J Mol Model (2010) 16:49–59
DOI 10.1007/s00894-009-0519-9

ORIGINAL PAPER

# Insight into ligand selectivity in HCV NS5B polymerase: molecular dynamics simulations, free energy decomposition and docking

**Tong Li · Matheus Froeyen · Piet Herdewijn**

Received: 6 March 2009 / Accepted: 4 April 2009 / Published online: 26 May 2009


**Abstract** Modeling studies were performed on HCV NS5B polymerase in an effort to design new inhibitors. The binding models of five different scaffold inhibitors were investigated and compared by using molecular dynamics simulations, free energy calculation and decomposition. Our results show Tyr448 plays the most critical role in the binding of most inhibitors. In addition, favorable contributions of residues Pro197, Arg200, Cys366, Met414 and Tyr448 in a deep hydrophobic pocket prove to be important for the selectivity of inhibitors. Furthermore, an optimized docking protocol was presented based on cross-docking the five inhibitors in the palm binding site of this enzyme using the Autodock program. This protocol was used later to virtually screen NCI and Maybridge diversity set libraries. The binding site was profiled *via* the statistics and analysis of the hydrogen bond networks formed between the receptor and the top-ranked diversity set compounds. Based on our detailed binding site analysis two useful rules were proposed to guide the selection of promising hits.

**Keywords** Cross-docking · Free energy decomposition · HCV NS5B Polymerase · MM-GBSA · Molecular dynamics · Structure-based drug design

T. Li (✉) · M. Froeyen · P. Herdewijn
Laboratory for Medicinal Chemistry,
Rega Institute for Medical Research,
Katholieke Universiteit Leuven,
Minderbroedersstraat 10,
B-3000 Leuven, Belgium
e-mail: tong.li@rega.kuleuven.be

## Introduction

Chronic hepatitis C virus (HCV) infection is emerging as a worldwide health crisis. There are approximately 3–4 million new cases of HCV infection each year, and current estimates suggest that approximately 170 million people are chronically infected [1]. Therefore, there is a growing unmet medical need to discover novel therapies for chronic HCV infection.

HCV NS5B polymerase is an essential enzyme for viral RNA replication. This biochemical activity is not present in mammalian cells, offering the opportunity to identify very selective inhibitors of the viral enzyme [2]. The structure of the NS5B polymerase can be viewed as a right hand, where the palm domain contains the active site of the enzyme and where the fingers and the thumb are responsible for the interaction with the RNA (Fig. 1) [3, 4]. NS5B inhibitors can be divided into two categories: nucleoside inhibitors (NIs) and non-nucleoside inhibitors (NNIs). In contrast to NIs, NNIs are a structurally and chemically heterogeneous class and do not induce premature termination of the RNA synthesis [3]. Moreover, NNIs are almost invariably allosteric inhibitors believed to block the enzyme, preventing a conformational transition needed for initiation of RNA synthesis [2]. NNIs can interact at a number of different allosteric sites such as NNI Site A, B, C, D [5].

Structure-based drug design has become more important as a dramatic increase of novel three-dimensional structures of biological targets is available [6–8]. Many NS5B/NNI complex structures have been solved recently. Thus, it is a challenge to discover new inhibitors against HCV NS5B polymerase using structure-based drug design methods. In this work, we have used molecular dynamics simulations, free energy calculation and decomposition methods [9] to investigate the binding modes of five different scaffold

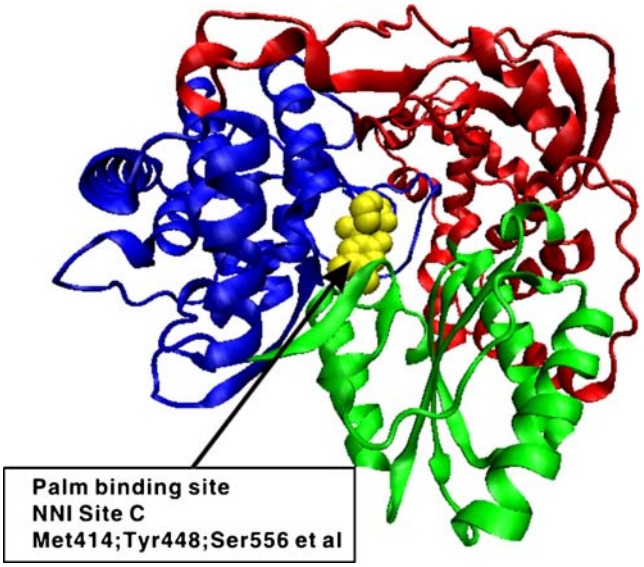


**Fig. 1** Structure of HCV NS5B polymerase. Thumb, palm and finger domains are colored blue, green and red respectively. NNI bound in the NNI Site C is colored in yellow

palm site inhibitors at the atomic level. The binding free energies were computed using the molecular mechanics/generalized Born surface area (MM/GBSA) method [10, 11]. In addition, we present an optimized protocol using the Autodock program based on cross-docking of the five inhibitors to the binding site. Furthermore, with the optimized docking protocol, virtual screening of NCI and Maybridge diversity set libraries was performed. The binding site was profiled *via* the statistics and analysis of the hydrogen bond networks formed between the receptor and the top-ranked diversity set compounds.

## Methods

### Molecular dynamics simulations

Five publicly available X-ray structures of HCV NS5B polymerase (1YVF, 2GC8, 2GIQ, 2JC0, 2QE5) complexed with different NNIs were obtained from the Protein Data Bank [12]. They were selected with three criteria: the protein is HCV NS5B polymerase genotype 1b; their NNIs all bind to the same palm binding site and these five NNIs represent five different scaffolds including acrylic acid derivative [13] (Fig. 2, NNI-1) , proline sulfonamide derivative [14] (Fig. 2, NNI-2), thiadiazine derivative [15] (Fig. 2, NNI-3), acyl pyrrolidine derivative [16] (Fig. 2, NNI-4), anthranilic acid derivative [17] (Fig. 2, NNI-5). The B chains, water molecules and other cofactors in the PDB files were removed from their X-ray structures. Missing residues were recovered with reference to other crystal structures of HCV NS5B polymerase. Proteins in the complexes were prepared using the *tleap* program in AMBER10 [18, 19]. The ligands were prepared by using the antechamber suite [20] in the AMBER package. Atomic charges were derived with the AM1-BCC charge method [21]. Two parameter sets were used, the biomolecular force field ff03 [22] for the protein and general AMBER force field (GAFF) [23] for the inhibitor. The complex was soaked in a rectangular box of TIP3P [23, 24] water molecules with a margin of 12 Å along each dimension. $Cl^-$ ions were added to neutralize the system. This yielded about 75,000 atoms for each system.

