_a	-9.2	Pocket 3	5. NCI_217697	-9.2	Pocket 1
6. NCI_84100_b	-9.1	Pocket 3	6. NCI_84100_a	-9.2	Pocket 1
7. NCI_25457	-9.0	Pocket 3	7. NCI_84100_b	-9.2	Pocket 1
8. NCI_670283	-9.0	Pocket 1	8. NCI_298892_b	-9.1	Pocket 1
9. NCI_97920	-9.0	Pocket 3	9. NCI_37553	-9.1	Pocket 5
10. NCI_58052	-8.8	Pocket 3	10. NCI_25457	-9.0	Pocket 3

a) In VST1, the grid box is centred at Pocket 1 of the nsP2 protease (12.4 Å, 34.3 Å, 28.6 Å) with dimensions of 66 Å × 86 Å × 60 Å. b) In VST2, the grid box is centred at the centre of the nsP2 protease (14.3 Å, 25.5 Å, 22.3 Å) with dimensions of 60 Å × 70 Å × 60 Å. The origin and axes of the coordinate systems were set the same as in the PDB structure (PDB id: 3TRK).

Additionally, the potential binding pockets were predicted with the MetaPocket algorithm. In total, there were five potential binding pockets (labeled as Pockets 1 to 5, shown in Figure 2) identified for the nsP2 protease. Pocket 4, containing the catalytic dyad Cys1013 and His1083, is the active site. The blind docking data initially revealed four potential binding sites in the nsP2 protease (Pockets 1, 3, 4, 5) where the hit compounds could potentially bind. In the docking VST1, six out of ten ligands preferred binding at Pocket 3 along with two ligands at Pocket 1 and two ligands at Pocket 4. In VST2, most of the ligands (five) occupied Pocket 1, two ligands occupied Pocket 3, two ligands occupied Pocket 4 and one ligand occupied Pocket 5. Encouragingly, the four binding pockets identified in the blind docking were reproduced by the MetaPocket method, despite the methods being based on very different algorithms. Additionally, a further pocket (Pocket 2) near Pocket 1 was found. Pocket 1 and Pocket 4 (Figure 2: behind) were groove-like while Pocket 2 (adjacent to Pocket 1, sharing some residues with Pocket 1 such as Tyr1177 and His1222), Pocket 3 (behind) and Pocket 5 (behind) were shallow. Pocket 3 was below Pocket 1, Pocket 4 was on the opposite site behind Pocket 1 and Pocket 5 was in the rear side behind Pocket 1. The binding affinities of all hit compounds were approximately -9 kcal/mol, which indicates favourable interactions between the compounds and the nsP2 protease. It was also interesting to note that some compounds could bind into different pockets with varying conformations such as NCI_670283 (Pocket 1 and 4), NCI_84100_b (Pocket 1 and 3) and NCI_37553 (Pocket 3 and 5).

8

In an effort to obtain more hit compounds for the nsP2 protease as well as to eliminate potential sampling issues with the blind docking, a focused docking centred at each of the potential pockets identified with the blind docking and MetaPocket method was carried out. The VST3/VST4, VST5, VST6, VST7 and VST8 screenings were focused at Pocket 1, Pocket 2, Pocket 3, Pocket 4 and Pocket 5, respectively. The results of the top ten compounds and their binding affinities are listed in Table S2 and their chemical structures are presented in Table S3. The top docked structures of each virtual screening are represented in Figure 2, showing the locations of five pockets (more details are illustrated in Figure S2 with hit compounds at each of focused dockings and in Figure S3 with top hit compounds and surrounding residues within 5 Å of each pocket).
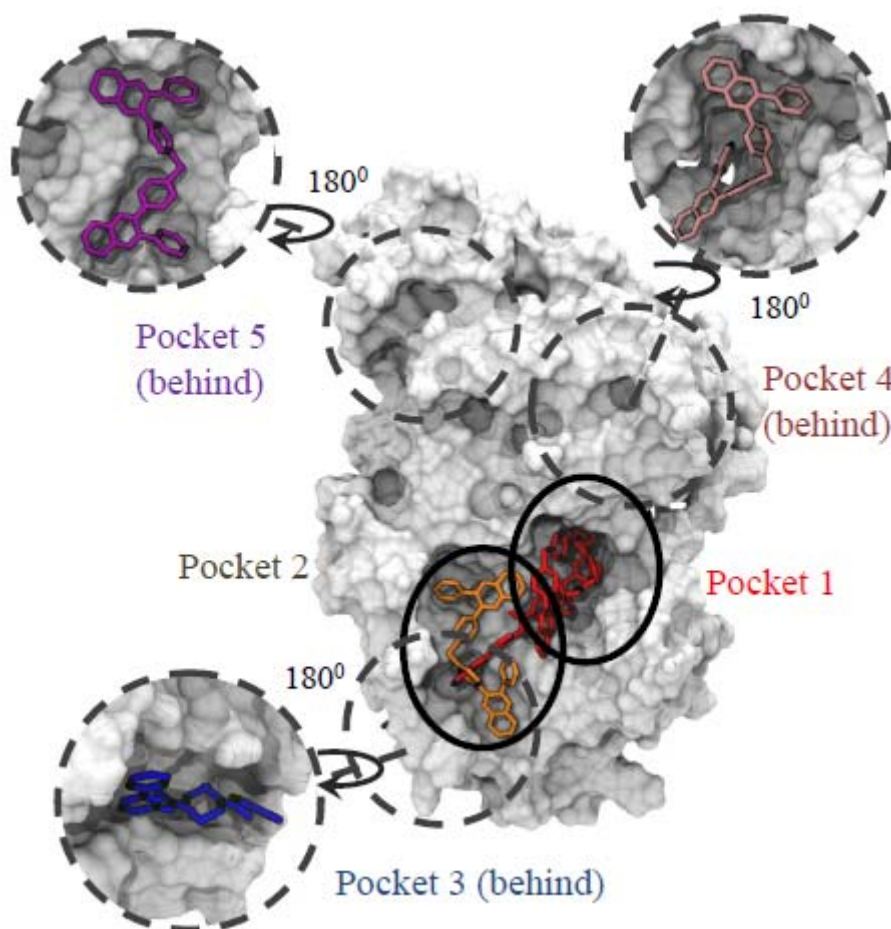

