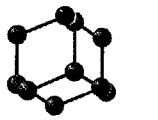
Current Bioinformatics, 2016, *11*, 382-391



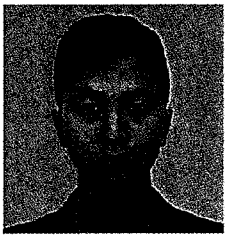
BENTHAM SCIENCE

A Computational Study of Three Frequent Mutations of EGFR and their Effects on Protein Dimer Formation and Non-Small Cell Lung Cancer Drug Resistance

Zhiyong Shen[1,*], Debby D. Wang[1], Lichun Ma[1], Hong Yan[1], Maria P. Wong[2] and Victor H.F. Lee[2]

[1]*Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong*

[2]*Li Ka Sing Faculty of Medicine, University of Hong Kong, Pokfulam, Hong Kong*



Zhiyong Shen

Abstract: Drug resistance is a major problem for non-small cell lung cancer (NSCLC) treatment due to mutations in patients' DNA sequences. It is now possible to obtain the human genome information easily based on the high-throughput sequencing technology, so personalized medicine can become a reality.

Based on mutation data of 168 patients with stage IIIB and IV NSCLC. We use computational method to predict the homo-dimers and hetero-dimers formation and compute the binding free energy of complexes (between drugs and proteins). For the gefitinib and erlotinib as two common drugs used in patient's therapy, we compute the possible 3D structure of epidermal growth factor receptor (EGFR) mutant- inhibitor complex. Rosetta and Amber are used for molecular dynamics analysis and simulation. The PRISM protocol is used to predict the binding energy based on similar protein-protein interaction surfaces. Multiple factors, including the mutant proteins surface geometry change, the number of hydrogen bonds change and the electronic change of the surface, are taken into account when in evaluating the binding free energy.

Our results suggest that the mutation position is very important for dimer formation and it affects the drug's binding strength with EGFR. Mutations such as L858R and T790M which do not happen on the protein interaction surface can hardly affect the formation of dimers. Patients with the delE746_A750 mutation can obtain a good therapy by using gefitinib instead of erlotinib. By comparing the binding free energy to form a homo- or heterodimers, we find that the L858R mutant will incline to form a hetero-dimer rather than a homo-dimer.

Keywords: Dimer formation, EGFR, drug resistance, mutation, hetero-dimer, homo-dimer.

Received: January 27, 2015 **Revised: October 30, 2015** **Accepted: March 21, 2016**

1. INTRODUCTION

There are different treatments for different types of lung cancers, and the non-small cell lung cancer (NSCLC) can be treated with surgery, chemotherapy or radiotherapy etc., depending on the stage of the cancer [1]. Especially in the stage 4, the treatment aims to control the cancer for as long as possible and to reduce the size of tumor. Drugs, such as erlotinib and gefitinib, can be applied to those patients, who have NCLC due to the mutation of the protein epidermal growth factor receptor (EGFR) [2]. Although these drugs are effective initially, almost all patients develop drug resistance after several months to a year due to another mutation of EFGR [3]. Using modern sequencing technology, it is possible to obtain gene mutation information of a patient easily [4], thus it is useful to analyze how the EGFR mutation affect the drug resistance and prepare personalized treatment of NCLC. Thus it is useful to analyze how the EGFR mutation affects the drug resistance and prepare personalized treatment of NCLC.

Protein dimerization and oligomerization are very common, which can bring advantages in terms of stability and specificity to the proteins involved [5-7]. In this work, we study how different EGFR mutations affect the dimer

formation of ERBB family members, ERBB1 (EGFR), ERBB2, ERBB3, and ERBB4, which then lead to drug resistance. The computational approaches as an efficient technique will be used here to investigate drug resistance by computing the binding free energy between the mutants and drug ligands [8]. In our study, we developed a pipeline for predicting possible EGFR mutation- induced drug resistance based on multiple factor evaluation by computer simulation. Our reference experiment data came from the Queen Mary Hospital of Hong Kong, and the whole data set consists of 168 patients whose stage in IIIB and IV non-small cell lung cancer (NSCLC) and their mutations are focused on exons 18 ~ 21 of the EGFR tyrosine kinase domain. We will choose the most common mutation types (L585R, T790M, V948R) from these patients for computer simulation and apply molecular mechanics to calculate the binding affinities of EFGR protein families with a total of 2 drug ligands (erlotinib and gefitinib). During solvent MD simulation, two conventional methods (MM-PBSA and MM-GBSA) were applied to estimate the binding free energy. The hydrogen bond contribution, including dimer conformational restriction and protein-ligand binding was evaluated using different methods [9]. The overall efficiency and accuracy found from this study suggest that a molecular simulation is a viable

*Address correspondence to this author at the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, China; Tel: +852-56242611; E-mail: szypanther@gmail.com

2212-392X/16 $58.00+.00

approach for predicting the drug resistance in specific mutant patients.

Specifically, the electrostatic [10] and van der Waals forces [11] in protein-protein interactions, provide essential conditions for EGFR to form an asymmetric dimer of two kinases first. The homo- or hetero-dimerzaition of the kinase domain of an EGFR-family receptor, which consists of an N-lobe face and a C- lobe face, can be activated by an allosteric mechanism and switch on the downstream signaling [12]. The NSCLC caused by EGFR mutations are initially responsive to small molecule tyrosine kinase inhibitors (TKIs), such as gefitinib (IRESSATM) and erlotinib (TARCEVA) [13]. A well-known mutation L858 of EGFR shows a good response to the drugs. However, the second mutations, such as T790M can cause cause drug resistance. Over 50% of resistance is caused by the mutation in the ATP binding pocket of the EGFR kinase domain, and this will prevent the binding of the drug erlotinib through steric hindrance [14]. That is, the amino acid threonine at position 790 acts like a "gatekeeper" residue and its mutation strengthens ATP binding but reduces the binding of erlotinib [14]. All these situations can be explained by evaluating the binding free energy of the mutant-drug complex, including the difficulty level to form a dimer and the total number of potential intermolecular hydrogen bonds. In this process, how to obtain a reliable dimer structure will become an important step for further investigation. Here, we will carry out statistical analysis of six mutation types from one to three point mutations and a deletion mutation. The specific mutations involved are as follows L858R, T790M, V848R, L858R_T790M, T790M_L858R_V948R and delE746_A750. Most patients have been included by the above mutation types, with 80 cases of L858R and 38 cases of delE746_A750. The flowchart of our study is shown in Fig. **1**.