The system was minimized by 1000 steps of steepest descent followed by 1000 steps of conjugate gradient to remove the bad contacts in the crystal structure. Then, during a 300 ps equilibration period, each system was brought to 300 K in 25 K increments at 25 ps intervals. Harmonic restraints with a force constant of 5 kJ $mol^{-1}$ $Å^{-1}$ were applied to the backbone of proteins in this initial period. A subsequent 10 ns production run for each system was performed at a constant temperature of 300 K and a constant pressure of 1 atm. The particle mesh Ewald (PME) method [25] was applied to calculate long-range electrostatic interactions. The SHAKE method [26] was applied to

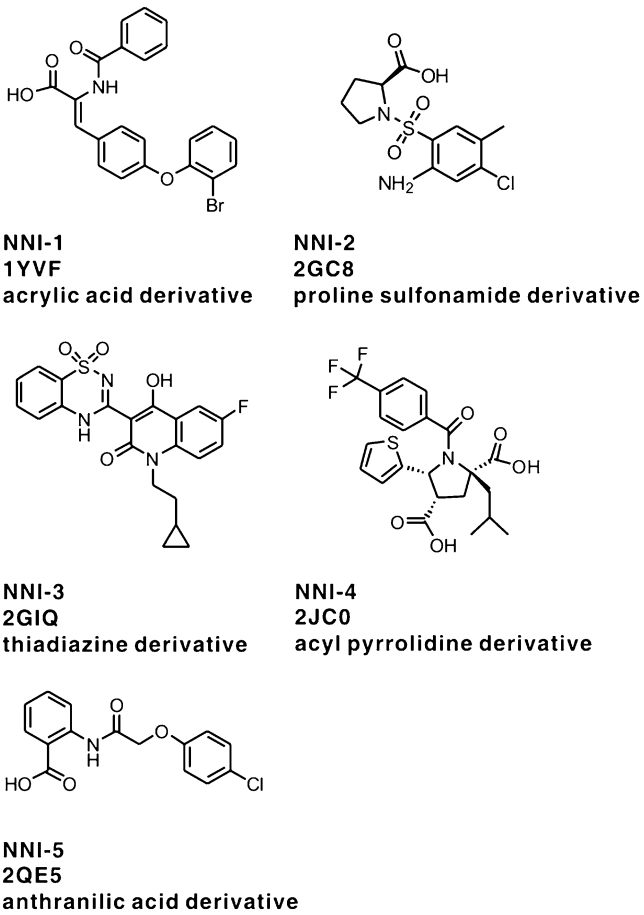


**Fig. 2** Chemical structures of the five known inhibitors

constrain all of the covalent bonds involving hydrogen atoms. Periodic boundary conditions were applied to all dimensions. No constraint was applied to either the protein or the ligand during the production simulations. Coordinates were saved every 2 ps.

## MM/GBSA calculation and free energy decomposition

Due to the high computational demand of the PB calculations, the interaction energy between each inhibitor and protein was computed using MM/GBSA methods. A total number of 500 snapshots were taken from the last 8 ns of the MD trajectory with an interval of 16 ps. The MM-GBSA method can be conceptually summarized as:

$$\begin{aligned}\Delta G_{bind} &= \Delta G_{complex} - \Delta G_{protein} - \Delta G_{ligand} \\ &= \Delta E_{MM} + \Delta G_{GB} + \Delta G_{np} - T\Delta S\end{aligned} \quad (1)$$

where $\Delta E_{MM}$ is the molecular mechanics interaction energy between the protein and the inhibitor; $\Delta G_{GB}$ and $\Delta G_{np}$ are the electrostatic and nonpolar contributions to desolvation upon inhibitor binding, respectively; and $-T\Delta S$ is the conformational entropy change, which was not considered because of the high computational cost and low prediction accuracy. The dielectric constant was set to 1 for the interior solute and 80 for the surrounding solvent. The LCPO method [27] was used to calculate the solvent accessible surface area (SASA) for the estimation of the nonpolar solvation free energy ($\Delta G_{np}$) with (=0.0072 kcal mol$^{-1}$ Å$^{-2}$ and (=0.00 kcal mol$^{-1}$ [28]. The polar contribution ($\Delta G_{GB}$) of desolvation was computed using a modified GB model developed by Onufriev et al. [29].

Evaluation of the contribution of each residue to the binding free energy has been made at the atomic level by means of free energy decomposition. This decomposition was for molecular mechanics and solvation energies but not for entropies. MM/GBSA method was used for the binding energy calculation. The details of this protocol were described in previous studies [9].

## Protein and ligand preparation for docking

The same five X-ray structures of NS5B polymerase complexed with different NNIs (Fig. 2) were used for our docking experiment. After removing B chains, water molecules and other cofactors, the complexes were superposed on the unliganded NS5B polymerase (1C2P) using the Chimera [30] program and the new coordinates were kept in order to facilitate the cross-docking studies. Missing residues were recovered with reference to other crystal structures of NS5B polymerase in the PDB database. All the hydrogens were added to the five proteins, obtained by deleting the corresponding NNIs in their complexes, using the Maestro program [31]. Protonation states were assumed to be those most common at pH 7, *i.e.*, Lysine, Arginine, Aspartates, and Glutamate were considered in the ionized form.

The 3D coordinates of the five NNIs were extracted from the corresponding PDB files and hydrogens were added using Maestro program and Gasteiger charges [32] were assigned in AutoDock Tool 1.4.6 [33]. These conformations were used for the cross-docking. An external conformational analysis of the five inhibitors was made using a random conformation as the starting point instead of crystal conformations for a further validation docking. In this case, the ligands were converted to SMILES strings by the OpenBabel program [34] and then the LigPrep module of Maestro was used to produce their 3D structures. During the preparation, the force field was set to MMFFs and all the combinations of stereoisomers were generated. One conformation was produced for NNI-1, NNI-3 and NNI-5 respectively. Meanwhile, there were two and four conformations with different stereochemistry for NNI-2 and NNI-4 respectively and only one conformation for each with the same stereochemistry as that from the crystal structure was kept for docking.