Figure 2. Representation of docked structures of top hit compounds in different virtual screenings at five different binding sites of the nsP2 protease with Pocket 4 being the active site of the nsP2 protease.

Interestingly, a change of grid box size and its location affected the searching process in the Vina program. In VST4 with an increased box size, larger compounds were identified as top hits despite not being present in VST3 for example NCI_293778 (-9.8 kcal/mol), NCI_84100_a (-9.2 kcal/mol) and NCI_84100_b (-9.2 kcal/mol). It is likely that the larger grid box accommodates the binding of

9

these molecules better. Some hit compounds were present in different virtual screenings, such as NCI_293778 (Pocket 1, 2, 3, 4 and 5), NCI_37553 (Pocket 1, 2, 3, 4 and 5) and NCI_61610 (Pocket 1, 4 and 5).

Binding modes obtained from docking at potential binding sites were analysed to show details of interactions and to identify key residues through analyses of hydrogen bonds (H-bonds) or hydrophobic contacts. The residues surrounding ligands within 5 Å which make up each pocket are presented in Table 2. As mentioned earlier, the active site was Pocket 4 with Cys1013 and His1083 in good positions to be a catalytic dyad. The residue Trp1084 is close to His1083 in Pocket 4 which is essential for interactions, as previously reported in the study of the nsP2 protease of SINV [23]. The results of the focused dockings showing the important residues involved in forming H-bonds and hydrophobic contact for each pocket, are listed in Table 2. Among the five pockets, with respect to ligands containing aromatic rings, π-stacking interactions and π-network interactions were often present with His1222 at Pocket 1, Tyr1177 at Pocket 2 and Tyr1079 at Pocket 4.
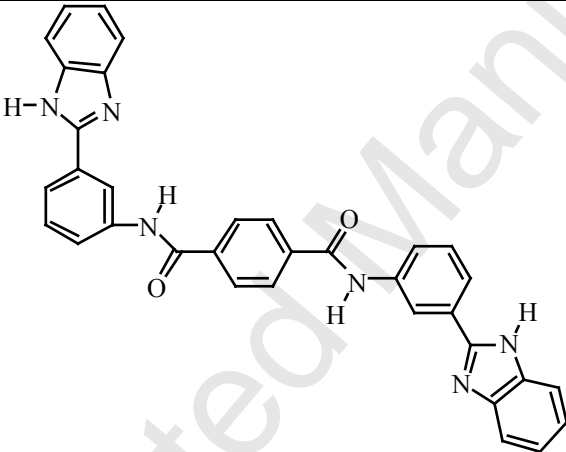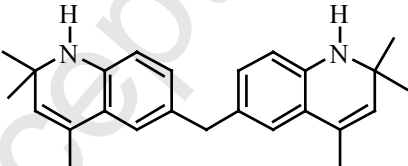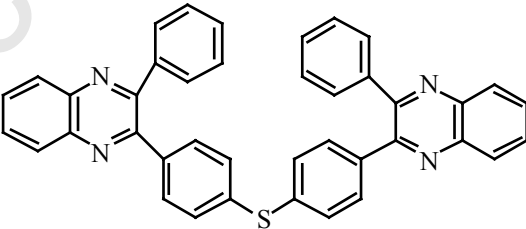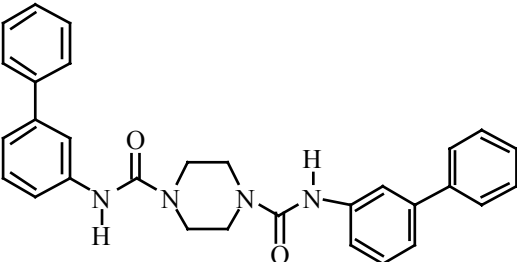
Table 2. The important residues in each pocket of the nsP2 protease with the key residues involved in forming H-bonds and hydrophobic contacts between the protein and ligands in bold.