2. METHODS

2.1. Reference Patient Data Collection

The reference patient data are obtained from the Queen Mary Hospital in Hong Kong. The dataset consists of 168 non-small cell lung cancer (stages IIIB and IV) patients with mutations on exons 18~21. of the EGFR tyrosine kinase domain. They have a total of 37 EGFR mutation types (Table **1**), including residue substitution, deletions, insertions and double-point mutations. Two common therapy drug geftinib and erlotinib have been used in NSCLC treatment [15]. Here, we choose to collect some of the major mutation types (L858R, delE746_A750, delL747_P753insS) in the clinical observation data, with the PFSs of the corresponding inhibitors. In the analysis, if the number of patients is more than five, we will filter out the highest value of PFS and the lowest value of PFS and then to compute the mean value of remaining PFSs as the corresponding mutant type's PFS.

2.2. EGFR Monomeric Mutants Generation

Rosetta and Amber are used to build the protein structure and analysis. Here, we use Rosetta 3.4 to produce possible EGFR monomer mutants. The method of homology modeling is employed to model a protein mutant using a known similar structure as template. In order to obtain a more reliable mutant structure, we find the crystal structure of these mutant types in PDB database first, and then use the minirosetta comparative modeling protocol of Rosetta to predict the complete 3D protein structure of the mutant. Here, we will focus on six EGFR monomeric mutants structure (L858R, T790M, V948R, L858R_T790M, L858R_T790M_V948R, and DelE746_A750) for further analysis. The reference template crystal structures including "3W2O","4I1Z","2ITU" and "2GS2" are consider for each

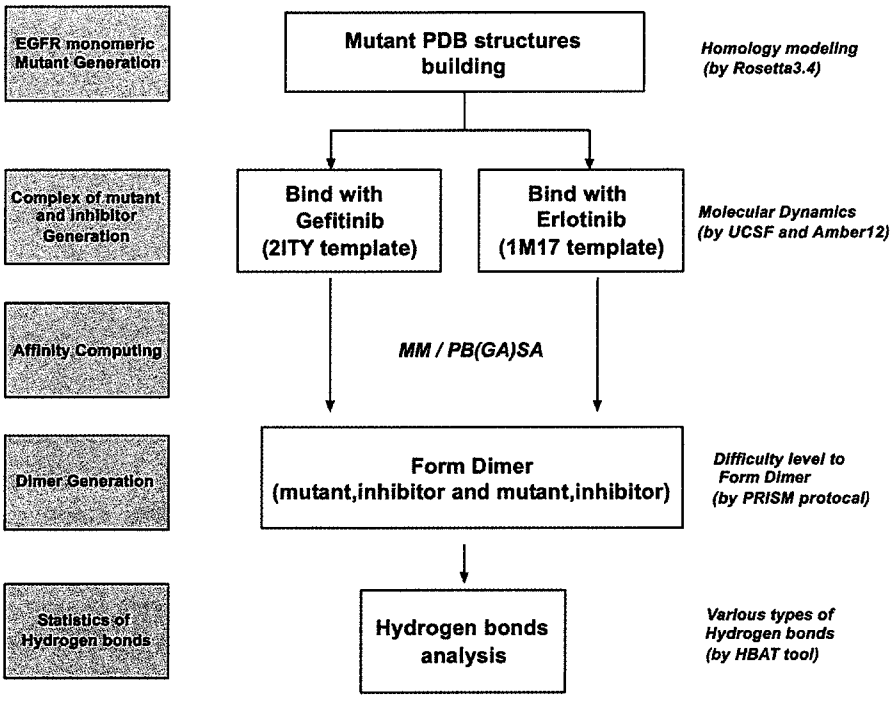


Fig. (1). The flowchart of our computational method.

Table 1. Binding affinity calculation (complex-receptor ligand) (with parameters *mmpbsa_debby.in: Inp=1, end frame=50).* The complex-receptor-ligand structures come from the production period MD results.

Name	Poisson Boltzmann	Generalized Born
prod_L858R_gef.pdb	-29.532	-33.4123
prod_delE746_gef.pdb	-37.2888	-43.6836
WT_2ITY_gef.pdb	-23.2195	-28.3240
prod_T790M_gef.pdb	-18.3718	-21.6586
prod_L858R_T790M_gef.pdb	-21.0164	-24.5047
prod_V948R_gef.pdb	-31.8136	-39.791
prod_L858R_T790M_V948R_ gef.pdb	-20.0102	-25.5585
prod_L858R_T790M_V948R_ ero.pdb	-25.6769	-29.9581
prod_delE746_ero.pdb	-28.9963	-39.6788
prod_L858R_T790M_ero.pdb	-30.1159	-33.5697
prod_V948R_ero.pdb	-24.3458	-29.2192
WT_1M17_ero.pdb	-31.6901	-39.7356
prod_L858R_ero.pdb	-28.1155	-34.4142
prod_T790M_ero.pdb	-27.0228	-34.0539

corresponding mutant generation. In the process of model construction, we use the multiple-sequence alignment program ClustalW [16] to align the target sequence to the template, and at the same time, a fragment library is produced. We submit the target sequences to the official website service (http://robetta.bakerlab.org/fragmentsubmit.jsp) to create 3mer and 9mer fragment libraries. These fragment libraries will be used in fragment insertion during the corresponding 3D protein structure prediction separately. Subsequently, the 3D structures of the monomer mutants are evaluated by their physics-based energies. The minimum energy structure will be chosen as the most accurate model for further analysis. In this process, the full atom energy scoring function will be employed to identify the best prediction result. This scoring function is a model generated using various contributions including Lennard-Jones interactions, residue pair interaction, van der Waals self-energy, hydrogen bonding and unfolded state reference energy etc. The predicted total score of the structure is evaluated by computing the weighted sum of the above terms. After obtaining all of the best mutant 3D protein structures, the minimization step for each structure will be carried out using Amber 12.

