8-2016

Molecular Mechanics Study of Protein Folding and Protein-Ligand Binding

Tingting Han
Clemson University

MOLECULAR MECHANICS STUDY OF PROTEIN FOLDING AND PROTEIN-LIGAND BINDING

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Chemistry

---

by
Tingting Han
August 2016

Accepted by:
Dr. Brian Dominy, Committee Chair
Dr. Dvora Perahia
Dr. Emil Alexov
Dr. Leah Casabianca

ABSTRACT

In this dissertation, molecular dynamics (MD) simulations were applied to study the effect of single point mutations on protein folding free energy and the protein-ligand binding in the bifunctional protein dihydrofolate reductase-thymidylate synthase (TS-DHFR) in plasmodium falciparum (pf). The main goal of current computational studies is to have a deeper understanding of factors related to protein folding stability and protein-ligand binding.

Chapter two aims to seek solutions for improving the accuracy of predicting changes of folding free energy upon single point mutations in proteins. While the importance of conformational sampling was adequately addressed, the diverse dielectric properties of proteins were also taken into consideration in this study. Through developing a three-dielectric-constant model and broadening conformational sampling, a method for predicting the effect of point mutations on protein folding free energy is described, and factors of affecting the prediction accuracy are addressed in this chapter.

The following two chapters focus on the binding process and domain-domain interactions in the bifunctional protein pfDHFR-TS. This protein usually plays as the target of antimalarial drugs, but the drug resistance in this protein has caused lots of problems. In chapter three, the mechanism of the development of drug resistance was investigated. This study indicated that the accumulation of mutations in pfDHFR caused obvious changes of conformation and interactions among residues in the binding pocket, which further weakened the binding affinity between pfDHFR and the inhibitor drug. Furthermore, the high rigidity and significantly weakened communications among key residues in the

protein binding pocket were exhibited in the pfDHFR quadruple mutant. The rigid binding site was associated with the failure of conformational reorganization upon the binding of pyrimethamine in the quadruple mutant. Chapter four investigated the effect of the N-terminus in pfDHFR-TS on enzyme activity and domain-domain communications. This is the first computational study that focuses on the full-length pfDHFR-TS dimer. This study provided computational evidence to support that remote mutations could disturb the interactions and conformations of the binding site through disrupting dynamic motions in pfDHFR-TS.

DEDICATION

To my family

ACKNOWLEDGMENTS

First and foremost, I would like to express my great appreciation to my advisor Dr. Brian Dominy for his support, valuable comments and continuous encouragement throughout my ph.D study. His way of mentoring guided me to develop as a chemist.

I also give my thanks to Dr. Steven Stuart and his research group for valuable discussions and suggestions to my research work. Special thanks to Dr. Dvora Perahia for her sincere encouragement during my ph.D study. I would like to acknowledge Dr. Emil Alexov and Dr. Leah Casabianca for their courses which helped me to understand computational biophysics in the perspective of theories and techniques.

I thank my labmates Vibhor Agrawal, Zhe Jia, Yinling Liu, and Richard Overstreet for the discussions, support, and friendship.

I am also very grateful for the opportunity to use the Palmetto cluster offered by Clemson Computing & Information Technology. Without this resource, I could not conduct my research.

TABLE OF CONTENTS

Table of Contents (Continued)

Page

LIST OF TABLES

LIST OF FIGURES

List of Figures (Continued)

Figure Page

CHAPTER ONE

INTRODUCTION

1.1 Overview of protein folding and protein ligand binding

The problem of protein folding[1] was first posed more than one half-century ago, when the Nobel Prize in Chemistry 1962 was awarded jointly to Max Ferdinad Perutz and John Cowdery Kendrew for their studies of the structures of globular proteins. Since then, scientists have raised the question of how to explain protein structures by physical principles[2, 3]. What are the driving forces for a protein to fold to its 3D-folded native structure from its denatured state[4]? Can scientists compute a protein's native structure from the amino acid sequence[5]? After decades of research, the mystery in protein folding starts to unfold. The major contributing forces which drive proteins to fold are hydrogen bonding, van der Waals interactions, electrostatic interactions, hydrophobic interactions and chain entropy. Despite the huge number of conformations between denatured and native structure, proteins can precisely fold by determined pathways and mechanisms, without sampling all possible conformations. The protein folding energy landscapes[6] are funnel-shaped, where low-energy conformational ensembles have fewer conformations. With a correctly folded structure, a protein can carry out its remarkable molecular functions.

Today, the field of protein folding could not be framed as the folding process any more, since old questions in the field of protein folding have generated even more new

questions, which are related to physics, chemistry, biology and medicine. In this field, there are so many problems to be solved, such as developing algorithms that can accurately predict stability of folded proteins, solving the mechanism of protein folding related diseases, and developing methods to accurately calculate binding affinity between small ligands and proteins[7-9].

Proteins often achieve their specific biological functions through direct interactions with other ligands, such as peptides, nucleic acids, substrates and drugs. Therefore, a prerequisite for understanding or modifying the cellular activities is to obtain a good knowledge of the mechanism of the protein-ligand binding process, including the local or non-local interactions, the conformational changes, and the energies which play as major driving forces for the formation of protein-ligand complexes[10]. Furthermore, in the field of drug discovery, understanding protein-ligand binding is essential for exploring the mechanism of drug resistance and providing guidance in the development of new drugs[11, 12].

1.2 Overview of the simulation methods

1.2.1 Molecular dynamics simulation

In order to reveal the mystery of protein folding and protein-ligand binding, computational approaches have been developed to explore these problems from the insight of the molecular level. Molecular dynamics (MD) is one of these computational approaches in reproducing the behavior of molecules' motion. The molecular dynamics simulation is

based on Newton's equation of motion, as described in equation (1.1), where the force exerted on each atom is essential to describe the physical system.

$$F_i = m\frac{d^2\boldsymbol{r}_i}{dt^2} \tag{1.1}$$

The acceleration of each atom could be determined from the force on each atoms or the position of each atoms. The force is usually evaluated by the potential energy $U(\boldsymbol{r}_1, \dots, \boldsymbol{r}_N)$ for *N* interacting atoms, where $\boldsymbol{r}_i$ represents the individual atom's position. The force acting on the *i*th atom could be written as equation (1.2):

$$F_i = -\frac{\partial U(\boldsymbol{r}_1, \dots, \boldsymbol{r}_N)}{\partial \boldsymbol{r}_i} \tag{1.2}$$

Therefore, it is necessary to find an accurate force field to describe the potential energy of the physical system. The CHARMM force field is one of the typical force fields used in MD simulations, as shown in equation (1.3)[13]:

$$\begin{aligned} U(\vec{r}) = &\sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{Urey-Bradley} K_{UB}(S-S_0)^2 \\ &+ \sum_{dihedrals} K_\varphi(1+\cos(n\varphi-\delta)) + \sum_{impropers} K_\omega(\omega-\omega_0)^2 \\ &+ \sum_{\substack{non-bonded \\ pairs}} \left\{ \varepsilon_{ij}^{min} \left[(\frac{R_{ij}^{min}}{r_{ij}})^{12} - 2(\frac{R_{ij}^{min}}{r_{ij}})^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\} \end{aligned} \tag{1.3}$$

The first five terms on the right hand side of the above equation describe the internal terms, which include bond length, bond angle, Urey-Bradly, dihedral angle, and improper angle. The internal terms, except for the dihedral angle term, are described in the harmonic form so that the molecules are in the correct chemical structure, and b_0, θ_0, S_0, ω_0 are values at

the equilibrium position. The last term represents the non-bond interactions between atom pairs (*i*, *j*), which exclude adjacent atoms with covalent bond and atom pairs with two covalent bonds in between. In this force field, the non-bond interactions include Coulombic electrostatic interactions and van der Waals interactions calculated by the Lennard-Jones 6-12 term. In the van der Waals term of the equation (1.3), R_{ij}^{min} is the distance where the potential reaches a minimum, ε_{ij}^{min} is the well depth, and r_{ij} is the distance between centers of the two atoms *i* and *j*. In the Coulombic potential energy term, $q_i \; and \; q_j$ is the partial charge of the two atoms *i* and *j*, respectively, ε is the relative dielectric constant, which is set to 1 in explicit solvent, and ε_0 is the electrical permittivity in space. Molecular dynamics simulations utilize the CHARMM force field are applied all through the three projects discussed below.