Binding pocket	Residues making up the pocket
Pocket 1	Gln1039, Ala1040, Glu1043, **Lys1045**, Ala1046, Tyr1047, Gly1176, Tyr1177, Pro1191, Leu1192, Gly1193, Val1194, Asn1202, Leu1203, Glu1204, Ile1221, **His1222**, Thr1223, Asp1235, His1236, Met1238, **Lys1239**, Met1242 and Cys1233
Pocket 2	His1151, Pro1153, Val1154, **Lys1155**, Gly1156, Glu1157, Arg1158, Met1159, **Glu1160**, **Tyr1177**, Asn1178, Leu1179, Ala1180, **His1222**, Thr1223, Pro1224, Phe1225, **Gln1232**, Arg1260, Tyr1262, Ser1293, Thr1292 and **Thr1295**
Pocket 3	Glu1157, Arg1158, Met1159, **Trp1161**, Leu1162, Lys1165, Ile1166, **Asn1167**, Gly1254, **Ser1256**, Arg1281, Ser1282, Arg1284, Leu1286, Lys1287, Pro1288, Pro1289, **Cys1290**, Leu1300, Ser1302 and Asn1303
Pocket 4	Cys1013, Ala1046, Tyr1047, Ser1048, Pro1049, Glu1050, Val1051, Tyr1078, **Tyr1079**, Asn1082, His1083, **Trp1084**, Gly1090, **Lys1091**, Phe1093, Tyr1201, Asn1202, Glu1204, **Leu1205**, Gly1206, Pro1208, Ala1209, Met1238, Gln1241, Met1242, Gly1245, Asp1246 and Arg1249
Pocket 5	Asn1040, Glu1050, Leu1053, Asp1064, Leu1065, **Asp1066**, **Ser1067**, Gly1068, Leu1069, **Phe1070**, Ser1071, Lys1091, Phe1093, Asn1096, Glu1098, **Ala1099**, Ile1102, Leu1103, Lys1106, Tyr1107, **Asn1140**, Arg1141, **Arg1142**, Leu1143, Pro1144, Arg1267, **Glu1270**, **Arg1271**, Cys1274, Val1275, Arg1278, Thr1313, His1314 and Asn1317

Molecular dynamics simulations investigating the stability of the protein and its complexes

The top hit compounds NCI_217697, NCI_61610, NCI_37553 and NCI_293778 from virtual screenings were subjected to molecular dynamics simulations using the NAMD package (Table 3). In particular, ligand NCI_293778 with different conformations at the different binding sites (NCI_293778vst4 at Pocket 1, NCI_293778vst5 at Pocket 2, NCI_293778vst2 and NCI_293778vst7 at Pocket 4 and NCI_293778vst8 at Pocket 5) was investigated.

Table 3. Chemical structure of five top hit compounds and their properties.

	Compound	Structural formula	ΔG^a	Physical properies
1	**NCI_61610** ($C_{34}H_{24}N_6O_2$)		-10.6	LogP[b]: 5.31 H-D[c]: 4 H-A[d]: 4 MW[e]: 548.6
2	**NCI_217697** ($C_{25}H_{30}N_2$)		-9.3	LogP: 4.97 H-D: 2 H-A: 0 MW: 358.5
3	**NCI_293778** ($C_{40}H_{26}N_4S$)		-9.8[f], -10.4[g] and -10.2[h]	LogP: 10.87 H-D: 0 H-A: 4 MW: 594.73
4	**NCI_37553** ($C_{30}H_{28}N_4O_2$)		-9.6	LogP: 5.03 H-D: 2 H-A: 2 MW: 476.57

11

a) ΔG refers to the binding affinities by Vina in kcal/mol. b) The logarithm of the partition coefficient between n-octanol and water using ChemBioDraw Ultra 14.0. c) Number of hydrogen bond donor atoms. d) Number of hydrogen bond acceptor atoms. e) Molecular weight. f) Located in Pocket 1 and Pocket 2. g) Located in Pocket 4. h) Located in Pocket 5.

The overall stability of the nsP2 protease and its complexes was evaluated by the values of backbone atomic positional root-mean-square-deviation (RMSDs). Analysis of the backbone RMSD profile (Figure S4) showed that most of the systems were stable during the simulations within 1-2 Å after reaching equilibrium within 50 ns. In simulation of the complex of ligand NCI_293778vst8 at Pocket 5, a large RMSD was observed around 47 ns (Figure S4). Examination of the structures revealed that it was due to the relative movement between two domains and the fluctuation of loops. This ligand was also found to gradually move from a position whereby part of the ligand was positioned out of the pocket, to the whole ligand being out of the pocket but still stuck at the rear after 40 ns and finally dissociated from the pocket after 47 ns.