2.3. Complex of Mutant-Inhibitor Generation

The optimized 3D mutant protein structure will be aligned to the template to build a protein-ligand complex, using the structure comparison tool of the UCSF Chimera [17]. The templates here are 2ITY and 1M17 for the two different drugs (gefitinib and erlotinib). Then we replace the corresponding alignment structure and combine the drug with the mutant to obtain the mutant-drug complex. The complex is computationally solvated into an common octahedron water box (TIP3P water model) [18] with a 10 angstrom (compensate the large computational costs) buffer around the complex in each direction and the ff99SB force field and gaff force filed will be applied in this stage [19]. The ff99SB force field is good for management of the helical and extended regions of protein backbones, while the gaff force filed will handle small ligands well. After obtaining the protein-ligand structures, molecular dynamic (MD) simulations will be used for further analysis.

In the process of MD simulation, we use Amber (version 12) to execute a rough 1000-step minimization for each mutant-drug solvated complex based on homolog modeling. We equilibrate the solvated complex by implementing 50ps of heating and 50 ps of density equilibration with a little restraints on the structure which is followed by 500ps of constant pressure equilibration at 300K. All the simulation will be run with shake on hydrogen atoms and a fix 2 fs time step, and also include the Langevin dynamics for temperature control. After checking the system has equilibrated through parameters involving temperature, density, total energy and the backbone RMSD, we proceed with the MM-PBSA production MD run. This production simulation is also performed with the same conditions as the final phase of equilibration in order to defend an abrupt jump in the potential energy owing to a change in simulation conditions. Lastly, we obtain the result of the production MD simulation at an interval of 2 fs for 2 ns and collect the trajectory frames at a step of 10 ps and 200 frames in each trajectory. In the end, to evaluate the binding free energy of protein-ligand complex reliably, the stabilization of each system should be confirmed by the complex's backbone RMSD curve [20].

2.4. Binding Free Energy Calculation Using MM/PB (GB)SA Model in Amber

The two methods MM-PBSA (molecular mechanics/ Poisson-Boltzmann surface area) and its complementary tool MM-GBSA (molecular mechanics/generalized Born surface area) are popular used to calculate the free energy difference between two states, which most often represent the bound and unbound state of two solvated molecules. Based on the MM/PB(GB)SA model [21], the production MD simulations produce the motion trajectories of each solvated system, and the binding free energy of the involved mutant-drug complex is calculated based on these trajectories. The binding affinity of a solvated protein-ligand or protein-protein system can be estimated by the binding free energy value in a solvent environment. Its fundamental theory is the thermodynamic cycle, which can be characterized according to the following equation:

$$\Delta GBind, Solv = \Delta GBind, Vacuum + (\Delta G Solv, Complex - \Delta G Solv, Receptor - \Delta G Solv, Ligand) \quad (1)$$

or

$$\Delta GBind, Solv = \Delta GBind, Vacuum + \Delta\Delta GSolv \quad (2)$$

In the above equation, ΔGBind,Solv and ΔGBind, Vacuum represent the free energy difference between the bound and unbound states of a complex in a solvent and vacuum environment respectively. ΔΔGSolv comes from the computation result in the brackets of Equation (1). Δ G Solv,Complex represent the solvation free energy of complex and its two binding partners are Δ G Solv, Receptor and Δ G Solv,Ligand respectively. The free energy of each element comes from a sum of four terms (ΔEMM, ΔGpol ,ΔGnonpol and TΔSnma). Taking the ΔGBind, Solv as example, we have

$$\Delta GBind, Solv = \Delta EMM + \Delta Gpol + \Delta Gnonpol - T\Delta Snma \quad (3)$$

Here, ΔGBind , Solv is the binding free energy in solution state, ΔEMM is the molecular mechanics energy which consists of a van der Waals and an electrostatic contribution, ΔGpol stands for the electrostatic/polar interaction and ΔGpol stands for nonpolar interactions. Considering the high computational complexity, the entropy contribution ΔSnma is usually neglected for states with similar entropies. The electrostatic solvation energy is evaluated using the finite difference Poisson-Boltzmann or generalized Born method, and the nonpolar contribution is obtained using the solvent-accessibility surface method. The last term ΔEMM is obtained from the following equation:

$$\Delta EMM = \Delta Eele + \Delta Evdw \quad (4)$$

where ΔGnonpol represents the hydrophobic contribution to the solvation free energy, which can be obtained from a function of the solvent-accessible surface-area and the following equation:

$$\Delta Gnonpol = \gamma A + b \quad (5)$$

where A is the solvent-accessible surface area and γ and b are empirical parameters.

In our work, a parallel version of the python script of MMPBSA.py.MPI is used to run on a server with multiple processors to accelerate the computations. Each previously obtained MD trajectory, representing a number of conformations, can be recognized as a major input for MMPBSA.

2.5. Dimer Generation by PRISM

Prediction of protein-protein interactions at the structural level is important because it involves the protein function and is useful drugs discovery. Specifically, dimerization and autophosphorylation are the critical events in EGFR activation [22]. How to reliably predict the possible dimer formation of EGFR, including homo-dimers and hetero-dimers, is a crucial aspect for drug efficacy evaluation. Here, a protocol named PRISM (protein interactions by structural matching) is adopted to predict the protein complex structures. The protocol determines if the two target protein's surfaces are similar with the two complementary sides of a known template interface, and predict if these two proteins can interact with each other by using this template interface architecture. PRISM predicts the binding residues by using structural similarity and evolutionary conservation of putative binding residue 'hot spots', and its flexible refinement by using a docking energy function.

In our work, we using PRISM to obtain all possible dimers of mutant-drug complexes based on EGFR related 6 different known templates (2gs2AB,2gs6AC,2jitAB,3ikaAB, 3pp0AB and 3rcdAB), and then the one with the minimum energy will be chosen as the dimer structure for further hydrogen bond analysis.

2.6. Hydrogen Bond Analysis

Hydrogen bond is the electrostatic attraction between polar molecules. It usually occurs when a hydrogen (H) atom bound to a highly electronegative atom (nitrogen, oxygen or fluorine etc.) experiences attraction to some other nearby highly electronegative atom. A hydrogen bond is usually stronger than the van der Waals interaction, but weaker than covalent or ionic bonds. This type of bonds can occur in inorganic molecules such as water and in organic molecules, such as proteins and DNA. That is why we should also analyze the hydrogen bond in the mutant-drug complex dimer. A tool named HBAT is used here to automate the analysis of the potential hydrogen bonds and some other similar type of weak interactions such as halogen bonds and non-canonical interactions in macromolecular structures which is available in the PDB file format.