1.2.2 Implicit solvent simulation

In contrast to explicit solvent method, the implicit solvent method treats the solvent as continuous medium, thus it neglects the large number degree of freedom in explicit solvent method. Assume that the absolute values of the solvation free energy of ions of the same size and opposite charge are not identical[14], one common model for estimating solvation free energy ΔG_{solv} is:

$$\Delta G_{solv} = \Delta G_{polar} + \Delta G_{nonpolar} \qquad (1.4)$$

Equation (1.4) implies that the process of solvating molecules involves two steps: first, moving the solute to solvent with the removal of all charges ($\Delta G_{nonpolar}$); and second, transfer all the partial changes to the continuum solvent (ΔG_{polar})[15]. In one relatively

rigorous approach, $\Delta G_{nonpolar}$ includes generating a cavity in the solvent and inserting the solute into the cavity, which contains attractive dispersion and repulsion interaction between solute and solvent[16]. A common method for estimating $\Delta G_{nonpolar}$ is through calculating solvent accessible surface area (SASA), More details about calculating SASA are provided in Chapter 2.

In the implicit solvent framework, the polar solvation free energy (ΔG_{polar}) can be solved through the Poisson-Boltzmann model. When mobile ions are absent in solvent, the Poisson equation (PE)[17] for calculating electrostatic potential is:

$$\nabla[\varepsilon(\boldsymbol{r})\nabla\phi(\boldsymbol{r})] = -4\pi\,\rho(\boldsymbol{r}) \tag{1.5}$$

where $\phi(\boldsymbol{r})$ is the electrostatic potential, $\rho(\boldsymbol{r})$ is the charge density, $\varepsilon(\boldsymbol{r})$ is the position dependent dielectric constant. However, when the effect of salt is considered in the continuum solvent, equation (1.5) becomes more complicated. With the presence of mobile ion, the charge density is described as[18]

$$\rho(\boldsymbol{r}) = \rho_f(\boldsymbol{r}) + |e|\sum_j n_j\, z_j \exp\left(-\frac{\phi(\boldsymbol{r})|e|z_j}{kT}\right) \tag{1.6}$$

where n_j and z_j are the bulk density and ion charge, respectively. $|e|$ is the elementary charge, and $\rho_f(\boldsymbol{r})$ is the charge density for a set of fixed partial charges q_j at position $\boldsymbol{r}_j$ inside the dielectric boundary. We can obtain the non-linear Poisson-Boltzmann equation through substituting the charge distribution described in equation (1.6) into PE. If the exponential term in the above equation is linearized, then we can get the linear Poisson-Boltzmann (PB) equation:

$$\nabla[\varepsilon(\boldsymbol{r})\nabla\phi(\boldsymbol{r})] = -4\pi\rho_f(\boldsymbol{r}) + \kappa^2\varepsilon(\boldsymbol{r})\phi(\boldsymbol{r}) \tag{1.7}$$

The numerical solutions of PB equation are usually obtained through finite-difference method. Once the electrostatic potential is calculated, the polar solvation free energy is computed through equation (1.8). Here, $\phi(r_i)_{vac}$ is the electrostatic potential when ε equals to exterior dielectric constant.

$$\Delta G_{polar} = \frac{1}{2}\sum_i q_i[\phi(r_i) - \phi(r_i)_{vac}] \tag{1.8}$$

In the case of an ion with radius *a*, the equation can be reduced to the Born formula, as shown in equation (1.9):

$$\Delta G_{Born} = -\frac{q^2}{2a}\left(1 - \frac{1}{\varepsilon_{ext}}\right) \tag{1.9}$$

where ε_{ext} is the exterior dielectric constant. Thus, for a molecule consisting of *N* spherical atoms with radii a_i and charge q_i, if the distance between any two atoms is sufficiently larger than atom radii, the polar solvation free energy the generalized Born (GB) model is given by the summation of individual Born terms and pair-wise Coulombic terms[19]:

$$\Delta G_{polar} = \sum_{i=1}^{N}\frac{q_i^2}{2a_i}\left(\frac{1}{\varepsilon_{ext}} - 1\right) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{q_i q_j}{r_{ij}}\left(\frac{1}{\varepsilon_{ext}} - 1\right) \tag{1.10}$$

where r_{ij} is the distance between two different atoms *i* and *j*. However, atoms in real molecules are not spheres as mentioned above.

In order to capture the physics of PE for real molecule geometries, a function f_{GB} is introduced in the GB theory. Substituting f_{GB} to equation (1.10), the polar solvation free energy could be represented as equation (1.11):

$$\Delta G_{polar} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{GB}}\left(\frac{1}{\varepsilon_{ext}} - 1\right) \tag{1.11}$$

$$f_{GB} = \left[r_{ij}^2 + R_i\, R_j \exp(-\gamma r_i r_j / R_i\, R_j)\right]^{1/2} \tag{1.12}$$

$$R_i = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon_{ext}}\right)\frac{q_i^2}{\Delta G_{ii}^{polar}} \tag{1.13}$$

where $\gamma = 1/4$ is the most common form[20], R_i is the effective Born radii of the *i*th atom, and ΔG_{ii}^{polar} is obtained from the self-contribution of every atom in the molecule.

1.3 Overview of projects

In this dissertation, we aim to improve the accuracy of predicting the protein stability, which is necessary for understanding the relationship between protein structure and function, and designing new proteins. We are also interested in understanding the mechanism of drug resistance development to provide a guidance for the future drug discovery.

In chapter 1, we developed a novel model to improve the accuracy of predicting changes of protein folding free energy upon single point mutation. For more than 30 years, there has been a great focus on understanding the interactions dominating protein folding and the driving forces maintaining protein stability[21]. The major research interests include the contribution of hydrophobic effect, the polar or charge interactions, local interactions and non-local interactions. To answer these questions, chemists have generated site-directed mutations to identify the essential interactions and residues in maintaining protein structural stability. There are a lot of experimental data regarding the effect of site-directed

mutations on protein folding free energy[22, 23]. However, the accuracy of theoretical methods in predicting this effect still need to be improved[24]. Since protein conformations are constantly shifting from one to another, it is very difficult to capture all protein conformations in one computational study. The conformational sampling problem has been a major concern[25]. In this dissertation, we presented a method to compute changes of protein folding free energy upon site-directed single mutation with an attempt to broaden conformational sampling through generating diverse conformations and applying molecular dynamics simulations subsequently. On the other side, considering the diverse dielectric environments in proteins and, protein dielectric properties were also taken into consideration in this study. A method was developed here to predict the changes of protein folding free energy upon single point mutations through conformational sampling, the Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) approach, and the recognition of multiple protein dielectric environments.