The trajectories after 3 ns were used for further analysis. For the flexibility of the nsP2 protease and its complexes, the root-mean-square-fluctuation (RMSF) curves of the Cα atoms in the nsP2 protease and its complexes were calculated (Figure 3). Comparative analysis of the RMSF values of the apo protein and the complexes was focused on the binding sites of ligands. Most of the residues making up the binding sites were quite stable during the simulation with fluctuations within 2 Å. Furthermore, no significantly reduced fluctuations were found in the binding pockets of the protein in comparison with the nsP2 protease (apo protein) and its complexes of the nsP2 protease-ligand.
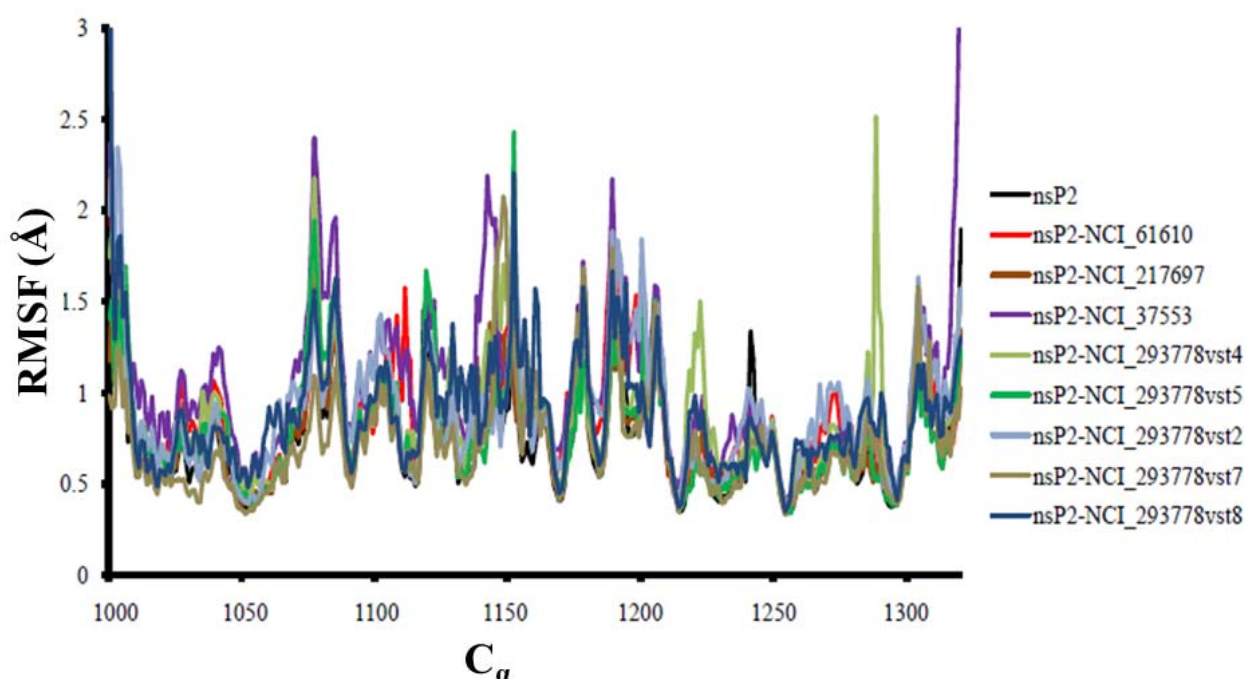
12

Figure 3. RMSFs values of C_α atoms of the apo protein nsP2 and its different complexes during MD simulations.

Atomic details of interaction between the nsP2 protein and hit compounds

More details were obtained by analysis of the molecular interactions between the ligands and nsP2 through hydrogen bonding interactions and hydrophobic contacts. The criterion for a hydrogen bond in a given structure was a maximum distance of 3.5 Å between hydrogen donor (D) and acceptor (A) atoms and the angle D-H…A must be greater than 120 degrees [17]. A cutoff distance of 4.0 Å was used in monitoring hydrophobic contacts between the non-polar parts of residues of the nsP2 protease and ligands [40, 41]. For most of the investigated hit compounds, their interactions with the nsP2 protease were from hydrophobic contacts rather than H-bonds, as most of the H-bonds occupancy was found to be low (less than 10%). The residues involved in forming hydrophobic contacts based on the trajectories of the MD simulations are listed in Table 4 and H-bonds occupancies are listed in Table S4.

13

Table 4. Hydrophobic contact analyses on the trajectories sampled in the MD simulations.

Ligand	Interacting non-polar parts of the residues
NCI_61610	Ala1040, Tyr1177, Pro1191, Leu1203, Ile1221, Leu1243
NCI_217697	Pro1191, Leu1243
NCI_37553	Trp1161, Leu1162, Leu1286, Pro1288, Pro1289, Leu1300
NCI_293778vst4	Pro1191, Leu1192, Leu1203, Ile1221, Leu1243
NCI_293778vst5	Tyr1177, Pro1224
NCI_293778vst2	Ala1046, Tyr1079, Trp1084, Leu1205
NCI_293778vst7	Ala1046, Val1077, Tyr1079, Trp1084, Leu1205
NCI_293778vst8	Ile1102, Leu1103