3. RESULTS AND DISCCUSIONS

3.1. The Template PDB Structures Choice

This Homology modeling is currently the most widely used protein structure prediction method. The more similar the template protein with the target one, the more accurate the predict result will be. In our MD simulation, we choose five corresponding PDB files as the template structures (2GS2, 2ITU,3IKA,3W2O and 4I1Z), and the target EGFR monomeric mutant sequence structure can be obtained by homology modeling based on their most similar sequence protein 3D structure. In the end, we use 2ITU as L858R mutant template, 3IKA as T790M mutant template, 4I1Z as both V948R related mutant template, and 3W2O as the two-point mutations (L858R and T790M) protein template. The inhibitor's (gefitinib and erlotinib) structure can be extracted from 2ITY and 1M17 respectively and all of these 3D structures are shown in Fig. **2**.

3.2. EGFR Monomeric Mutant Generation

Roseatta comparative modeling is an effective way to predict protein structures [23]. It usually includes four steps. The first step is the alignment generation based on one or more template structures. The second step is to produce an incomplete model based on the template and then rebuild the missing parts using loop modeling. The third step is to refine a full-atom protein model by using the Rosetta full- atom energy function. The last step is to select the best models based on clustering, and the clustering method is recognized as a conventional way to identify the best structure among those predicted ones. The reason is that the lowest energy structure, which belongs to the largest cluster, can be considered as being closest to the native structure. Here, we use the minirosetta method [24] to obtain the corresponding 3D structures of the six EGFR monomeric mutants (L858R, T790M, V948R, L858R_T790M, L858R_T790M_V948R,

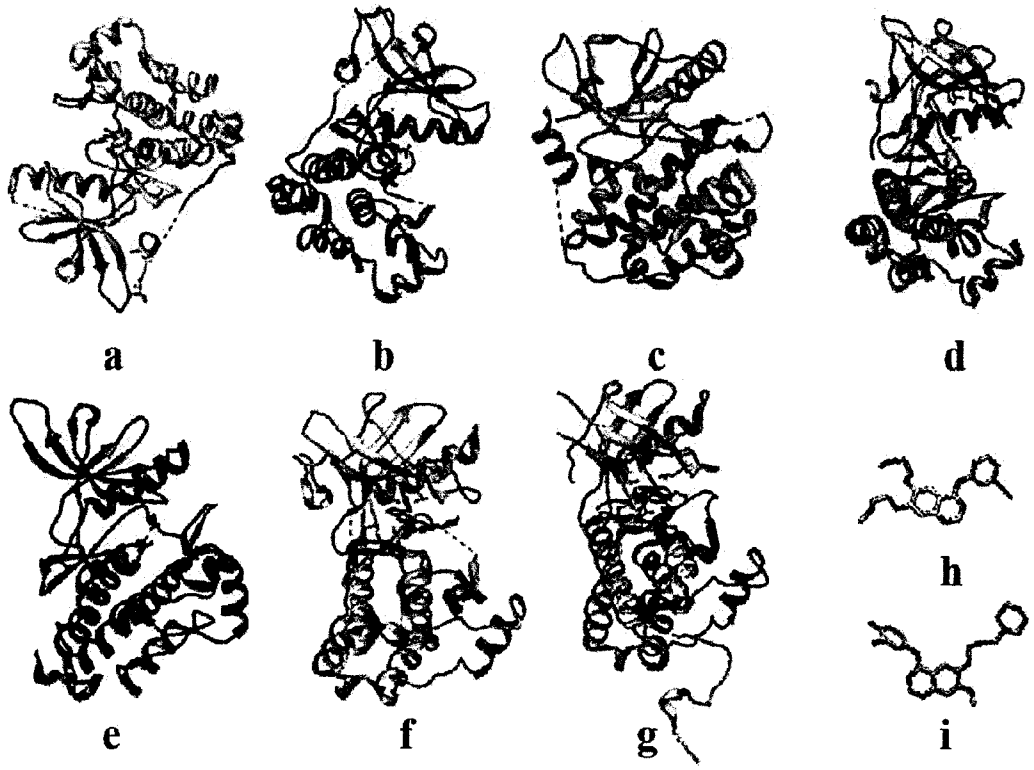


Fig. (2). The 3D templates used in our structural analysis. (a) to (e): the structures of the kinase domains of 2GS2, 2ITU, 3IKA, 3W2O and 4I1Z, respectively. (f) and (g): the complex of 2ITY(composed of an active WT EGFR kinase and gefitinib inhibitor) and 1M17(composed of an active WT EGFR kinase and erlotinib inhibitor). (h) and (i): two 3D structure of inhibitors (erlotinib and gefitinib) respectively.

and DelE746_A750) for the next step of complex prediction. In our study, L858R means the residue substitution of L with R at the 858 residue site. In T790M, the residue substitution of T with M occurs at the 790 residue site, and in V948R, the residue substitution of V with R occurs at the 948 residue site. DelE746_A750 means the deletion of residues at sites from 746 (residue E) to 750 (residue A). L858R_T790M and L858R_T790M_V948R represent a double-point and triple-point mutations respectively.

3.3. Complex of Mutant and Inhibitor Generation

The mutants from the above model-building process are processed in an energy minimization step to remove the structural artifacts before molecular dynamics simulation for computing the binding energy of mutant and inhibitor. Here, we will use the "positional restraint" method in energy minimization, in which the atoms will move some distance away from the reference positions to which they are restrained in the model, and then the new positions will be used as the reference positions for the next iteration. After several iterations, the model will reach a local energy minimum status, where each atom's position will not change further. Unlike the molecular dynamics simulation, none of the atoms has any velocities associated with the energy minimization. The minimization energy state mimics an environment where the temperature is 0K. However, this is totally different from what we usually observe in reality. Before the production step, we need to heat the system up to 300K to make it equilibrate further. The following diagrams of temperature, density, total energy and backbone RMSD [25] will be evaluate to determine whether the system have equilibrated or not. Taking the T790M mutant-gefitinib results as an example (Fig. **3**), we can see that these diagrams (density, temperature and total energy) have converged by the end of equilibration period. The others corresponding results including L858R, V948R, L858R_T790M, L858R_T790M_V948R, and DelE746_A750 mutations can be found in the supplementary figures. All of them can obtain the similar trend for each diagrams. Show the Fig. **3** as follows:

After the equilibration phase of the simulation, we will perform the production runs further. In order to avoid an unexpected jump in the potential energy due to changes in simulation conditions, the production phase of the simulation should be run under the same conditions as the equilibration phase. Here we run a total of 2ns and then check density, temperature, total energy and backbone RMSD again as we did for the equilibration step. These results are shown in Fig. **4**.