In the following two chapters, different aspects of the bifunctional protein dihydrofolate reductase-thymidylate synthase in plasmodium falciparum (pfDHFR-TS) have been studied. The pfDHFR-TS plays an important role in the folate pathway, which is reducing dihydrofolate (DHF) to tetrahydrofolate (THF)[26]. Because THF is essential for purine, pyrimidine, and amino acid production, inhibiting the activity of pfDHFR can lead to the failure of DNA production or cell division. Therefore, pfDHFR has been a target for the treatment of malaria[27]. However, during the course of antimalaria drug treatment, mutations occur and lead to antimalarial resistance. In chapter 3, we provide an insight into the mechanism of pfDHFR resistance to pyrimethamine (Pyr) from a computational

approach. Despite the large amount of experimental studies regarding pfDHFR drug resistance[28, 29], quite little theoretical work is published to explain the mechanism of the drug resistance in pfDHFR[30]. We performed molecular dynamics (MD) simulations for wild type pfDHFR, C59R/S108N, N51I/C59R/S108N/I164L mutant, and explored how the accumulation of mutations affect binding affinity between pfDHFR-TS and Pyr. The results are consistent with the experimental data, and indicate that antimalarial resistance is related to significant conformation and flexibility changes upon mutations in *pf*DHFR. It is also found that the weakened communications among key residues may cause the failure of conformational reorganization upon the binding of Pyr, which lead to weak binding between pfDHFR and Pyr.

Compared to DHFR in eukaryotes, the structure of pfDHFR-TS is unique[26]. The unique structural features of this bifunctional protein, which include the junction region connecting the DHFR and TS domain, the N-terminal tail in DHFR domain, and the two extra inserts (residue 20 to 36 and residue 64 to 99) in the DHFR domain, may be important in the protein function. It is reported that the N-terminus, even though remote from pfDHFR active site, plays an important role in maintaining pfDHFR activity[31] and domain-domain communication in pfDHFR-TS[32]. A hypothesis was made in a previous experimental study that the N-terminal tail may influence the enzyme catalytic function through the interactions with the Insert II and the αβ loop (residue 141-184)[31]. However, this hypothesis is not tested yet. There are adequate studies focusing on studying mutations near or in the active site, but only quite a few studies are about the role of the unique N-terminal tail in pfDHFR-TS. In chapter 4, we provide the computational study of the role

of N-terminal tail, and reveal that the deletion of this tail could perturb the conformation of active site, which may be a reason for the decreased pfDHFR activity upon deleting the N-terminal tail. This is also the first computational study that focuses on the full-length pfDHFR-TS dimer to understand the domain-domain interactions. The N-terminal tail not only can modulate the pfDHFR activity, it also contributes to maintaining the communications among different domains in this bifunctional dimer protein.

In summary, the dissertation presented here provides a novel model to improve the accuracy of predicting changes of protein folding free energy upon single point mutations. It also brings a mechanism of the drug resistance development in pfDHFR-TS, which can help to guide the future discovery of new drugs in treating malaria. More progresses about the research has been discussed in the next chapters.

CHAPTER TWO

PREDICTING FOLDING FREE ENERGY CHANGES: A STRUCTURAL ENSEMBLE AND MOLECULAR DYNAMICS SIMULATION APPROACH

ABSTRACT

A method to calculate changes in the folding free energy upon single point mutations was presented here through an MM/PBSA approach that attempts to broaden conformational sampling through the generation of diverse "seed" conformations and subsequent molecular dynamics (MD) sampling in the phase space surrounding these seed conformations. This approach was applied to 150 mutants from 9 independent different proteins, and changes in protein stability upon mutation were calculated. The seed conformations of all mutants were generated using the program CONCOORD, and the changes in the folding free energy of each mutant were evaluated by the Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) method. The role of conformational sampling was investigated by comparing the methods of using single minimized structures, ensembles generated using CONCOORD, and ensembles generated using a combination of CONCOORD and MD. Recognizing the important role of electrostatics in protein stability, we also examined the dielectric properties of the proteins and their impact on model accuracy. Of the models being investigated, a three dielectric region model in which the dielectric constant increases from a value of 4 in the protein core

to 6 on the protein periphery showed the best agreement with experiment. The three dielectric constant model resulted in a correlation coefficient of 0.7 and a root mean square error of 2.09 kcal/mol between the computational and the experimental results. A subsequent analysis of the error within the model indicated significant challenges remain in the characterization of the electrostatic environment surrounding charged or polar residues.

2.1 Introduction

Single point mutations can affect protein stability and function, which are frequently related to human diseases, such as Alzheimer's disease and Rett syndrome. In order to better understand how protein stability will change upon single mutations, a number of computational estimation techniques have been developed to obtain information about protein dynamics that could not be easily gained through experimental techniques. Through these computational techniques, the changes in folding free energy of protein are calculated, and valuable information for protein stability study is provided.

One of the protein folding free energy prediction methods is based on detailed atomic models coupled to physical force fields[33,34,35,36,37]. Bash et al. [33] implemented the free energy perturbation (FEP) method in molecular simulation to estimate the trypsin stability upon a point mutation, and achieved good agreement with the experimental results. Such methods are based on precise physical models, but are computationally expensive and could not estimate contribution from each energy components. The statistical potential based technique is another method which is popularly used to study protein

stability[38,39,40,41] , and could successfully predict the change of protein thermal stability upon mutations. The free energy is computed by the linear combination of different statistical potential terms, such as distance potential, torsion potential, residue contact, dihedral angle, and solvent accessibility[39]. Statistical potentials are extracted from protein structure databases, and rely on chain length or composition, where true physical potentials do not. Similar to the statistical potential based methods, empirical energy functions rely on the combination of physical energy terms and empirical experimental data[42,43,44] , and involve weights that are fit to the experimental data. These methods are computationally efficient but do not provide accurate physical information of protein structure. Structure based methods, such as FoldX[45], Eris[46], CC/PBSA[47] and SAAFEC[48], while predicting the change of folding free energy upon mutations at different accuracy level, they are also capable of estimating the structural change upon single point mutations. However, applying multiple weighing factors make such models less physically rational.

It is desirable to develop a model that could both accurately and physically rationally predict protein folding free energy upon single point mutations. Physical potential approach (e.g., molecular dynamics), which simulates all atom force fields, is found physically precise. However, such approach is computationally expensive. Generally, a MD simulation of microseconds is required to reveal the folding reaction of a small protein, and a short simulation (less than 10 ns) will encounter unavoidable sampling problem[49]. Therefore, from the perspective of simulation efficiency and accuracy, it will be a robust choice to replace one long timescale simulation by several parallel short timescale simulations through sampling conformational space[50,51], which overcomes the

drawback of physical potential approach. The generation of diverse "seed" conformations not only conserves the computational resources, but also improves the probability of overcoming the multiple energy barriers. de Groot et.al [52] developed a sampling technique, named CONCOORD. In this technique, the structural coordinates are generated randomly, and then corrections are applied iteratively to search for structures that fulfill all predefined distance restrictions in CONCOORD. CONCOORD has been successfully applied as a sampling technique to predict the protein folding free energy changes upon single mutations in the CC/PBSA method[47]. However, those sampled structures donot agree with Boltzmann distribution. In addition, four weighing factors were applied in this model, so it is not completely physically rational. Thus, it is necessary to develop an approach that could follow both Boltzmann sampling and physical potential.