The NCI diversity set [35] and Maybridge HitFinder Collection [36] were used as libraries for a virtual screening experiment. The NCI diversity library is a reduced set of 1990 compounds selected from the original NCI-3D structural database for their unique scaffolds. The HitFinder Collection comprises 14400 premier compounds representing the drug-like diversity of the Maybridge Screening Collection. Both libraries were used to make the ligand database more diverse and representative. The two libraries were downloaded in 3D SDF format and prepared using the LigPrep module of Maestro and then converted into PDB format by the OpenBabel program. Polar hydrogens were added and Gasteiger charges were assigned to both the proteins and ligands in AutoDock Tool 1.4.6. Of the compounds from NCI diversity set, 139 could not be processed properly and were eliminated due to the presence of metal atoms or multiple fragments. An additional 32 were removed due to extremely poor binding, yielding 1819 compounds in total for further analysis. In the meanwhile, 14,094 compounds in the HitFinder Collection were kept after removing problematic and extremely bad binding compounds.

## Optimized docking protocol

Autodock 4.0.1 [37] was used for all the docking calculations. Proteins and ligands were converted into the same and special file format named “PDBQT” for docking. The grid box was centered at the center of the

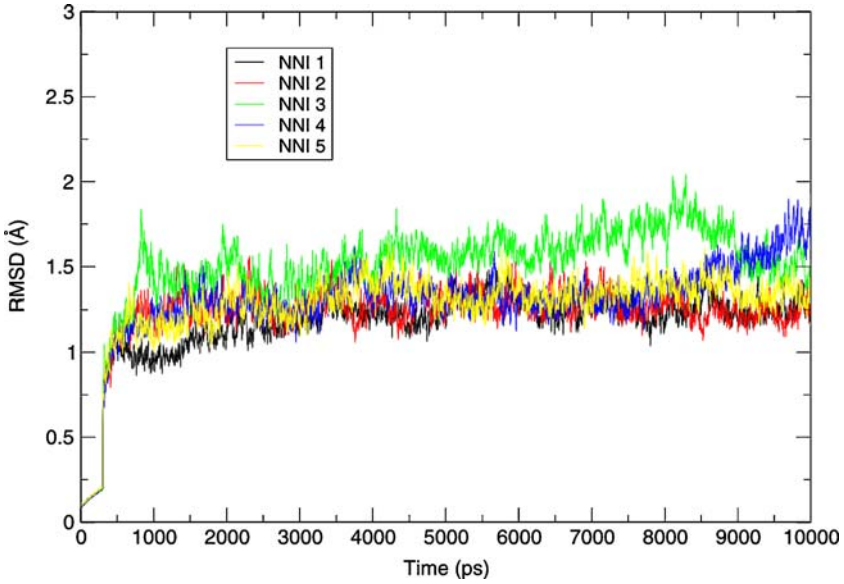


**Fig. 3** RMSDs of Cα atoms for the five simulated trajectories

Site C which was defined as all protein residues within 5 Å of the atoms of any of the five known active inhibitors (Fig. 2). In a docking study, the main issue is to get the correct ligand pose and to have it in the very first ranked position. The calculated lowest binding energy is intensively used as the standard criteria to rank the different ligand poses. However, the correct poses are not always found at the highest position which can easily induce false positive and false negative, especially during virtual screening which is heavily dependent on the first ranked ligand pose. In addition, due to the availability of several protein receptors, a representative receptor structure in which the proper conformations of known inhibitors can be recovered as many as possible was needed if there is no significant protein flexibility involved. Thus, we investigated deeper two important docking parameters of Autodock: the grid box size and the number of runs by determining whether the docked poses of the five known active inhibitors with the lowest binding energy were predicted correctly in their cognate and noncognate proteins. RMSD values were calculated between heavy atoms of the docked structures and their initial structures in the crystal structure.

Except for the grid box size and number of runs, all the other docking parameters were set to their default values. First two grid box sizes of 60×60×60 and 30×30×40 with a spacing of 0.375 Å between the grid points were used to build the affinity grids of different atom type respectively. Based on the two grid box sizes, the docking was performed with 10 docking runs. After comparing the results, the grid built from the smaller grid box size was selected to evaluate the different number of docking runs (20, 30, and 50). Finally, the best combination of a grid size of 30×30×40 and 50 docking runs was used as the docking parameters for the virtual screening of the compounds from NCI diversity set and the HitFinder Collection.

Dockings were carried out on an Intel Xeon 2.33 Ghz Linux workstation with 4 Gb RAM and 4 CPUs. We achieved an average throughout of about 1 ligand/35 min/CPU with the optimized docking protocol.

The analysis of hydrogen bonding for the docking results was performed by HBPLUS [38] and LIGPLOT [39] programs. This calculation was automated by a small in-house developed linux script.

**Table 1** Hydrogen bonds based on the last 8 ns simulations

| Hydrogen bond | | % Occupied | Distance (Å) |
|---|---|---|---|
| NS5B residues | NNI-1 | | |
| Gly449-NH | O | 66.25 | 3.131 (0.20) |
| Tyr448-NH | O | 32.89 | 2.881 (0.15) |
| NS5B residues | NNI-2 | | |
| Tyr448-NH | O2 | 67.49 | 2.968 (0.17) |
| Cys366-O | HN12 | 60.36 | 2.866 (0.14) |
| NS5B residues | NNI-3 | | |
| Ser556-OGH | O13 | 61.42 | 2.754 (0.15) |
| Tyr448-OH | O18 | 35.91 | 3.123 (0.24) |
| NS5B residues | NNI-4 | | |
| Ser367-OGH | O25 | 65.88 | 2.674 (0.16) |
| Tyr448-NH | O13 | 63.07 | 3.164 (0.18) |
| NS5B residues | NNI-5 | | |
| Tyr415-OH | O3 | 99.62 | 2.787 (0.15) |
| Arg386-NH11 | O3 | 48.13 | 2.973 (0.19) |