3.4. Binding Affinity Calculation Using MM/PB(GB)SA Model in Amber

To obtain the production MD simulation trajectories results, we can carry out the binding energy calculation using two methods, MM-GBSA and the MM-PBSA in Amber, for comparison [26]. MMPBSA.py of the AMBERTOOLS can

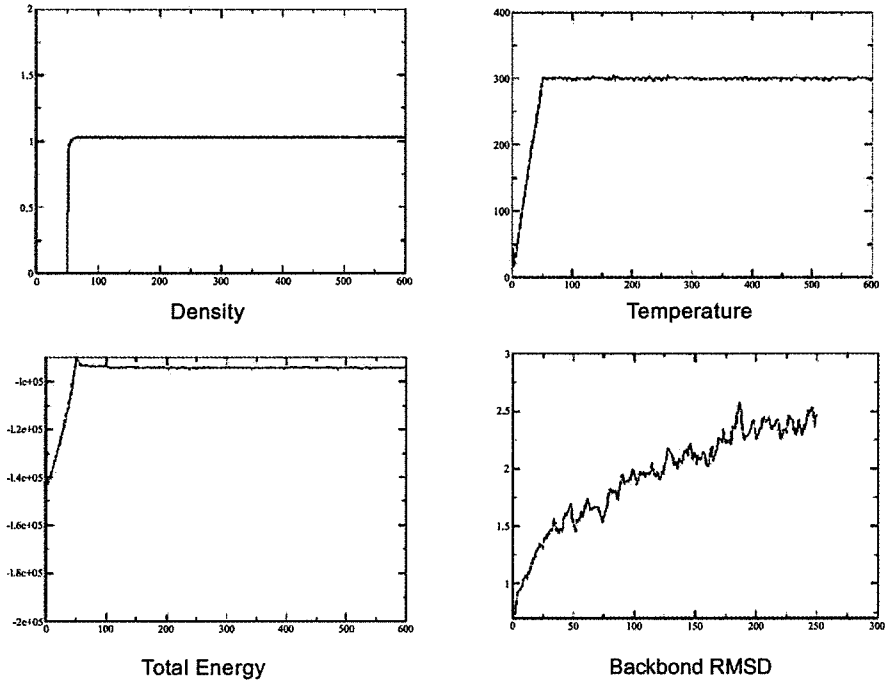


Fig. (3). Density, temperature, total energy and backbond RMSD at the end of equilibration period in T790M mutant-gefitinib complex.

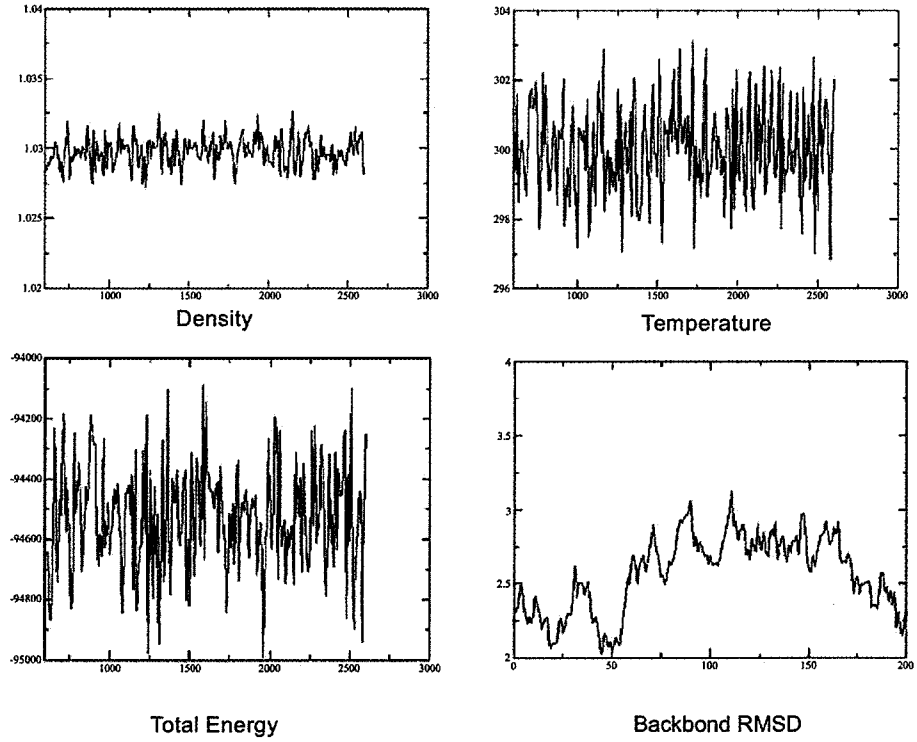


Fig. (4). Density, temperature, total energy and backbond RMSD at the end of production period in T790M mutant-gefitinib complex.

be used to evaluate the free energy of the binding, and these scripts will also create three unsolvated mdcrd files (complex, receptor, and ligand) for using ptraj to generate per-residue RMSF (root mean square fluctuation) values for a complex 3D structure. The process details can be found in the Method part. Here we present the binding affinity of each mutant- inhibitors complex in the Table **1** as follows.

We can find that the mutation type of T790M will decrease the affinity with drugs, especially in the situation of the protein iteraction with gefitinib ligand. The mutation's binding free energy is reduced from -23.2195 kcal/mol to -18.3718 kcal/mol compared with the widetype's. At the same time, even the ligand replaced with erlotinib ligand, we also get the same conclusion that the drug's affinity is lower in T790M mutation than widetypes's, obtain the binding free energy with -27.0228 kcal/mol and -31.6901 kcal/mol respectively. Other study indicates that due to the change of electron distribution, the T790M also will increase ATP affinity as its primary mechanism for drug resistance [27].