Here we applied the Molecular Mechanics/Poisson-Boltzmann Surface Area approach, which attempts to broaden conformational sampling through the generation of diverse "seed" conformations and subsequent molecular dynamics (MD) sampling in the phase space surrounding these seed conformations, to estimate the folding free energy of the wild type, the mutants, as well as the unfolded state. Structural ensembles were generated separately for the wild type and the mutants using both the program CONCOORD[24] and MD simulation. The combination of CONCOORD and molecular dynamics sampling could significantly increase the structural diversity in the ensembles. Comparing with the free energy perturbation method, while it may be less accurate, it's computationally more efficient with the capability to evaluating the contribution from each energy compoents. Comparing with statistical potential methods and structure based

methods even though less efficient, it does not need a large training set to fit different parameters for each energy term, or a number of weighing factors in evaluating the energy term.Thus, our approach provides a physically rational way to predict protein folding free energy change upon point mutations with comparable accuracy with other popular prediction methods. The dielectric constant was also an important parameter, since inner and outer regions of protein respond to the external electric field differently. In this work, models applying single, double, and three dielectric constants were evaluated. In the three dielectric constants model, the values of changes of protein folding free energy for 150 mutants from nine structurally unrelated proteins were calculated and compared against the experimental data. The correlation coefficient is found to be 0.70 and the standard deviation was 2.09 kcal/mol.

2.2 Methods

2.2.1 Input structures and Structure ensembles

Nine structurally unrelated proteins were used in this study. The initial coordinates of the wild type of these nine proteins were obtained from the crystal structure in the Protein Data Bank (PDB ID 1AYI, 1PGA, 1STN, 1YPC, 2LZM, 1VQB, 1CSP, 1APS, 4LYZ)[53,54,55,56,57,58,59,60,61]. Single point amino acid mutations were made to the wild type protein structure by applying the program MODELLER[62], and 150 mutant structures were generated. The structures of the nine wild type proteins and the 150 mutants were taken as the initial structures. These 150 mutants were randomly chosen from the 582 mutants in the paper by Benedix et al.[47] and the ProTherm database[63]. Proportions of different

mutation types within 150 mutants, including uncharged to uncharged mutations, uncharged to charged mutations, charged to uncharged mutations, and charged to charged mutations, are the same as those in the ProTherm database.

2.2.2 Structure ensembles

Diverse "seed" conformations of the 150 mutants' structures were generated by the program CONCOORD[52]. Twenty different conformations were independently generated for each protein. When using CONCOORD, the crystal structure from the protein data bank or the mutant structure from MODELLER is used as the reference structure. All of the pairwise interatomic distances, *d*, were measured for the reference structure, and then the upper and lower interatomic distances were clearly defined for all pairs of atoms. CONCOORD uses a set of parameters for the pairs involving interaction, but for the other pairs of atoms, the upper and lower distances were set to be $d \pm 1$ nm. In order to generate the ensemble of structures, a structure was generated with initial coordinates, where the coordinates were randomly chosen from a cube with edges of 2 nm. Then corrections of these coordinates were iteratively applied so that all the interatomic distances could be between the upper and lower distances. For each structure, 1000 iterations of corrections were performed.

2.2.3 Unfolded structures

The denatured state free energy is difficult to estimate, and here we are applying an idea which was discussed in Seeliger's paper [64] to generate the tripeptides GXG[65], where

X is any one of the twenty standard amino acids. In this GXG tripeptide, residue "X" is surrounded by glycine residues, only backbone atoms can affect internal motions of side chain of residue "X". When the protein is in unfolded state, ideally, there is no interaction among each side chains of amino acid or among different chains of protein, and only backbone interactions between adjacent residues are considered, thus the GXG model can be considered as an effective approximation for the unfolded state. The tripeptides, GXG, are first generated using CHARMM with the CHARMM27 parameters. Then the structural ensembles of tripeptides are obtained through the CONCOORD program. The free energy of each tripeptide is calculated by applying the same method as applied for calculating the folded state free energy, dielectric constant of 4 was applied for all GXG free energy calculation. For each GXG, all energies terms are averaged over all the CONCOORD structures.

2.2.4 Explicit solvent simulation

Conformational free energy differences were calculated for both the wild type and the mutant proteins using the molecular mechanics Poisson-Boltzmann surface area MM/PBSA method[66]. Before performing MM/PBSA analysis, all "seed" conformations of the wild type protein structures and the 150 mutants structures were undergone the process of explicit solvent simulation.

Before the energy minimization and MD simulation of the whole system, the structures of the proteins are prepared in three steps. First, all the coordinates of the CONCOORD structures were converted into the format that CHARMM could read using

the convpdb.pl script in the package MMTSB[67]. CHARMM version c35b6 was then used to place the hydrogen and build other missing atoms. Next, all structures were minimized applying harmonic restraints to hold atoms near a desired location. The first 500 steps of minimization were performed using the steepest descent (SD) algorithm, and there were the other 500 steps of minimization using the Adopted Basis Newton-Raphson (ABNR) algorithm. Then each protein was placed in a periodic cubic box containing TIP3P water with a density of 0.0334 moelcules/$Å^{-3}$($1g/cm^3$) and the water molecules overlapping with the solute were removed.

All energy minimizations and molecular dynamics simulations were carried out with the package CHARMM applying the CHARMM27 force field and the TIP3P explicit water model as the CHARMM force field is parameterized with respect to TIP3P water model. To ensure the correct usage of Particle Mesh Edward (PME), the system was neutralized by adding Na^+ cation, and Cl^- ion with a concentration of 0.15 M. After the system was solvated, the systems (including the protein solute and the explicit solvent) were first minimized for 500 steps using the steepest descent (SD) algorithm, followed by another 500 steps minimization using the Adopted Basis Newton-Raphson (ABNR) algorithm. During the minimization, harmonic restraints were applied with a force constant of 20 kcal/mol/$Å^2$, which could prevent the atoms in the protein from large motions. During the explicit solvent simulation, the dielectric constant is set to 1, which is corresponding to the permittivity of vacuum. The short-range non-bond interaction was described by Lennard-Jones potential with a cutoff of 12 Å using a switch function. PME was applied to calculate the electrostatic force during the simulation. Then the systems were heated

without applying harmonic restraints from 100 K to 300 K over a period of 20 ps using 1 fs time step, and equilibrated at 300 K for a further 1.6 ns using NPT ensemble.. The constant pressure was maintained at the pressure of 1 atm with the Langevin piston method. In the simulation, all covalent bonds involving hydrogen were constrained by the SHAKE command.

2.2.5 Minimized structure

In order to obtain minimized structure, the energy minimization of crystal structures and CONCOORD structures was performed by CHARMM with CHARMM27 force field as mentioned above. The harmonic restraints were applied during energy minimization, and the force constant was gradually reduced from 20 kcal/mol/$Å^2$ to 1 kcal/mol/$Å^2$. Under each force constant, 5000 steps of minimization was performed using SD algorithm, and then energy minimization with ABNR algorithm was performed until the energy change is less than or equal to 1.0E-9 kcal/mol.

2.2.6 MM/PBSA and Energetic analysis

The folding free energy change upon mutations could be represented by the thermodynamic cycle in Figure 2.1.