**Table 2** Binding free energy calculation between the five NNIs and HCV NS5B polymerase (All energies are in kcal $mol^{-1}$)

| NNIs | $\Delta E_{ele}$ | $\Delta E_{vdW}$ | $\Delta G_{nonpolar}$ | $\Delta G_{GB}$ | $\Delta G_{bind}$ | $\Delta G_{bind,exp}$ |
|---|---|---|---|---|---|---|
| NNI-1 | -272.01±13.23 | -48.87±3.01 | -5.84±0.22 | 287.77±12.28 | -38.95±3.49 | -9.53 |
| NNI-2 | -254.27±15.92 | -33.67±2.40 | -4.28±0.17 | 261.54±16.02 | -30.67±4.14 | -7.50 |
| NNI-3 | -128.83±8.01 | -39.86±2.74 | -5.07±0.22 | 137.28±6.52 | -36.48±2.65 | -8.97 |
| NNI-4 | -415.61±22.51 | -43.27±3.79 | -5.97±0.39 | 434.47±21.37 | -30.38±5.34 | -6.40 |
| NNI-5 | -253.94±21.94 | -30.25±2.23 | -4.98±0.22 | 265.82±20.73 | -23.34±3.31 | -7.88 |

## Results and discussion

### Molecular dynamics

Binding free energy calculations using the MM-PBSA/MM-GBSA methods, which are based on molecular dynamics of the given protein-ligand complex in explicit solvent, have been widely used to predict binding affinities of biomolecular complexes. Previously reported studies suggest longer MD trajectories may be required for the reliability of MM-PBSA/MM-GBSA energy estimates [40]. Thus, we performed 10 ns of unrestrained dynamics for each protein/NNIs complex to sample adequate conformational space for energy calculation. The $C_\alpha$ RMSDs of all five simulated systems compared to the starting minimized structures were monitored and plotted in Fig. 3. It shows that all the RMSD curves increase quickly in the first 500 ps and tend to be stable after 2 ns. Thus, we selected the last 8 ns snapshots of each system to be used in the binding free energy calculation and free energy decomposition analysis.

As it is well known, hydrogen bond plays a critical role in the ligand/protein interaction. Thereby, we examined the hydrogen bonds between the NNIs and binding site residues of the protein. The hydrogen bonds that occupy more than 30% time in the last 8 ns period are listed in Table 1. We can see from the table that for each complex there is at least one stable hydrogen bond with occupancy more than 60% which are responsible for the binding stability. Moreover, the strong hydrogen bond between NNI-5 and Tyr415 occupies nearly 100% time during the 8 ns simulation. Interestingly, the hydrogen bonds between residue Tyr448 and the NNIs occur most frequently, with 4 out of 5 NNIs forming this hydrogen bond with an occupancy more than 30%, which indicates the key role of Tyr448 for the NNIs binding. In addition, Cys366, Ser367, Arg386, Tyr415, Gln446, Gly449 and Ser556 have been targeted at least one time by the NNIs by means of hydrogen bonding.

### Binding free energy calculation and free energy decomposition

The calculated relative binding free energy and contributions of vdW, electrostatic interaction and solvation energy using the single trajectory MM-GBSA method are listed in Table 2. Because entropic contribution is excluded, the calculated binding free energy is lower than that derived from the experiment. Detailed analysis suggests that major contributions favorable to binding are vdW and electrostatic

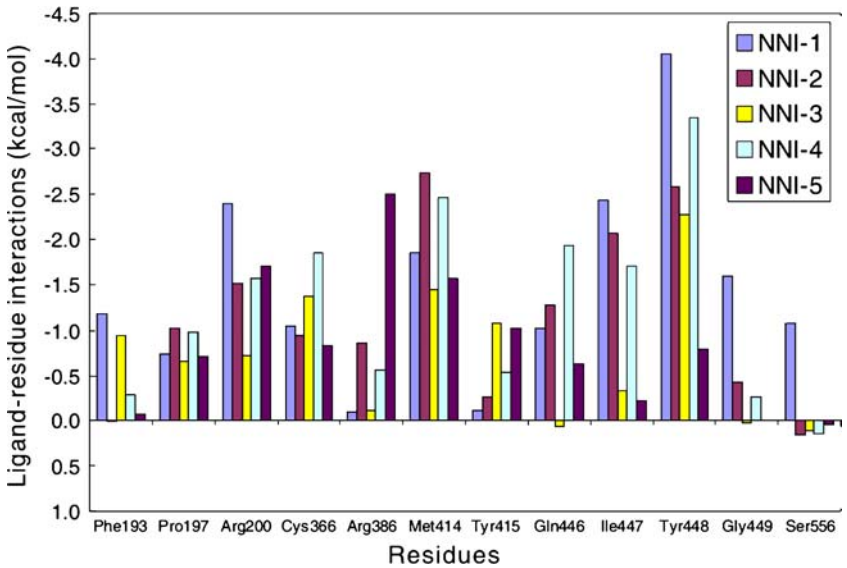


**Fig. 4** Binding free energy contributions of key binding site residues calculated from free energy decomposition

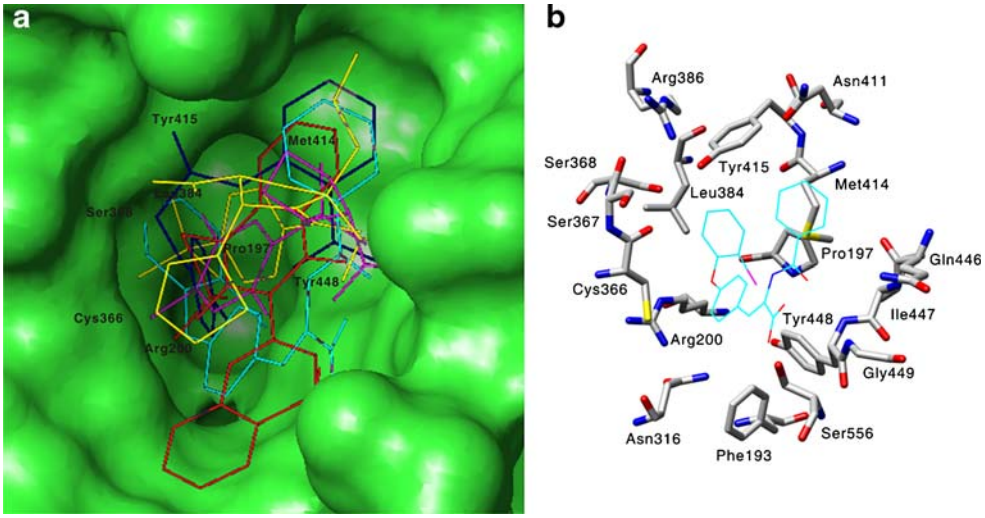


**Fig. 5** (**a**) The deep lipophilic binding pocket in the palm binding site. Residues which form this pocket are labeled. The five NNIs are shown in stick (**b**) Palm binding site residues shown in stick with NNI-1

energies, whereas polar solvation energies opposed the binding. On the other hand, nonpolar solvation energies contribute slightly favorably. The predicted binding free energy successfully ranks the NNI-1 (-38.95 kcal $mol^{-1}$) and NNI-3 (-36.48 kcal $mol^{-1}$) in the first two positions which is consistent with the result of experimental analyses (-9.54 and -8.97 kcal $mol^{-1}$, respectively). However, it does not discriminate the difference between NNI-2, NNI-4 and NNI-5, which implies that the predicted relative binding energy is better used for distinguishing high affinity compounds from low affinity compounds than for sorting similar affinity compounds.