3.5. Results for the Forming Dimer by PRISM

A powerful protein-protein interactions prediction protocol (PRISM) will be applied here for our accurate predict the dimer structures [28]. PRISM which uses a multiple protein structural alignment algorithm (MultiProt) engine can predict interface regions from spatial similarities of amino acids disregarding the residue's order on protein chains [29]. In the results, only the partners with at least 50% of the residues match with the target surfaces and similar evolutionary patterns (the number of identical hotspots between them) will be considered. At last, the candidate pairs whose corresponding target proteins of left and right interfaces should not intercept with each other in their possible complex state. The results are listed in Table **2** based on the above multi-step filtering process.

Based on the affinity measure, the dimer with the lowest free binding energy on specific template will be chosen as the best 3D structure of the corresponding dimer. These structures will be used in HBAT [30] to further analysis

Table 2. Dimer affinity prediction results by PRISM based on different templates. Surface match template (pdb self homodimer binding energy)

Mutant Name	*2gs2AB*	*2gs6AC*	*2jitAB*	*3ikaAB*	*3PP0AB*	*3rcdAB*	Best Match
equil_L858R.pdb	*15.51*	*-22.01*	*-29.95*	*-28.83*	NA	NA	***2jitAB***
equil_T790M_L858 R.pdb	-38.19	-48.6	-61.98	-62.99	-62.72	NA	***3ikaAB***
equil_L858R_T790	NA	NA	NA	NA	NA	NA	NA
M_V948R.pdb							
equil_T790M.pdb	-37.61	-39.14	-42.81	-30.91	-20	35.35	***2jitAB***
equil_V948R.pdb	NA	NA	NA	NA	1609.79	NA	NA
equil_delE746.pdb	-69.19	-34.23	-32.85	-35.99	-35.11	NA	***2gs2AB***
3bce(ERBB4)	NA	-70.16	-61.5	-71.25	-39.64	-41.47	***3ikaAB***
3pp0(ERBB2)	-79.78	-70.16	-87.96	-84.92	-79.75	-81.28	***2jitAB***
2ity(ERBB1)	-91.79	-92.44	-88.43	-103.42	-95.53	-19.27	***3ikaAB***

Note: NA means the corresponding results cannot be produced by PRISM. Based on the corresponding similar templates, it is not always possible to form homo-dimers.

Table 3. Hetero-dimer affinity prediction results by PRISM based on 2jitAB template.

Template	Part1	Part2	Binding Free Energy	Dimer Type
2jitAB	2ity	equil_T790M.pdb	-96.48	ERBB1/ERBB1
2jitAB	**3pp0**	**equil_L858R.pdb**	**-80.35**	ERBB1/ERBB2
2jitAB	2ity	equil_L858R.pdb	-75.93	ERBB1/ERBB1
2jitAB	2ity	equil_T790M_L858R.pdb	-74.74	ERBB1/ERBB1
2jitAB	2ity	equil_L858R_T790M_V948R.pdb	-68.81	ERBB1/ERBB1
2jitAB	**3pp0**	**equil_V948R.pdb**	**-63.22**	ERBB1/ERBB2
2jitAB	dell	equil_T790M_L858R.pdb	-62.19	ERBB1/ERBB1
2jitAB	2ity	equil_V948R.pdb	-55.98	ERBB1/ERBB1
2jitAB	**3bce**	**equil_T790M_L858R.pdb**	**-53.63**	ERBB1/ERBB4
2jitAB	**3pp0**	**equil_T790M_L858R.pdb**	**-48.74**	ERBB1/ERBB2
2jitAB	3pp0	equil_T790M.pdb	-50.77	ERBB1/ERBB2

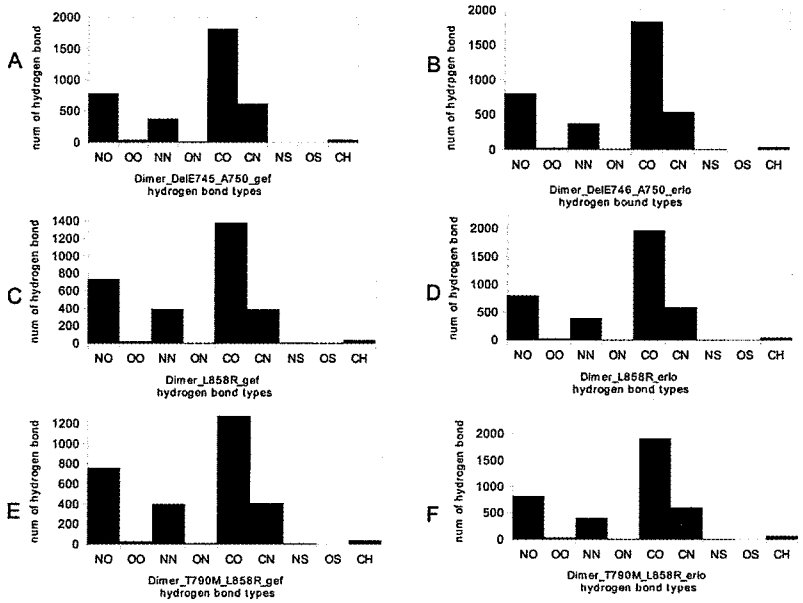


Fig. (5). The hydrogen bond distribution in each dimer structure. NO: N—H...O type bonds, OO: O—H...O type bonds, NN:N—H...N type bonds, ON:O—H...N type bonds, CO:C—H...O type bonds, CN: C—H...N type bonds, NS:N—H...S type bonds, OS:O—H...S type bonds, CH:C—H...S type bonds. (A) and (B): delE746_A750 with gefitinib and delE746_A750 with erlotinib. (C) and (D): L858R with gefitinib and L858R with erlotinib. (E) and (F): T790M_L858R with gefitinib and T790M_L858R with erlotinib.

hydrogen bonds in the dimer. Comparing the two tables (Tables **2** and **3**), we can find that ERBB1 with L858R mutation will form a hetero-dimer ERBB1/ERBB2 with ERBB2 (3pp0) instead of forming a homo-dimer ERBB1/ ERBB1. The binding free energy is -80.35 kcal/mol for the hetero-dimer while it is -75.93 kcal/mol for the one mutation homo- dimer and -29.95 kcal/mol for itself homo-dimer in Table **2**. Similarly, ERBB1 with the structure surface point mutation V948R will also form a hetero-dimer with ERBB2 instead of forming a homo-dimer ERBB1/ERBB1.