Considering the structure of the ligands, the presence of both donor and acceptor atoms in the ligand NCI_61610 and NCI_37553 could form more H-bonds than ligands NCI_217697 and NCI_293778. As expected, H-bonds were mostly formed with NCI_61610 and NCI_37553. Ligand NCI_61610 showed H-bond interactions with residues Lys1239, Glu1204, Leu1203, Tyr1177, Gly1176 and Lys1045, although these interactions were weak (their occupancies less than 10%). The complex nsP2 protease-ligand NCI_37553 at Pocket 3 was maintained through strong H-bonds between its oxygen and with Arg1284 (80.5%), or between its nitrogen and Arg1284 (26.3%), or this oxygen with Ser1302 (10.4%). Interestingly, a comparison of NCI_293778 bound at different binding sites revealed that the nitrogen of this ligand could form one H-bond (with an occupancy of 10.7%) with residue His1222 at Pocket 1 accompanied by surrounding hydrophobic interactions (Figure 4) but in contrast, NCI_293778 did not often have strong H-bonds at other pockets (with low occupancies, see Table S4). This indicates that NCI_293778vst4 was more likely to be accommodated in Pocket 1 than in other pockets. However, the aromatic ring of this ligand could form π-stacking interactions or a π-network with Tyr1177 (Pocket 2) and Tyr1079, Trp1084 (Pocket 4) to maintain the interactions between the protein nsP2 protease and ligand.
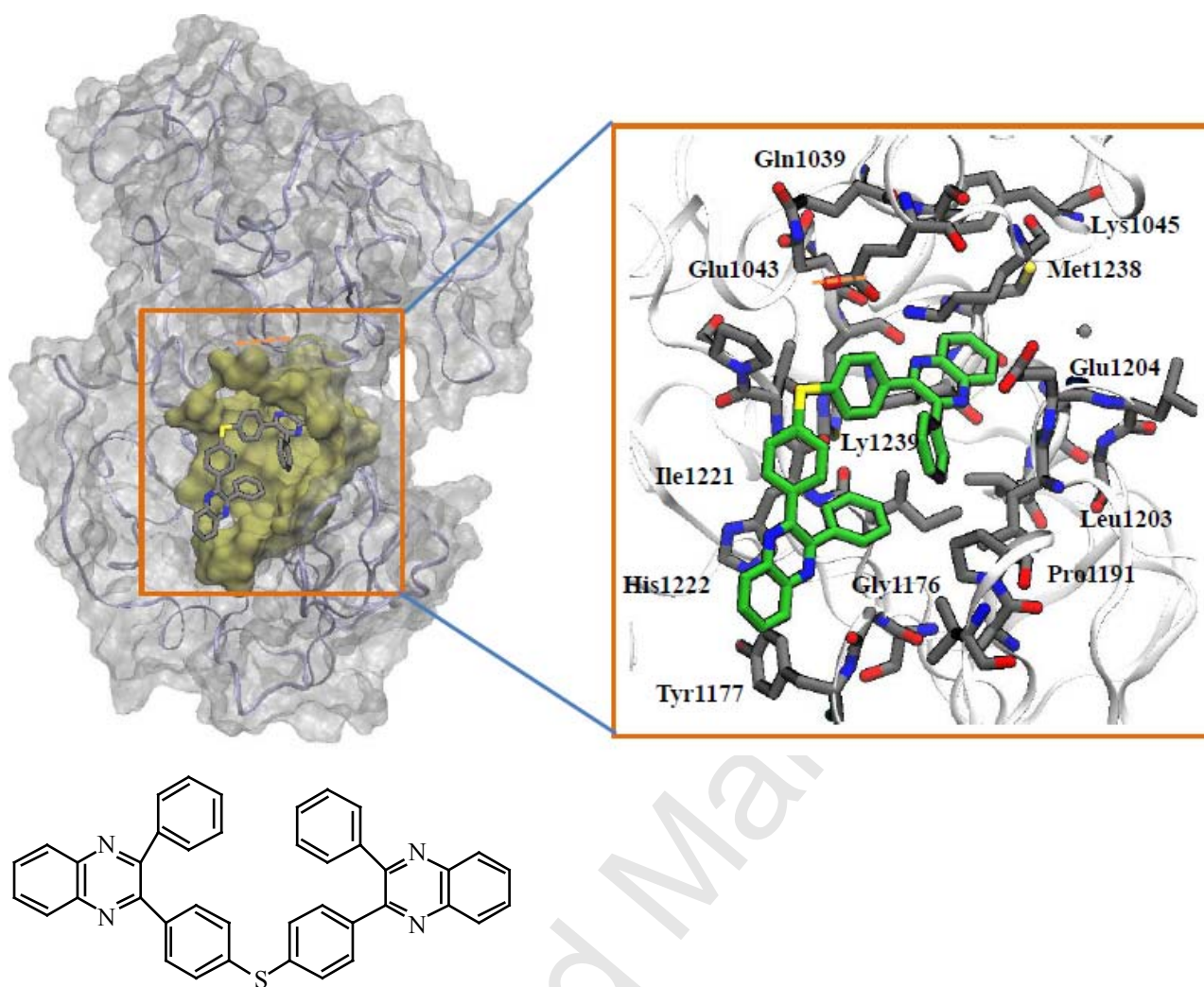
14

Figure 4. The docked structure of ligand NCI_293778vst4 (in green) in Pocket 1 showing the residues forming Pocket 1 and the compound-protein interactions (in grey).