To further elucidate the key residues for NNI binding and examine their contribution modes, the binding free energy was decomposed on per residue with the MM-GBSA method. Figure 4 depicts the important residues for NNI binding identified by the free energy decomposition. In general, Pro197, Arg200, Cys366, Met414, and Tyr448 show more significant and common favorable contributions for the binding than other residues. Previous replicon based experiments using benzothiadazine NNI have shown that mutations of M414, Y448 were sufficient to confer resistance to these compounds [41, 42]. Consistent with this observation, our analysis suggests that the resistance by these two mutations results most possibly from the decreased binding affinity of the inhibitors with the enzyme because of the loss of their favorable contributions during mutations. In addition, previous experiment proved the lipophilic binding pocket which is formed by residues Pro197, Arg200, Cys366, Ser368, Leu384, Met414, Tyr415, and Tyr448 (Fig. 5 and Fig. 6) affords substantial hydrophobic interactions. For example, the bromobenzene group of NNI-1 extends into this deep, lipophilic pocket affording substantial hydrophobic interactions [13]. The aryl group of the sulfonamide region in compound NNI-2 packs between the hydrophobic side-chains of Tyr448 and Tyr415 and above the side chain of Met414 [14]. The cyclopropyl group of NNI-3 extends deeper into this lipophilic pocket [15]. The 4-$CF_3$ group in the benzamide of NNI-4 locates also in this lipophilic pocket [16]. The phenoxy ring of compound NNI-5 and its substituents reside in a spherical hydrophobic pocket of 3.3–3.9 Å defined by Pro197, Leu384, Tyr448, Arg200, and Met414 [17]. In agreement with these experimental data, these favorable contributions of residues Pro197, Arg200, Cys366, Met414 and Tyr448 in this pocket, according to our free energy decomposition analysis, partially explain their tight binding between inhibitors and the enzyme.

## Docking protocol and cross-docking

We have used molecular dynamics and free energy decomposition to perform a thorough investigation of different roles of the NS5B palm binding site residues

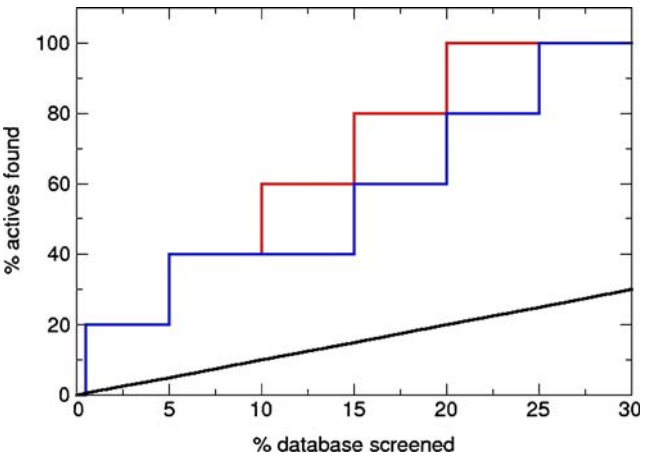


**Fig. 6** Enrichment curve. Red: from NCI diversity set. Blue: from Maybridge diversity set. Black: random

**Table 3** The RMSD summary of cross-docking five known inhibitors with a 60×60×60 grid box and 10 runs

| | | Ligands | | | | |
|---|---|---|---|---|---|---|
| | | 1YVF (NNI-1) | 2GC8 (NNI-2) | 2GIQ (NNI-3) | 2JC0 (NNI-4) | 2QE5 (NNI-5) |
| Proteins | 1YVF | 0.67 | 3.22 | 3.85 | 1.53 | 1.65 |
| | 2GC8 | 4.66 | 1.44 | 3.84 | 1.40 | 1.93 |
| | 2GIQ | 4.75 | 2.35 | 4.18 | 3.10 | 0.51 |
| | 2JC0 | 4.62 | 4.57 | 4.15 | 0.56 | 3.62 |
| | 2QE5 | 4.75 | 2.76 | 4.11 | 6.02 | 1.07 |
| | | RMSD < 2.5 | 2.5 ≤ RMSD | | | |

binding to five different scaffold inhibitors. In the next step, we explored the feasibility of docking-based drug design in this binding site. Autodock which is a free and widely-used program was selected for our docking experiment. Although most of the default parameters in Autodock are good enough, some parameters should be optimized to different specific ligand/receptor systems. The optimization can be based on the ability of recovering the crystal poses of known inhibitors in their lowest energy docked conformations, and also on the computation time. Meanwhile, with the availability of more and more crystal structures of the same protein target, not only self-docking but also cross-docking are critical metrics for evaluating the performance of docking.

Two important parameters of Autodock: grid box size and number of runs were optimized, which proved to be appropriate for our system. We have tried two box sizes: 60×60×60 and 30×30×40. Docking runs were varied as 10, 30, and 50 runs. We have also considered changing ga_num_evals (maximum numbers of energy evaluation number) from medium $2.5\times10^6$ to long $2.5\times10^7$, but the computation time was extremely increased. Thus, we set all ga_num_evals to medium $2.5\times10^6$. Normally, a successful docking is defined with the ligand heavy atom RMSD threshold of 2.0 Å mostly in self-docking. For cross-docking evaluations, given the uncertainty from protein superposition and the fact that the binding mode can be properly conveyed even for RMSD values above 2.0 Å [43, 44], a threshold of 2.5 Å is reasonable for recognizing correctly docked structures. In our study, we retain the definition of 2.5 Å as our evaluation threshold.