3.6. Statistics of Hydrogen Bonds in the Dimers

Hydrogen bonds can be found in many strengths and geometries and they play an important role in determining the protein 3D structures including the dimer structure. In proteins, hydrogen bond donors include O-H, N-H or S-H interacting with the nonbonding electrons of acceptor atoms such as O, N or S. Hydrogen bonds can mediate drug-receptor binding and are important in drug development [31]. We use HBAT to analyze potential hydrogen bonds in the dimers. The distribution of the hydrogen bonds in each dimer is summarized in Fig. **5**.

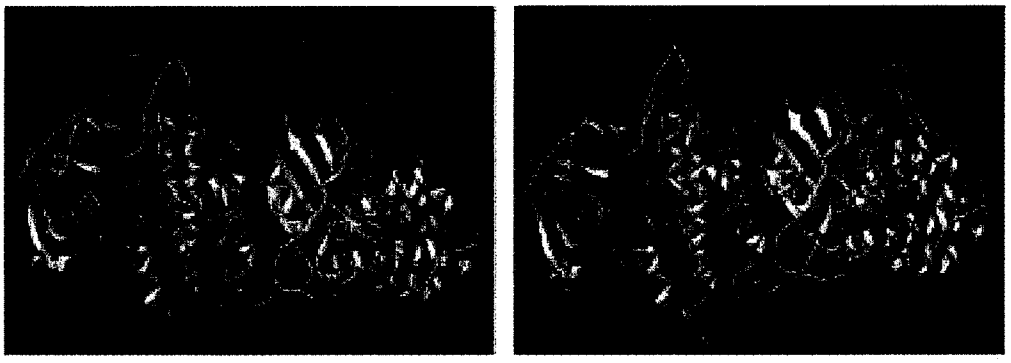
From the above distributions, we can see that many hydrogen bonds are located on the strong (N— H...O, NN:N—H...N) and weak (CO:C—H...O,CN: C—H...N) bonds in protein-ligand complexes. Both dimers share almost the same distribution for each hydrogen bonds type. Multifurcation of hydrogen bonds is also very common in these complex structures. The number of main hydrogen bonds (LD-BA and LA-BD) and the total potential intermolecular hydrogen bonds in each dimer structure is summarized in Table **4**.

The homo-dimer for the mutation delE746_A750 with gefitinib (Dimer_DelE746_gef) has 18 main hydrogen bonds (LD-BA and LA-BD), more than twice in the homo-dimer

Table 4. Hydrogen bond distribution in each dimer structure.

Dimer Name	Main H-bonds (LD-BA and LA-BD)	Total Potential Intermolecular Hydrogen Bonds
Dimer_DelE746_erlo	8	2895
Dimer_DelE746_gef	18	2913
Dimer_L858R_erlo	10	2970
Dimer_L858R_gef	33	3006
Dimer_T790M_L858R_erlo	14	2870
Dimer_T790M_L858R_gef	12	2984

Note:LD=Ligand Donor, BD=Backbone or Main Chain Donor, BA=Backbone or Main Chain Acceptor. LA=Ligand Acceptor.



A **B**

Fig. (6). The 3D structure of Dimer with ligands shown. (**A**), mutation delE746_A750 with gefitinib dimer structure. (**B**), mutation delE746_A750 with erlotinib dimer structures. The blue line represent the found H-bonds in the dimer structures.

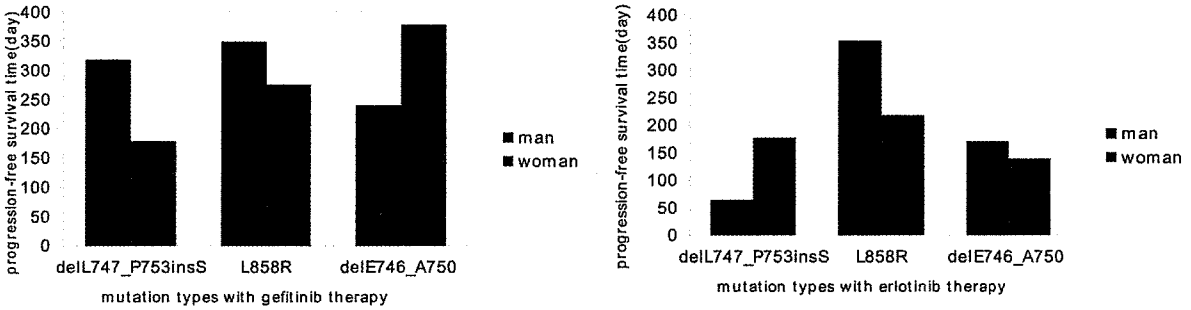


Fig. (7). The relationship between mutation types of patients with progression-free survival (PFS) time under different drug therapy. (**A**) Three major mutation types (L858R: 19 man cases, 49 woman cases, delE746_A750: 7 man cases, 23 woman cases, delL747_P753insS: 4 man cases, 4 woman cases) of patients with gefitinib therapy. (**B**) Three major mutation types (L858R: 5 man cases, 7 woman cases, delE746_A750: 6 man cases, 2 woman cases, delL747_P753insS: 1 man case, 1 woman case) of patients with erlotinib therapy.

for the same mutation with erlotinib (Dimer_DelE746_erlo) which has only 8 main hydrogen bonds (Fig. **6**). This result is consistent with binding free energy values for the two dimers, -37.2888 kcal/mol and -28.9963 kcal/mol respectively. This implied that the patient with delE746_A750 mutation could get a better therapy by using gefitinib instead of erlotinib. For patients with the L858R mutation, the binding free energies when gefitinib and erlotinib are applied are -29.532 kcal/mol and -30.1159 respectively. In this case, the efficiencies of the two drugs are very close. Our following clinical data analysis confirmed this result.

3.7. Clinical Data Analysis

Three major mutation types ((L858R: 80 cases, delE746_A750: 38 cases, delL747_P753insS: 10 cases) cover most patients in our database. Considering the factors including the different gender and therapy drug(gefitinib and erlotinib), we use a bar chart to present the relationship between the progression-free survival (PFS) time with different drug therapy and compute the mean value as PFS value for each mutation type of patient (Fig. **7**).

We can find that the doctor are more inclined to use the gefitinib to treat patients who get the stage IIIB and IV non-small cell lung cancer (NSCLC), and their PFS time is also longer than those patients who use erlotinib. Especially, in the other two major mutations (delE746_A750 and delL747_P753insS), the PFS time of gefitinib therapy is larger than the PFS time of erlotinib therapy.