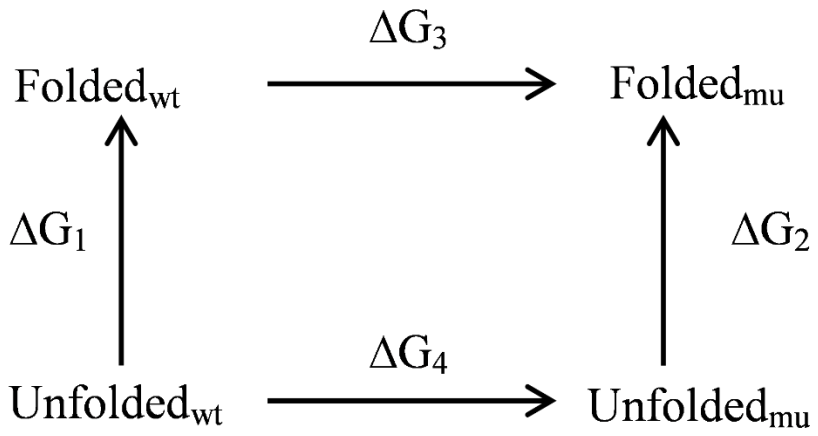

Figure 2.1: Thermodynamic cycle of protein folding upon point mutation

From the thermodynamic cycle, the folding free energy difference upon the mutations is calculated by $\Delta\Delta G = \Delta G_2 - \Delta G_1$, which is also equal to $\Delta G_3 - \Delta G_4$. Therefore, the folding free energy changes can be determined if ΔG_3 and ΔG_4 are obtained. The free energy function is given by equation (2.1) and (2.2).

$$\Delta G_{MMPBSA} = \Delta E_{MM} + \Delta G_{PB} + \Delta G_{SA} - T\Delta S \quad (2.1)$$

$$\Delta G_{MMPBSA} = \Delta E_{VDW} + \Delta E_{ELEC} + \Delta E_{INT} + \Delta G_{Solv}^{polar} + \Delta G_{SA}^{non-polar} - T\Delta S \quad (2.2)$$

Folded and unfolded free energy was calculated separately for the wild type and mutants using the MM/PBSA approach. The gas phase energies, including Van der Waals energy (ΔE_{VDW}), Coulombic energy (ΔE_{ELEC}) and other energy terms, such as bond energy, angle energy and so on, were calculated from the molecular mechanical energy function using charmm27 parameters without applying non-bond cutoff. The Poisson-Boltzmann polar solvation energy was calculated by solving the linear form of the PB equation applying the PBEQ module in the CHARMM package. The external dielectric constant

was set to 78. The internal dielectric constant used for proteins is different depending on the position of mutations. For all the 150 mutations in our study, dielectric constant of 4, 5 or 6 were applied to interior, partially exposed, or surface mutations respectively. The salt concentration was zero in this calculation. The non-polar part of the solvation energy was obtained by calculating the solvent accessible surface area (SASA) using a 1.4 Å radius probe: $\Delta G_{SA}^{non-polar} = \gamma SASA + b$, where the surface tension γ uses the value of 0.00542 kcal/mol/ Å^2, and the constant b adopts the value of 0.92 kcal/mol[68]. The folded protein energy was calculated with the snapshots extracted from the trajectory every 3 ps after the system was considered to achieve equilibrium. We assumed that the change of protein entropy upon single point mutation is quite small, and the entropy of wild type and mutants could be canceled out in equation (2.1) and (2.2). Each energy terms of chosen frames was calculated, and the final MMPBSA free energy was estimated based on ensemble averages of the associated energy terms.

$$\Delta G_{MM/PBSA} = \langle \Delta E_{MM} \rangle + \langle \Delta G_{PSOLV} \rangle + \langle \Delta G_{SA} \rangle \quad (2.3)$$

2.2.7 Three methods to address the significance of conformational sampling

In the first method, energy minimization was performed for the crystal structures of proteins, and then single-minimized structures were used to calculate the changes of folding free energy of each mutant. In the second method, 20 independent structures were sampled through the CONCORD package. All CONCOORD structures were minimized, and all energies of minimized structures were averaged to estimate protein stability. In the third method, CONCOORD were also applied to obtain 20 different protein conformations,

and all CONCOORD structures were subjected to energy minimization followed by explicit solvent molecular dynamics simulation, which was described in the method section. For each CONCOORD structure, 300 snapshots were extracted from the trajectory, and were used for the MM/PBSA calculation.

2.2.8 Data evaluation

The correlation between the computational results and the experimental data set[69,70,71,72,73,74,75] is evaluated through the Pearson linear regression correlation coefficient, and the equation is given by:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} \quad (2.4)$$

where x_i and y_i are the calculated data and experimental data separately, and n is the number of mutants.

The standard deviation (σ) between the computational data set and the experimental data set is calculated by the following equation (2.5):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (2.5)$$

where N is the number of mutants, x_i is the change of free energy obtained through computation, and μ is the mean of all the experimental value.

The RMS Error between the predicted value and the experimental value is shown in equation (2.6):

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{2.6}$$

where x_i and y_i are the calculated data and experimental data respectively, and n is the number of mutants.

The 95% confidence interval for the population correlation coefficient (ρ) was also evaluated. To obtain a confidence interval for ρ, we first calculated a 95% confidence interval for μ_V, where $\mu_V = \frac{1}{2}\ln[\frac{1+\rho}{1-\rho}]$. The interval for μ_V is

$$(v - \frac{z_{\alpha/2}}{\sqrt{n-3}}, v + \frac{z_{\alpha/2}}{\sqrt{n-3}}) \tag{2.7}$$

where $v = \frac{1}{2}\ln[\frac{1+r}{1-r}]$, and r is the sample correlation coefficient. This interval can yield a 95% confidence interval for ρ:

$$\left(\frac{e^{2c_1} - 1}{e^{2c_1} + 1}, \frac{e^{2c_2} - 1}{e^{2c_2} + 1}\right) \tag{2.8}$$

where $c_1$and c_2 are the left and right endpoints of the interval (2.7).

2.3 Results

2.3.1. Conformation sampling can improve the prediction accuracy

Three different methods were applied to evaluate the significance of conformational sampling in the prediction of folding free energy changes upon mutations. In all of the three methods, ε=5 is the optimal dielectric constant, where the RMSE is smaller comparing with the results applying any other value of dielectric constant in each method (more details are in section 3.2). The results of the three methods are shown in

Table 2.1. Through the comparison of the three approaches, the importance of applying conformational sampling is addressed. The combination of CONCOORD and MD sampling is chosen to investigate the effect of mutations on protein stability.

Table 2.1: Comparison of the three methods

Methods	RMSError (kcal/mol)	Correlation coefficient (95% confidence interval)*
Minimize/single	17.0±1.2	0.01(-0.15, 0.16)
CONCOORD/minimize	8.1±0.6	0.41(0.27, 0.54)
CONCOORD /MD	3.6±0.3	0.57(0.45, 0.67)

*: The value in parentheses is the 95% confidence interval of correlation coefficient.

2.3.1.1 The effect of CONCOORD sampling

The results in Table 2.1 showed that among the three methods, the method applying the single minimized structure produced the largest RMSE and smallest correlation between predictions and experimental results. When the sampling technique CONCOORD was applied, the root mean square error (RMSE) was greatly reduced and the correlation between the estimation and experimental data was also improved.

It is reported that there are multiple minima existing in proteins[76], therefore, the single-minimized structure of the wild type and the mutant may not correspond to the same minima, which make it difficult to correctly predict protein stability. In the CONCOORD/minimized method, we then analyzed the RMSD between each CONCOORD structures for the same protein, as well as the RMSD between the minimized structures of each CONCOORD conformations for the same protein. We found that the RMSD of the later is quite close to the former one for every CONCOORD structure, and

the difference is less than 0.1Å. The above observations indicated that the initial CONCOORD structures corresponded to different minima in the same protein. Therefore the application of CONCOORD sampling enabled the consideration of several conformations corresponding to different minima of the protein, and the unweighted averaged energies over several minima could more accurately predict the effect of single mutation on protein stability than the single-minimized structure did. Furthermore, the ensemble of states in protein is accordant with Boltzmann distribution, and proteins constantly shift from one conformation in one conformational (thermodynamic) state to another. In order to take into consideration the large number of conformational states, the averaged energies over several states could better reflect the changes in protein folding free energy.