First, we ran the docking with two grid box sizes: 60×60×60 and 30×30×40 respectively and the ga_runs were set to the same 10 runs. Tables 3 and 4 show the RMSD summary of cross-docking the five inhibitors with the two different box sizes respectively. The rows represent the protein models, while the columns correspond to the ligands. Only the first-ranked docked conformations (the lowest energy conformations) were considered, which means that the docking was deemed to be correct with only the correct docked conformation in the first position. The docking result from the smaller box size is apparently better than that from the larger one. The number of wrong dockings decreases from 15 to 10. With the larger box size, many RMSD deviations are much higher such as docking NNI-1 to protein structures of 2GC8, 2GIQ, 2JC0 and 2QE5 with 4.66, 4.75, 4.62, and 4.75 Å respectively, docking NNI-4 to protein structures of 2QE5 with 6.02 Å. Moreover, there is not any correctly docked conformation for NNI-3. After checking these wrong conformations with large deviations we found that most of the ligand poses are far away from the center of the binding site pocket, which indicates the larger box size is responsible for the wrong docking. It is reasonable because the palm binding site is located in a pocket of the

**Table 4** The RMSD summary of cross-docking five known inhibitors with a 30×30×40 grid box and 10 runs

| | | Ligands | | | | |
|---|---|---|---|---|---|---|
| | | 1YVF (NNI-1) | 2GC8 (NNI-2) | 2GIQ (NNI-3) | 2JC0 (NNI-4) | 2QE5 (NNI-5) |
| Proteins | 1YVF | 1.14 | 2.68 | 3.29 | 4.74 | 0.80 |
| | 2GC8 | 4.03 | 1.45 | 1.54 | 0.86 | 3.73 |
| | 2GIQ | 1.80 | 2.37 | 1.10 | 2.53 | 3.95 |
| | 2JC0 | 2.13 | 2.40 | 3.20 | 0.62 | 0.76 |
| | 2QE5 | 1.92 | 2.88 | 3.48 | 1.17 | 0.90 |

**Table 5** The RMSD summary of cross-docking five known inhibitors with a 30×30×40 grid box and 30 runs

| | | Ligands | | | | |
|---|---|---|---|---|---|---|
| | | 1YVF (NNI-1) | 2GC8 (NNI-2) | 2GIQ (NNI-3) | 2JC0 (NNI-4) | 2QE5 (NNI-5) |
| Proteins | 1YVF | 0.94 | 3.16 | 4.13 | 1.53 | 1.41 |
| | 2GC8 | 1.92 | 1.41 | 1.54 | 0.89 | 1.15 |
| | 2GIQ | 2.35 | 2.36 | 3.89 | 1.76 | 0.59 |
| | 2JC0 | 3.98 | 2.24 | 3.62 | 0.95 | 0.88 |
| | 2QE5 | 1.90 | 2.75 | 3.98 | 0.74 | 0.50 |

active site cavity of NS5B polymerase and if the box size is too large, the ligand could occupy other space in this cavity. Thus, an appropriate box size is necessary to cover the pocket in this case. By inspection, we selected the 30×30×40 box size, which has led to an improved docking performance.

Next, we investigated the number of runs to access its influence on the protocol performance. Extra 30 and 50 runs combined with the 30×30×40 box size were tested individually. Compared to 10 runs, the improvement is evident with 30 and 50 runs. The number of incorrect dockings decreases from 10 to 7 and then to 5 (Tables 5 and 6). We did not try more runs because the computation time for one docking with 50 runs has increased to about 40 minutes and more time is not appropriate for screening large compound library. Finally, the combination of grid size 30×30×40, ga_run 50, ga_num_evals $2.5\times10^6$ (all other parameters were left at the default values) were selected to be the optimal software conditions for our system.

Obviously, protein flexibility in the palm binding site has few impacts on the docking according to the cross-docking results. In addition, the final aim is to virtually screen and discover new leads; therefore, we selected the protein model of 2GC8 as a representative protein for our further screening research. The guideline we have used is to select the protein structure which can reproduce as many as correct conformations of known inhibitors in a reasonable computation time. All 5 inhibitors were correctly docked in the protein model of 2GC8 and 2GIQ with 50 runs. Considering the smaller predicted RMSDs for ligands from protein model of 2GC8, we determined to apply protein model of 2GC8 with the combination of 30×30×40 box size and 50 runs as our virtual screening tool to the virtual screening.

## Docking to the representative protein structure

As we can see, docking with input ligand geometries directly extracted from X-ray structures is relatively easy and biased. However, in most of the cases, the X-ray conformation of the ligand is unknown and a random conformation is used as starting ligand geometry for docking. To investigate the sensitivity of the docking protocol to ligand geometry, we have prepared a random conformation for each of the five inhibitors and docked them to the representative protein structure. Despite the drop off of docking accuracy compared to using the x-ray conformation, 3 out of the 5 NNIs were predicted and ranked correctly in the lowest-energy position (Table 7). The correctly predicted poses of NNI-3 and NNI-5 were obtained in the 50 output poses of Autodock, but they were not scored in the first position. In fact, it is well known that the accuracy of current scoring functions for small compounds is still poor [45, 46]. Up to now, it is not possible to find a scoring function which can correlate quite well with the observed activities.

**Table 6** The RMSD summary of cross-docking five known inhibitors with a 30×30×40 grid box and 50 runs

| | | Ligands | | | | |
|---|---|---|---|---|---|---|
| | | 1YVF (NNI-1) | 2GC8 (NNI-2) | 2GIQ (NNI-3) | 2JC0 (NNI-4) | 2QE5 (NNI-5) |
| Proteins | 1YVF | 0.85 | 2.70 | 3.86 | 1.51 | 1.36 |
| | 2GC8 | 1.99 | 1.40 | 1.50 | 1.41 | 1.94 |
| | 2GIQ | 2.24 | 2.37 | 1.80 | 1.68 | 0.65 |
| | 2JC0 | 2.19 | 2.24 | 3.17 | 0.67 | 1.19 |
| | 2QE5 | 1.84 | 2.90 | 3.77 | 1.16 | 0.63 |