4. DISCUSSION

In the paper, we use computational methods to confirm why some patients who use gefitinib can have a long PFS time than those patients who use erlotinib. We also find that the mutation types which make it hard to form the dimer structures will help to stop the downstream signaling. Some mutations change the drug binding affinity and cause drug resistance. The number of hydrogen bonds to form between the protein and the drug affect the drug resistance level. Our findings can help medical doctors to predict the efficacies of different drugs in NSCLC treatment.

5. CONCLUSION

The mutation position is very important for dimer formation and it affects the drug's binding strength with EGFR. Mutations such as L858R and T790M which do not happen on the protein interaction surface can hardly affect

the formation of dimers. However, the T790M mutation cause drug resistance due to the increase in the positive charge which can increase the affinity for ATP. Patients with the delE746_A750 mutation can obtain a good therapy by using gefitinib instead of erlotinib. By comparing the binding free energy to form a homo- or heterodimers, we find that the L858R mutant will incline to form a hetero-dimer rather than a homo-dimer.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is support by the Hong Kong Health and Medical Research Fund (Project 01121986) and the Hong Kong Research Grants Council (Project CityU 11200715)

REFERENCES

[1] Vansteenkiste J, De Ruysscher D, Eberhardt WE, *et al.* Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann oncol 2013; 24(Suppl 6): vi89-98.

[2] Shelton JG, Steelman LS, Abrams SL, *et al.* The epidermal growth factor receptor gene family as a target for therapeutic intervention in numerous cancers: what's genetics got to do with it? Expert Opin Ther Tar 2005; 9: 1009-30.

[3] Regales L, Gong YX, Shen RL, *et al.* Dual targeting of EGFR can overcome a major drug resistance mutation in mouse models of EGFR mutant lung cancer. J Clin Invest 2009; 119: 3000-10.

[4] del Viso F, Bhattacharya D, Kong Y, Gilchrist MJ, Khokha MK. Exon capture and bulk segregant analysis: rapid discovery of causative mutations using high-throughput sequencing. BMC Genomics 2012; 13: 649.

[5] Matthews JM, Sunde M. Dimers, Oligomers, Everywhere. Adv Exp Med Biol 2012; 747: 1-18.

[6] Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. Trends Biochem Sci 2004; 29: 618-25.

[7] Gonzalez-Magaldi M, Postigo R, de la Torre BG, *et al.* Mutations That Hamper Dimerization of Foot-and-Mouth Disease Virus 3A Protein Are Detrimental for Infectivity. J Virol 2012; 86: 11013-23.

[8] McHugh L, Arthur JW. Computational Methods for Protein Identification from Mass Spectrometry Data. Plos Comput Biol 2008; 4: e12.

[9] Yang TY, Wu JC, Yan CL, *et al.* Virtual screening using molecular simulations. Proteins 2011; 79: 1940-51.

[10] Butt HJ. Electrostatic Interaction In Atomic Force Microscopy. Biophys J 1991; 60: 777-85.

[11] Roth CM, Neal BL, Lenhoff AM. Van der Waals interactions involving proteins. Biophys J 1996; 70: 977-87.

[12] Bae JH, Boggon TJ, Tome F, Mandiyan V, Lax I, Schlessinger J. Asymmetric receptor contact is required for tyrosine auto-phosphorylation of fibroblast growth factor receptor in living cells. Proc Natl Acad Sci U S A 2010; 107: 2866-71.

[13] Comis RL. The current situation: Erlotinib (Tarceva (R)) and gefitinib (Iressa (R)) in non-small cell lung cancer. Oncologist 2005; 10: 467-70.

[14] Politi K, Zakowski MF, Fan PD, Schonfeld EA, Pao W, Varmus HE. Lung adenocarcinomas induced in mice by mutant EGF receptors found in human lung cancers respond to a tyrosine kinase inhibitor or to down-regulation of the receptors. Gene Dev 2006; 20: 1496-510.

[15] Riely GJ. The use of first-generation tyrosine kinase inhibitors in patients with NSCLC and somatic EGFR mutations. Lung Cancer 2008; 60: S19-S22.

[16] Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics 2002; Chapter 2: Unit 2 3.

[17] Goddard TD, Huang CC, Ferrin TE. Software extensions to UCSF Chimera for interactive visualization of large molecular assemblies. Structure 2005; 13: 473-82.

[18] Mark P, Nilsson L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J Phys Chem A 2001; 105: 9954-60.

[19] Li L, Li D, Chen H, Han JG. Studies on the binding modes of Lassa nucleoprotein complexed with m7GpppG and dTTP by molecular dynamic simulations and free energy calculations. J Biomol Struct Dyn 2013; 31: 299-315.

[20] Goette M, Stumpe MC, Ficner R, Grubmuller H. Molecular Determinants of Snurportin 1 Ligand Affinity and Structural Response upon Binding. Biophys J 2009; 97: 581-9.

[21] Gohlke H, Case DA. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. J Comput Chem 2004; 25: 238-50.

[22] Yarden Y, Schlessinger J. Self-Phosphorylation Of Epidermal Growth-Factor Receptor - Evidence for a Model Of Intermolecular Allosteric Activation. Biochemistry 1987; 26: 1434-42.

[23] DiMaio F. Advances in Rosetta structure prediction for difficult molecular-replacement problems. Acta Crystallogr D Biol Crystallogr 2013; 69: 2202-8.

[24] MacDonald JT, Maksimiak K, Sadowski MI, Taylor WR. De novo backbone scaffolds for protein design. Proteins 2010; 78: 1311-25.

[25] Mohanty S, Hansmann UHE. Folding of proteins with diverse folds. Biophys J 2006; 91: 3573-8.

[26] Hou TJ, Wang JM, Li YY, Wang W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. J Chem Inf Model 2011; 51: 69-82.

[27] Yun CH, Mengwasser KE, Toms AV, *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc Natl Acad Sci U S A 2008; 105: 2070-5.

[28] Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. Nucleic Acids Res 2005; 33: W331-W6.

[29] Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins 2004; 56: 143-56.

[30] Tiwari A, Panigrahi SK. HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures. In Silico Biol 2007; 7: 651-61.

[31] Sarkhel S, Desiraju GR. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. Proteins 2004; 54: 247-59.