2.3.1.2 The effect of molecular dynamics sampling

Comparing the second method with the third method, when the MD simulation was performed for each CONCOORD structure, the results were further improved. This result implies that applying the ensemble of conformations generated from MD simulation could predict the effect of single point mutations on protein folding free energy more accurately than just using the minimized structure.

It is worthy to notice that the protein behavior is dynamic. It's well known that the protein folding process is corresponding to an energy landscape, and the concept of energy landscape could also be applied to the folded state protein[77,78]. When the temperature is higher than glass transition temperature of proteins, the anharmonic motion of protein is

increased, which indicates that the protein is not trapped in a single energy well any longer. Only the multidimensional energy landscape could completely describe proteins, thus the minimized structure can merely represent one conformation of protein, and the dynamic properties of protein could not be reflected when only considering the minimized structures. Typically, even nanoseconds of MD simulation could not reach all the conformations of protein, since proteins usually have to take more than microseconds to move from one energy valley to another one with energy barriers of several kT[78]., Therefore, the combination of CONCOORD and MD can reveal a more accurate picture of protein dynamic motion.

2.3.2. Dielectric constant affects prediction accuracy

The results in Figure 2.2 indicate that the dielectric constant plays an important role in accurately predicting protein stability. In this section, we will present how the heterogeneous dielectric property of protein relates to the accuracy of predicting changes of protein folding free energy. The CONCOORD/MD method is applied on the analysis of the 150 mutants.

2.3.2.1 The single dielectric constant model

If we assume the protein is homogeneous, and only one dielectric constant is applied for the whole protein, the optimal dielectric constant in the single dielectric constant model is 5, where RMSE is the smallest and the correlation is the largest among the results applying dielectric constant from 1 to 70(Figure 2.2), with RMSE = 3.6±0.3

kcal/mol, and correlation R=0.57±0.015 (p-value < 0.0001), with a confidence interval (0.45, 0.67).

Coulomb energy and polar solvation free energy are anti-correlated with each other. When the dielectric constant is small enough, changes in protein folding free energy caused by mutations ($\Delta\Delta G_{MM/PBSA}$) will be dominated by the Coulomb energy, however, as the dielectric constant increasing, the Coulombic energy term and polar solvation free energy term begins to cancel out. Potential energy, other than Coulombic energy, is not affected by electric the force field, and is stable. Therefore, there is a minimum in Figure 2.2. When the dielectric constant is greater than 30, the RMSE is almost steady. The reason is that when the model applys a quite large dielectric constant, the Coulomb energy and the polar solvation free energy will disappear, and only the van der Waals potentials and the bonded energy terms, such as bond potential, angle potential, and dihedral potential, contribute to the protein folding free energy. Due to the importance of electrostatic energy terms over other energy terms in characterizing the effect of mutations on the stability of protein folding, the model with a large dielectric constant is a bad one.

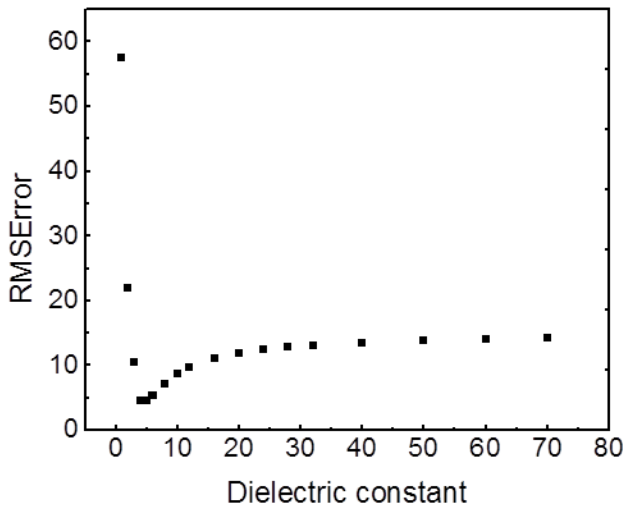

Figure 2.2: The RMSE between prediction and experiment results at different RMSE.

Figure 2.3 shows the correlation between the calculation data and the experimental results in the single dielectric constant model. If the deviation of calculated result from the regression line is greater than 2σ (σ=RMSE), this mutant is defined as an outlier. Based on this measure, there are 8 outliers in the single dielectric constant model (Figure 2.3). When the 8 outliers are excluded, the correlation rises to 0.65±0.010, with a 95% confidence interval of (0.54, 0.74), and the RMSE drops to 2.8±0.2 kcal/mol.

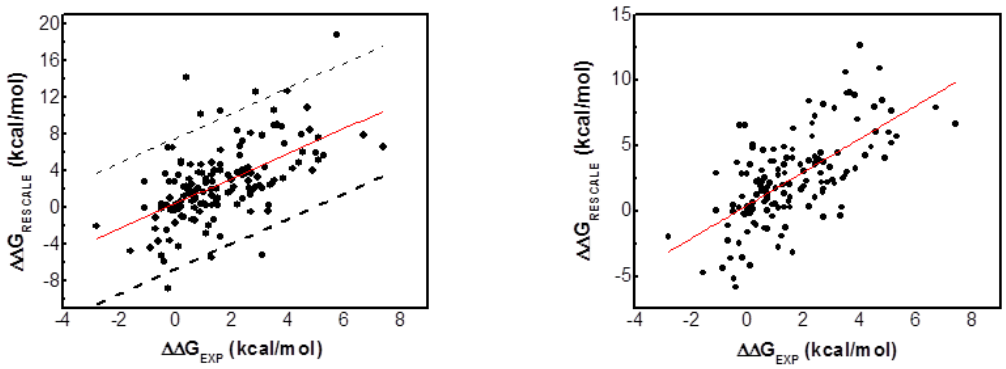

Figure 2.3: Calculated data compared with experimental data for 150 mutants applying the single dielectric model. In the left figure, the continuous line is the linear regression with an equation y=1.36x+0.47. The dash line is used to detect outliers, R=0.57, and σ=3.56 kcal/mol. In the right figure, the continuous line is the linear regression after discarding 8 outliers with an equation y=1.27x+0.45, R=0.65, and σ=2.73 kcal/mol.

Table 2.2 shows the analysis of the outliers in the single dielectric constant model. One explanation for some outliers is that we improperly assume the polarizability of each site is uniform all over the protein, which makes the single dielectric model unsuitable to estimate changes of the protein stability upon each point mutation. The importance of a proper electrostatic representation is further supported by the fact that 5 of the 8 outliers are charged residue related mutations. The mutation 1vqbL32R occurred in the hydrophobic core, and the residue Arg32 is even more destabilized by Lys46 through the repulsion effect in the mutant, and this might cause the overestimation of the coulomb energy. The similar situation happens to the mutation 1vqbY26R. This mutation also

involves a repulsion effect between Lys24 and Arg26 in the mutant, and changes of coulomb energy is overestimated. The mutation 1cspN10D leads to the formation of the salt bridge Asp10-Lys13 on the surface of proteins. The desolvation penalty of forming this salt bridge is underestimated, and thus the stabilizing effect of this mutation in the calculation is larger than experiment. Except for the charged site mutations, there are another two mutations 2lzmI58Y, which are highly buried in the hydrophobic core, and involved in large size change of the side chain. These mutations have caused movement of atoms and residues with respect to one another around the mutating position, which can also affect the electrostatic interactions. In this case, the reorganization of the local structure makes the changes of folding free energy difficult to be predicted.