**Table 7** Docking the 5 known inhibitors with random conformations to the protein model of 2GC8 (All energies are in kcal $mol^{-1}$)

| | Lowest energy pose | | | Lowest RMSD pose | | |
|---|---|---|---|---|---|---|
| Inhibitors | RMSD | Rank | Energy | RMSD | Rank | Energy |
| NNI-1 | 2.26 | 1 | -8.10 | 2.26 | 1 | -8.10 |
| NNI-2 | 1.46 | 1 | -7.24 | 1.35 | 8 | -7.14 |
| NNI-3 | 4.38 | 1 | -7.57 | 1.89 | 43 | -6.57 |
| NNI-4 | 1.33 | 1 | -7.71 | 1.33 | 1 | -7.71 |
| NNI-5 | 4.43 | 1 | -7.42 | 0.79 | 17 | -6.65 |

## Virtual screening against diversity sets

With the optimized docking protocol the single representative protein structure was used as the target for virtually screening two diversity sets: NCI diversity set and the HitFinder collection. Before the screening, we added the five known NNIs with the random conformations to the two diversity sets. It is common practice to see a library of drug-like molecules with known actives and use the enrichment factor obtained to evaluate the accuracy of the docking and scoring functions of the software [47]. Figure 6 shows the percentage of the number of known inhibitors retrieved when increasing the fraction of the ranked list. As can be seen, 40% of the actives were both recovered in the top 5% of ranked NCI and HitFinder diversity set. Overall, a state-of-the-art virtual screening tool rarely extracts 100% of the actives in the top 10% and even more rarely in the top 5%; 50–60% in the top 5% is more commonly observed with the best programs [47]. In addition, considering that all the actives are low micromolar lead compounds and new actives are possibly existing in the decoy set, the enrichment achieved by this docking protocol is reasonable and promising.

The top 10% (182) and 2% (282) compounds of NCI and HitFinder diversity set respectively obtained by our

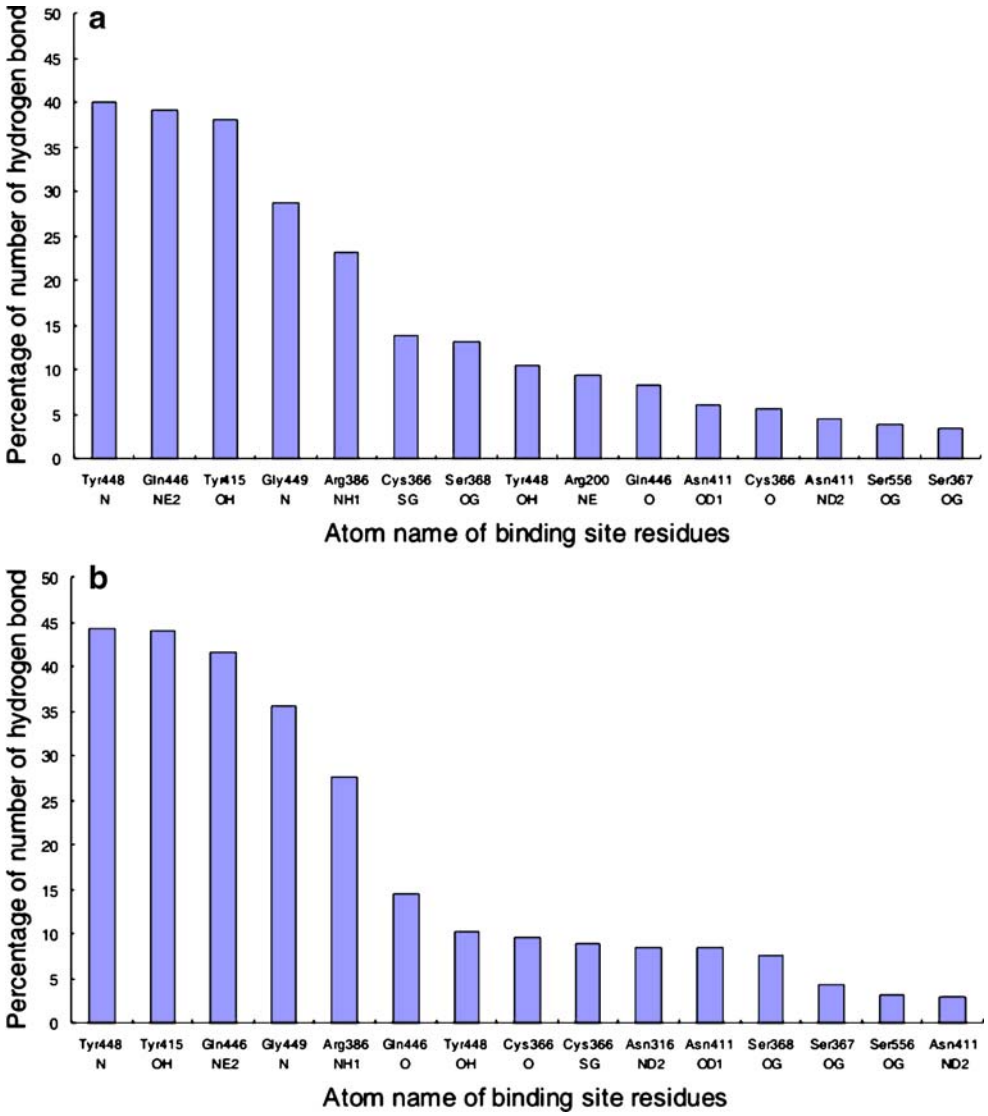


**Fig. 7** Statistics of hydrogen bonds formed between the top-ranked compounds of NCI and Maybridge diversity sets and the atom of binding site residues. (**a**) NCI diversity set (**b**) Maybridge diversity set

virtual screening, were extracted and their first-ranked binding poses were submitted to hydrogen bonding analysis. On one hand, we want to explore the binding site profile *via* the analysis and statistics of hydrogen bonds formed between the top-ranked diversity set compounds and the binding site residues. On the other hand, appropriate hydrogen bonds can be considered as an important factor to help select the good candidate hits. Figure 7 shows the statistics of hydrogen bonds formed between top-ranked compounds of the screening and the atoms of the binding site residues.

As can be seen, about 40% and 45% top-ranked compounds from NCI and HitFinder diversity set form at least one hydrogen bond with Tyr448 N. The critical role of Tyr448 for the recognition of known inhibitors is consistent with our MD and free energy decomposition analysis. The first five hotspot residues which are defined according to the number of forming hydrogen bonds with the diversity set compounds are Tyr448, Gln446, Tyr415, Gln449, Arg386 based on the NCI diversity set screening and Tyr448, Tyr415, Gln446, Gln449, Arg386 based on HitFinder diversity set screening. Although the order of ranking is a little different where Gln446 was ranked before Tyr415 based on the NCI diversity set, the key residues recognized are almost the same for the two screenings.