Furthermore, the interaction between these ligands and nsP2 at the different binding sites was examined to compare with docking and simulation results. The docking results were in close agreement with molecular dynamics simulations even though in order to maintain interactions there was flexibility of residues in the protein and ligands in forming H-bonds or creating hydrophobic contacts during simulations. There was no major change in the important residues at the binding sites of the nsP2 protease. Moreover, understanding the effects of ligand binding to other binding pockets on the active site was also of interest. The comparison of the initial and final structures in MD simulations revealed no significant structural changes in the residues in the active site. Experimental validation is required to confirm whether there was an allosteric site or a synergistic effect upon binding at different sites within the active site of the nsP2 protease.

15

Binding free energy calculations using the linear interaction energy (LIE) method

The prediction of binding free energy is usually a challenge for docking and scoring in computational drug design. In this study, the converged trajectories of complexes obtained from simulations were used to calculate the binding free energy for the hit compounds. The average values of van der Waals and electrostatic interactions for each ligand in the two states (bound state and unbound state) were calculated (Table 3). The van der Waals value of the ligands in the complex with the enzyme were much more negative than when only in solution, showing that the ligand has more favorable van der Waals interactions in the complex. However, it is striking that the electrostatic interactions of ligands in all the complexes were significantly less favourable than those of the ligand in water. In applying the LIE equation, the values of empirical parameters α, β and γ require an understanding of the chemical structure and also the binding site [37]. For the value of β, all of the top hit compounds in the study are neutral compounds and this value was estimated to be $\beta = 0.43$, based the empirical function proposed previously [37]. Initially, the test value of α and γ were set 0.18 and 0.0, respectively, which did not agree with the docking results (ΔG^1 in Table 3). However, as it has been shown that α and γ need to be recalibrated depending on the different systems; the α value has been suggested to be dependent on the system and the force field used in the LIE calculations while the γ relies on the nature of the binding site. In this case, due to the lack of experimental data of complexes, the chemical nature of the ligands and the binding sites were taken into the consideration. Most binding sites are composed of both polar and non-polar residues, so the definition of hydrophobicity in the selection of γ is not easy to define and in practice, the exact value of γ does not affect the relative ranking. A larger value of $\alpha = 1.043$ was adopted as it had been shown to provide a better estimate in the study of cytochrome P450-camphor analogue complexes [42]. The results of ΔG^2 with α set to 1.043 are shown in Table 3 and revealed better agreement with the binding affinity obtained from docking. This gave good explanations when comparing the binding free energies for different ligands at the same pocket, such as in the case of NCI_61610, NCI_217697 and NCI_293778vst4. For instance, ligand NCI_293778vst4 with the presence of stronger H-bonds (indicated by the higher value of H-bond occupancy) had higher binding affinities compared to the other ligands. H-bonds and hydrophobic contacts in some ligands (such as NCI_293778vst4 and NCI_37553) were maintained. These ligands were expected to gain lower binding free energies, however LIE predictions did not show this. It could be explained that there was a compromise between van der Waals and electrostatic results from hydrogen bonds interactions and hydrophobic contacts between the ligands and the enzyme nsP2 protease that contributes to the binding free energy. More work and results from experimental data are required to clarify the issue.

16

Table 3. Average binding free energies (kcal/mol) of top hit compounds calculated by LIE method using data trajectories from the MD simulations: ΔG^1 (in kcal/mol using $\alpha = 0.18$, $\beta = 0.43$ and $\gamma = 0$) or ΔG^2 (in kcal/mol using $\alpha = 1.043$, $\beta = 0.43$ and $\gamma = 0$) [37]. ΔG is the predicted binding affinity by Vina (in kcal/mol)

Hit compound	V_{bound}^{vdw}	$V_{unbound}^{vdw}$	V_{bound}^{elec}	$V_{unbound}^{elec}$	ΔvdW	$\Delta Elec$	ΔG^1	ΔG^2	ΔG
NCI_61610	-65.0	-54.4	-79.2	-91.0	-10.6	11.8	3.2	-6.0	-10.6
NCI_217697	-50.0	-39.1	-30.4	-36.2	-10.9	5.8	0.5	-8.9	-9.3
NCI_37553	-60.1	-48.6	-62.5	-73.1	-11.5	10.6	2.5	-7.4	-9.6
NCI_293778vst4	-69.6	-54.6	-35.6	-48.0	-15.0	12.4	2.6	-10.3	-9.8
NCI_293778vst5	-68.5	-54.4	-42.3	-47.8	-14.1	5.5	-0.2	-12.3	-9.8
NCI_293778vst2	-69.2	-54.7	-38.9	-47.7	-14.5	8.8	1.1	-11.3	-10.3
NCI_293778vst7	-72.7	-55.6	-40.9	-48.2	-17.1	7.3	0.1	-14.6	-10.4
NCI_293778vst8	-65.6	-54.4	-43.2	-47.4	-11.2	4.2	-0.2	-9.9	-10.2