Table 2.2: Analysis of the outliers in the single dielectric constant model

Mutations	$\Delta\Delta G_{EXP}$ (kcal/mol)	$\Delta\Delta G_{CALC}$ (kcal/mol)	Most likely explanation for outliers
1stnV39T	1.30	-5.42	Overestimation of Coulomb energy or underestimation of the solvation penalty
1pgaT53D	0.90	10.20	Improper dielectric parameter; underestimation of desolvation penalty of forming salt bridge
2lzmW126R	5.74	18.83	improper dielectric parameter
2lzmW138Y	2.87	12.64	Overestimation of the Coulomb energy
1vqbL32R	1.60	10.59	Highly buried mutation; improper dielectric parameter; overestimation of the Coulomb energy
1vqbY26R	0.40	14.17	overestimation of the Coulomb energy
1cspN10D	-0.26	-8.79	Improper dielectric parameter; underestimation of desolvation penalty of forming salt bridge
2lzmI58Y	3.11	-5.19	Difficult to predict structural reorganization

It is necessary to realize that the dielectric constant is not a universal constant[79], and instead differs in the different regions inside the protein[79,80]. Dielectric constant is the

ratio of the permittivity of the material to the permittivity of the vacuum, and it reflects the capability of dielectric polarizes being affected by external electric field. The dielectric constant of protein measures the protein polarizability. It is a parameter in models such as the one used here, but it is not a real constant, and its value may differ in the different regions inside the protein[79,80]. For example, the dielectric constant for the interior of the proteins is different from that of region on the surface. Simonson et.al[81,82] showed that the protein dielectric constant could vary from 2 in the interior of proteins to 13-30 in the outer region of proteins, and the flexible charged protein side chains at protein surface are associated with the large dielectric constant in the outer part. Therefore, in the following sections we will improve the prediction accuracy of the model by applying different dielectric constants based on the position of the mutations.

2.3.2.2 Double dielectric constants model

The classification is based on the solvent accessible surface area (SASA) and the Born radii of alpha carbon of residues. In the double dielectric constants model, The surface mutations are defined as: (1) any residues whose Born radius (R) of alpha carbon (C_α) is smaller than 4.0Å; (2) any residues with 4.0Å < $R_{C\alpha}$ < 5.0Å and percentage of solvent accessible surface area (SASA) of the residue is greater than 15%; (3) charged residues whose the average Born radius of the heavy atoms in side chain is smaller than 3.5Å. The rest of the mutations are buried mutations.

Based on the categorization methods, 87 mutants in the current dataset are surface mutations, and 63 mutants are interior mutations. To determine the optimal dielectric

constants for both surface mutations and interior mutations, RMSE between estimated value and experimental results were calculated for both categories at different dielectric constant, which is shown in Table 2.3.

Table 2.3 The RMSE of prediction results compared against experimental data in the double dielectric constants model

Dielectric constant	Buried (discard 2 outliers)*	exposed(discard 5 outliers)
ε=1	37.2±4.4 (37.2±4.4)	68.2±7.1 (57.7±6.2)
ε=2	12.9±1.6 (12.6±1.5)	26.0±2.7 (22.6±2.4)
ε=3	5.6±0.7 (5.2±0.6)	12.2±1.3 (11.0±1.7)
ε=4	**2.5**±0.3 **(1.9**±0.2**)**	5.6±0.6 (5.4±0.6)
ε=5	3.8±0.4 (3.6±0.4)	**3.3**±0.4 (2.8±0.3)
ε=6	4.5±0.6 (4.9±0.6)	4.0±0.4 (**2.4**±0.3)
ε=8	7.0±0.8 (7.1±0.8)	6.8±0.7 (4.6±0.5)
ε=10	8.1±1.0 (8.2±1.0)	8.7±0.9 (6.2±0.7)
ε=20	10.3±1.2 (10.5±1.2)	12.7±1.4 (9.5±1.0)
ε=40	11.5±1.4 (11.7±1.4)	14.8±1.6 (11.2±1.2)
ε=50	11.7±1.4 (11.9±1.4)	15.2±1.6 (11.6±1.2)
ε=60	11.9±1.4 (12.1±1.4)	15.4±1.6 (11.8±1.3)

*: The value in parentheses are the RMSE after discarding outliers.

It shows that ε=4 is the optimal dielectric constant for interior mutations, while 6 is the optimal dielectric constant for the surface mutations. Thus ε=4 and ε=6 is applied for mutations occurring in the buried or exposed regions, respectively. The smaller dielectric constant of the inner region and the larger dielectric constant of the outer region are also consistent with the previous studies[81-84], where they directly calculate the dielectric constant of some proteins (see section 6.2). The interior of a protein is much more hydrophobic and far less polarizable than the protein surface, thus, the dielectric constant in the inner region of proteins is lower than the surface region. The overall correlation between calculation and observation results is 0.45±0.025, with a 95% confidence interval

of (0.31, 0.57), and the RMSE is 3.4±0.3 kcal/mol for all 150 mutants in the current data set. Applying the outlier definition from previous section, there are 7 outliers (2 outliers in the buried region, and 5 outliers in the exposed region) in this double dielectric constants model. With the exclusion of outliers, the correlation is improved to 0.67±0.009 (p-value<0.0001), the 95% confidence interval for the population correlation coefficient of the mutations is (0.57, 0.75), and RMSE=2.3±0.2 kcal/mol, which is shown in figure 2.4.

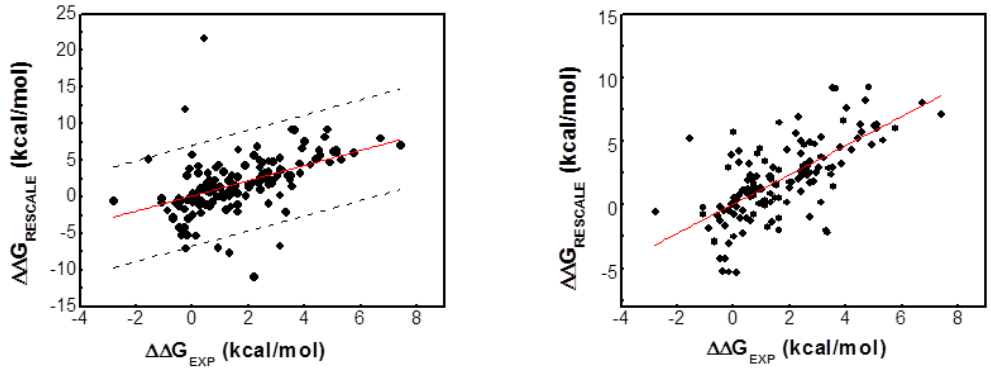

Figure 2.4: Calculated data compared against experimental data for 150 mutants applying the double dielectric constants model. In the left figure, the continuous line is the linear regression with an equation y=1.04x+0.16, R=0.45, σ=3.44 kcal/mol. The dash line is used to detect outliers. In the right figure, the continuous line is the linear regression after discarding 7 outliers with an equation y=1.15x+0.07, R=0.67, and σ=2.27 kcal/mol.

Comparing with the single dielectric constant model, the correlation between the prediction and the experimental data increases in the two dielectric constants model. There is also a significant decrease in the RMSE with the upper limit of RMSE in the two

dielectric constant model smaller than the lower limit of the single dielectric constant model.