Suggested by the hydrogen bonding profile of the top-ranked virtual screening compound, our MD and the free energy decomposition analysis, we propose two criteria which can help select the candidate hits. The first one is that there are at least two stable hydrogen bonds formed between the compound and any two residues of Cys366, Ser367, Arg386, Tyr415, Gln446, Tyr448, Gln449 and Ser556. The second criteria is that there should be tight binding in the deep hydrophobic pocket surrounded by Pro197, Arg200, Cys366, Met414 and Tyr448, such as strong vdW or electrostatic interaction.

## Conclusions

In this study we performed 10 ns fully unrestrained MD simulations for five inhibitors/protein complexes. These five inhibitors bound in the same palm binding site, represent five different scaffolds including acrylic acids, proline sulfonamides, thiadiazine, acyl pyrrolidine and anthranilic acid. For each system, we identified the key residues and characterized them structurally and energetically using MM-GBSA-based free energy calculation and decomposition. Our results show that Tyr448 has the most critical role in the binding of most inhibitors mainly by strong hydrogen bonding, which is consistent with previous experimental data. Moreover, favorable contributions of residues Pro197, Arg200, Cys366, Met414, and Tyr448 in a deep hydrophobic pocket prove important for the selectivity of inhibitors.

In an attempt to facilitate and accelerate the structure-based drug design for new inhibitors of NS5B polymerase, an optimized docking protocol was presented based on cross-docking the five known inhibitors into the palm binding site. Then, the optimized docking protocol with a selected representative protein model was used to virtually screen NCI and Maybridge diversity set libraries. The palm binding site was profiled *via* statistics and analysis of the hydrogen bond network formed between the receptor and the top-ranked diversity set compounds. According to the profile, Tyr448 was recognized as the most targeted residue with the highest frequency of hydrogen bond occurring between it and the promising compounds. The following residues Gln446, Tyr415, Gln449, and Arg386 were found forming hydrogen bond, with top-ranked compounds more often. We have proposed two useful rules to guide the selection of promising hits based on our detailed binding site analysis.

## References


1. Michael PM, Graham RF, Jurgen KR, Stefan Z, Fabien Z, Michael H (2007) Nat Rev Drug Discovery 6:991–1000
2. Clercq ED (2007) Nat Rev Drug Discovery 6:1001–1018
3. Gordon CP, Keller PA (2005) J Med Chem 48:1–20
4. Lesburg CA, Cable MB, Ferrari E, Hong Z, Mannarino AF, Weber PC (1999) Nat Struct Biol 6:937–943
5. Francesco RD, Carfi A (2007) Adv Drug Deliver Rev 59:1242–1262
6. Amzel LM (1998) Curr Opin Biotechnol 9:366–369
7. Thiel KA (2004) Nat Biotechnol 22:513–519
8. Jorgensen WL (2004) Science 303:1813–1818
9. Tong L, Froeyen M, Herdewijn P (2008) J Mol Graph Model 26:813–823
10. Wang W, Kollman PA (2000) J Mol Biol 303:567–582
11. Hou TJ, Yu R (2007) J Med Chem 50:1177–1188
12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucl Acids Res 28:235–242
13. Pfefferkorn JA, Greene ML et al. (2005) Bioorg Med Chem Lett 15:2481–2486
14. Gopalsamy A, Chopra R, Lim K et al. (2006) J Med Chem 49:3052–3055
15. Pogam SL, Kang H et al. (2006) J Virol 80:6146–6154
16. Slater MJ, Amphlett EM et al. (2007) J Med Chem 50:897–900
17. Nittoli T, Curran K, Insaf S et al. (2007) J Med Chem 50:2108–2116
18. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE III, DeBolt S, Ferguson D, Seibel G, Kollman PA (1995) Comput Phys Commun 91:1–41
19. Case DA, Pearlman DA et al. (2008) Amber 10, University of California
20. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) J Comput Chem 25:1157–1174
21. Jakalian A, Bush BL, Jack DB, Bayly CI (2000) J Comput Chem 21:132–146
22. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman PA (2003) J Comput Chem 24:1999–2012

23. Jorgensen WL (1982) J Chem Phys 77:4156–4163
24. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) J Chem Phys 79:926–935
25. Darden T, York D, Pedersen L (1993) J Chem Phys 98:10089–10092
26. Miyamoto S, Kollman PA (1992) J Comput Chem 13:952–962
27. Weiser J, Shenkin PS, Still WC (1999) J Comput Chem 20:217–230
28. Sitkoff D, Sharp KA, Honig B (1994) J Phys Chem 98:1978–1988
29. Onufriev A, Bashford D, Case DA (2000) J Phys Chem B 104:3712–3720
30. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) J Comput Chem 25:1605–1612
31. Schrödinger (2006) Maestro, Version 7.5. LLC, New York, NY
32. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219–3228
33. Sanner MF (1999) J Mol Graphics Model 17:57–61
34. The Open Babel Package, Version 2.0.2. http://openbabel.sourceforge.net/ Accessed Oct 2006
35. Repositories, http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html. Accessed Feb 2008
36. Repositories, http://www.maybridge.com/. Accessed Jul 2008
37. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) J Comput Chem 28:1145–1152
38. McDonald IK, Thornton JM (1994) J Mol Biol 238:777–793
39. Wallace AC, Laskowski RA, Thornton JM (1995) Prot Eng 8:127–134
40. Stoica I, Sadig SK, Coverney PV (2007) J Am Chem Soc 130:2639–2648
41. Tomei L, Altamura S et al. (2004) Antivir chem chemother 16:225–245
42. Nguyen TT, Gates AT et al. (2003) Antivir chem chemother 47:3525–3530
43. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R (2005) Proteins 60:325–332
44. Cavasotto CN, Abagyan RA (2004) J Mol Biol 337:209–225
45. Wang R, Lu Y, Wang S (2003) J Med Chem 46:2287–2303
46. Ferrara P, Gohlke H, Price DJ, Brooks CL III, Klebe G (2004) J Med Chem 47:3032–3047
47. Corbeil CR, Englebienne P, Yannopoulos CG et al. (2008) J Chem Inf Model 48:902–909