4. Conclusion

In this work, the X-ray crystal structure of nsP2 protease was utilized to conduct a combination of molecular docking, virtual screening and molecular dynamics simulations to search for potential inhibitors of the nsP2 protease. Starting with blind dockings and published information, we identified top hit compounds, together with the five potential binding pockets of the nsP2 protease. Subsequently, the focused dockings into these different binding sites were investigated to discover more hit compounds and look at further binding modes at these pockets. The top hit compounds were then subjected to molecular dynamics simulations for 50 ns after equilibration. The simulation results demonstrated the different binding affinities of different ligands through the number of H-bonds and hydrophobic contacts. Previous studies largely focused on Pocket 1 as the active site. However, in this study, Pocket 4 in the N-terminal domain of the nsP2 protease was identified as the active site by the presence of catalytic residues, Cys1013 and His1083. Importantly, the current work offers more opportunities to identify potential inhibitors. The effect upon the active site of ligands binding into different pockets requires further experimental work. Finally, the trajectories data of MD simulations were utilized for the linear interaction energy to obtain accurate binding free energy. The suggestion of increasing the value of α from 0.18 to 1.043 provided more reliable results. More experimental data is required to elucidate the issue in regard to achieving compromise between van der Waal and electrostatic energy. Finally, it is worth emphasizing that this is the first instance where molecular modelling with more accurate data from X-ray structure has been studied. Our findings open up a promising approach in combining docking and molecular dynamics

17

simulations to assist in rational drug design, especially providing useful information for the design of inhibitors for CHIKV that can help to combat the disease.

References

[1] Cavrini, F., Gaibani, P., Pierro, A.M., Rossini, G., Landini, M.P., Sambri, V. Chikungunya: an emerging and spreading arthropod-borne viral disease. J. Infect. Dev. Ctries. 2009, 3, 744.

[2] Gould, E.A., Delogu, I., Forrester, N., Khasnatinov, M., Gritsun, T., de Lamballerie, X., et al. Understanding the alphaviruses: Recent research on important emerging pathogens and progress towards their control. Antivir. Res. 2010, 87, 111-24.

[3] Townson, H., Nathan, M.B. Resurgence of chikunguna. Trans. R. Soc. Trop. Med. Hyg. 2008, 102, 308-9.

[4] Staples, J.E., Breiman, R.F., Powers, A.M. Chikungunya fever: an epidemiological review of a re-emerging infectious disease. Clin. Infect. Dis. 2009, 49, 942-8.

[5] Burt, F.J., Rolph, M.S., Rulli, N.E., Mahalingam, S., Heise, M.T. Chikungunya: a re-emerging virus. The Lancet. 2012, 379, 662-71.

[6] Schwartz, O., Albert, M.L. Biology and pathogenesis of chikungunya virus. Nat. Rev. Microbiol. 2010, 8, 491-500.

[7] Robinson, M.C. An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952-53. I. Clinical features. Trans. Roy. Soc. Trop. Med. Hyg. 1955, 49, 28-32.

[8] Mohan, A. Chikungunya fever: clinical manifestations & management. Indian J. Med. Res. 2006, 124, 471.

[9] Sudeep, A.B., Parashar, D. Chikungunya: an overview. J. Biosci. 2008, 33, 443.

[10] Rashad, A.A., Mahalingam, S., Keller, P.A. Chikungunya Virus: Emerging Targets and New Opportunities for Medicinal Chemistry. J. Med. Chem. 2014, 57, 1147-66.

[11] Her, Z., Kam, Y.-W., Lin, R.T.P., Ng, L.F.P. Chikungunya: a bending reality. Micro. Infect. 2009, 11, 1165-76.

[12] Singh, S.K., Unni, S.K. Chikungunya virus: host pathogen interaction. Rev. Med. Virol. 2011, 21, 78-88.

[13] Xavier de, L., Laetitia, N., Remi, C. Antiviral Treatment of Chikungunya Virus Infection. Infect. Disord. Drug Targets. 2009, 9, 101-4.

[14] Sourisseau, M., Fsihi, H., Frenkiel, M.-P., Blanchet, F., Afonso, P.V., Ceccaldi, P.-E., et al. Characterization of reemerging chikungunya virus. Plos Pathog. 2007, 3, 0804-17.

18

[15] Khan, A.H., Morita, K., Parquet Md, M.d.C., Hasebe, F., Mathenge, E.G.M., Igarashi, A. Complete nucleotide sequence of chikungunya virus and evidence for an internal polyadenylation site. J. Gen. Virol. 2002, 83, 3075-84.

[16] Pastorino, B.A.M., Peyrefitte, C.N., Almeras, L., Grandadam, M., Rolland, D., Tolou, H.J., et al. Expression and biochemical characterization of nsP2 cysteine protease of Chikungunya virus. Virus Res. 2008, 131, 293-8.

[17] Rungrotmongkol, T., Nunthaboot, N., Malaisree, M., Kaiyawet, N., Yotmanee, P., Meeprasert, A., et al. Molecular insight into the specific binding of ADP-ribose to the nsP3 macro domains of chikungunya and venezuelan equine encephalitis viruses: Molecular dynamics simulations and free energy calculations. J. Mol. Graph. Modell. 2010, 29, 347-53.

[18] Malet, H., Lafitte, D., Ferron, F., Lescar, J., Gorbalenya, A.E., de Lamballerie, X., et al. The crystal structures of Chikungunya and Venezuelan equine encephalitis virus nsP3 macro domains define a conserved adenosine binding pocket. J. Virol. 2009, 83, 6534.

[19] Singh, K.D., Kirubakaran, P., Nagarajan, S., Sakkiah, S., Muthusamy, K., Velmurgan, D., et al. Homology modeling, molecular dynamics, e-pharmacophore mapping and docking study of Chikungunya virus nsP2 protease. J. Mol. Model. 2012, 18, 39-51.

[20] Voss, J., Vaney, M.C., Duquerroy, S., Vonrhein, C., Ginard-Blanc, C., Crublet, E., et al. Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. Nature. 2010, 468, 709-12.

[21] Rashad, A.A., Keller, P.A. Structure based design towards the identification of novel binding sites and inhibitors for the chikungunya virus envelope proteins. J. Mol. Graph. Modell. 2013, 44, 241-52.

[22] Nguyen, P.T.V., Yu, H., Keller, P.A. Discovery of in silico hits targeting the nsP3 macro domain of chikungunya virus. J. Mol. Model. 2014, 20, 1-12.

[23] Strauss, E.G., De Groot, R.J., Levinson, R., Strauss, J.H. Identification of the active site residues in the nsP2 proteinase of Sindbis virus. Virol. 1992, 191, 932-40.

[24] Bora, L. Homology modeling and docking to potential novel inhibitor for Chikungunya (37997) protein nsP2 protease. Proteomics Bioinform. 2012, 5, 054-9.

[25] Bassetto, M., Silvestri, R., Tron, G.C., Neyts, J., Leyssen, P., Brancale, A., et al. Computer-aided identification, design and synthesis of a novel series of compounds with selective antiviral activity against chikungunya virus. Antivir. Res. 2013, 98, 12-8.

[26] Accelrys Software Inc. Discovery Studio Modeling Environment. Accelrys Software Inc., San Diego, 2013.

[27] Trott, O., Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 2010, 31, 455-61.

[28] Huang, B. MetaPocket: a meta approach to improve protein ligand binding site prediction. Omics. 2009, 13, 325-30.

[29] Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 2001, 46, 3-26.

[30] Phillips, J.C., Schulten, K., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., et al. Scalable molecular dynamics with NAMD. J. Comput. Chem. 2005, 26, 1781-802.

[31] MacKerell, A.D., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B. 1998, 102, 3586-616.

[32] Case, D.A., Woods, R.J., Cheatham, r.T.E., Darden, T., Gohlke, H., Luo, R., et al. The Amber biomolecular simulation programs. J. Comput. Chem. 2005, 26, 1668-88.

[33] Ulrich, E., Lalith, P., Max, L.B., Tom, D., Hsing, L., Lee, G.P. A smooth particle mesh Ewald method. J. Chem. Phys. 1995, 103, 8577-93.

[34] Brooks, B.R., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., et al. CHARMM: the biomolecular simulation program. J. Comput. Chem. 2009, 30, 1545-614.

19

[35] Humphrey, W., Dalke, A., Schulten, K. VMD: Visual molecular dynamics. J. Mol. Graph. Modell. 1996, 14, 33-8.

[36] Aqvist, J., Marelius, J. The linear interaction energy method for predicting ligand binding free energies. Comb. Chem. High T. Screen. 2001, 4, 613-26.

[37] Guitierrez-de-Teran, H., Aqvist, J. Linear Interaction Energy: Method and Applications in Drug Design. In: Comput. Drug Dis. Design, Springer New York, 2012, Vol. 819, pp. 305-23.

[38] Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M. MUSTANG: a multiple structural alignment algorithm. Proteins. 2006, 64, 559-74.

[39] Russo, A.T., White, M.A., Watowich, S.J. The Crystal Structure of the Venezuelan Equine Encephalitis Alphavirus nsP2 Protease. Structure. 2006, 14, 1449-58.

[40] Bruno, A., Guadix, A.E., Costantino, G. Molecular dynamics simulation of the heterodimeric mGluR2/5HT(2A) complex. An atomistic resolution study of a potential new target in psychiatric conditions. J. Chem. Inform. Modell. 2009, 49, 1602-16.

[41] Rajgaria, R., McAllister, S.R., Floudas, C.A. Towards accurate residue-residue hydrophobic contact prediction for alpha helical proteins via integer linear optimization. Proteins. 2009, 74, 929-47.

[42] Almlöf, M., Brandsdal, B.O., Aqvist, J. Binding affinity prediction with different force fields: examination of the linear interaction energy method. J. Comput. Chem. 2004, 25, 1242-54.

20