Three of the outliers in the two dielectric constant model (1cspN10D, 1vqbY26R, 2lzmI58Y) also appeared in the single dielectric model, and the explanation about these outliers is discussed above. The other four outliers in current model are 1stnR87A, 1stnR35G, 1stnV39T, and 4lyzD101R, and they are most likely caused by the overestimation of coulomb energy or the underestimation of the solvation penalty. Also, the value of the electrostatic energy is related to the dielectric constant, and an improper dielectric constant used for predicting the changes of folding free energy upon point mutations may also cause large discrepancy from the experimental results.

2.3.2.3 Three dielectric constants model

When applying this three dielectric constants model, the RMSE is further decreased relative to the two dielectric constant model. In this model, the exposed residues mentioned in the last section were further divided into two groups: exposed and partially exposed amino acids. As a result, 51 mutations are fully exposed, 36 mutations are partially exposed, and 63 mutations are in the interior of proteins. The RMSE between prediction and experimental data was calculated for the buried, partially exposed and exposed region, respectively. The optimal dielectric constants for each region is determined from the value of RMSE, and the RMSE is the smallest at the optimal dielectric constant. The optimal dielectric constants for buried, partially exposed, and exposed mutations are 4, 5, and 6, respectively. The results of the RMSE calculation are shown in Table 2.4.

Using the three dielectric constants model, the predicted results correlated with the experimental results with R=0.49±0.02, with a 95% confidence interval of (0.36, 0.60) and RMSE=3.2±0.3 kcal/mol. There are five outliers (2 outliers are buried mutations, and 3 outliers are exposed mutations) in this model. After discarding these five outliers, the correlation coefficient rises to 0.69±0.008 (p-value<0.0001), and RMSE=2.1±0.2 kcal/mol. The results are shown in Figure 2.5. In the three dielectric constant model, the 95% confidence interval for the population correlation coefficient of the mutations is (0.59, 0.77). The current model produces a higher correlation coefficient between predicted results and experimental results than the double dielectric constant models, with R increasing from 0.67 to 0.69.

Because the above results were obtained through the rescaling results using equation (2.7) and (2.8), MM/PBSA calculation were performed in the following step to test the consistence between the rescaling results and the MM/PBSA calculation results.

$$\frac{G_{COUL}^{a}}{G_{COUL}^{b}} = \frac{\varepsilon_P^b}{\varepsilon_P^a} \tag{2.7}$$

$$\frac{G_{SOLV}^{a}}{G_{SOLV}^{b}} = \left(\frac{1}{\varepsilon_P^a} - \frac{1}{\varepsilon_w}\right) / \left(\frac{1}{\varepsilon_P^b} - \frac{1}{\varepsilon_w}\right) \tag{2.8}$$

Table 2.4 RMSE of prediction results compared against experimental data in the three dielectric constants model

Three categories	Buried (discard 2 outliers)*	Partially exposed	Exposed (discard 3 outliers)
ε=1	37.2±4.4 (37.2±4.4)	73.2±11.4	65.2±9.0 (59.5±8.1)
ε=2	12.9±1.6 (12.6±1.5)	27.6±4.3	25.2±3.4 (23.7±3.2)
ε=3	5.6± 0.7(5.2±0.6)	12.6±2.0	12.1±1.6 (11.8±1.6)
ε=4	**2.5**±0.3 **(1.9**±0.2**)**	4.8±0.8	6.2±0.8 (6.1±0.8)
ε=5	3.8±0.4 (3.6±0.4)	**2.5**±0.4	**3.8**±0.5 (3.0±0.4)
ε=6	5.0±0.6 (4.9±0.6)	3.7±0.6	4.1±0.6 (**2.1**±0.3)
ε=8	7.0±0.8 (7.1±0.8)	7.3±1.2	6.4±0.9 (4.0±0.6)
ε=10	8.1±1.0 (8.2±1.0)	9.5±1.5	8.1±1.1 (5.6±0.8)
ε=20	10.4±1.2 (10.5±1.2)	13.9±2.2	11.8±1.6 (9.0±1.2)
ε=40	11.5±1.4 (11.7±1.4)	16.1±2.5	13.8±1.9 (10.7±1.5)
ε=50	11.7±1.4 (11.9±1.4)	16.6±2.6	14.1±2.0 (11.1±1.5)
ε=60	11.9±1.4 (12.1±1.4)	16.9±2.7	14.4±2.0 (11.3±1.6)

*: The value in parentheses is the RMSE after discarding outliers.

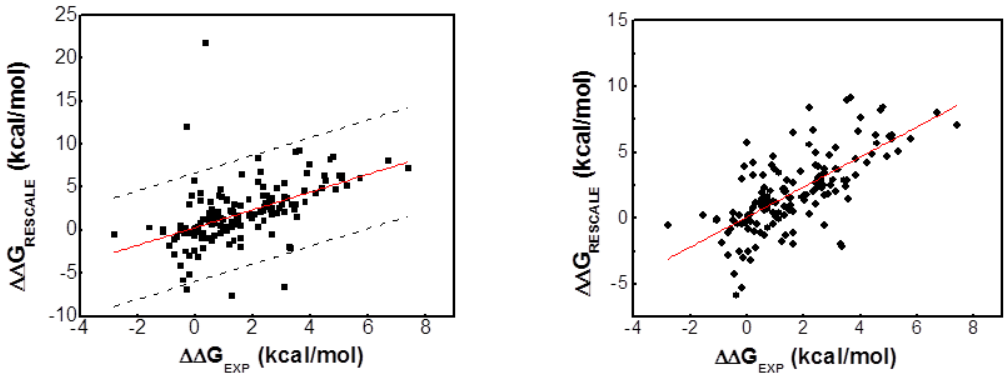

Figure 2.5: Calculated data compared against experimental data for 150 mutants applying the three dielectric constants model. In the left figure, the continuous line is the linear regression with an equation y=1.03x+0.33, R=0.49, and σ=3.15 kcal/mol. The dash line is used to detect outliers. In the right figure, the continuous line is the linear regression after discarding 5 outliers with an equation y=1.14x+0.11, R=0.69, σ=2.14 kcal/mol.

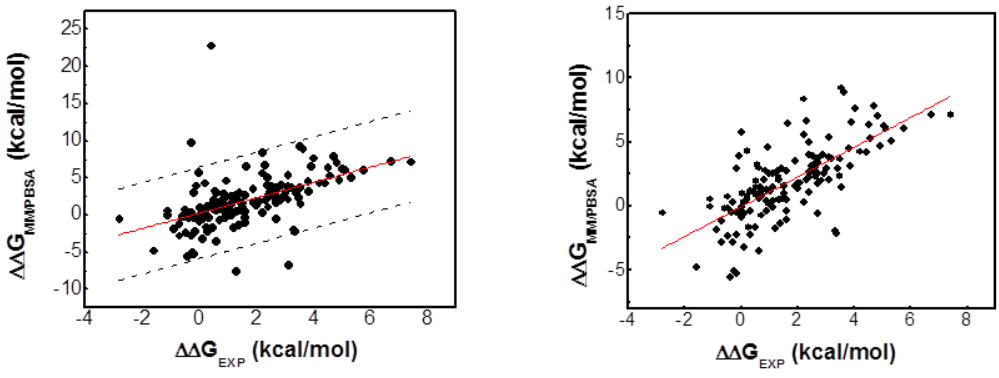

Figure 2.6: Calculated data compared against experimental data for 150 mutants applying the three dielectric constants model. In the left figure, the continuous line is the linear regression with an equation y=1.03x+0.25, R=0.50, and σ=3.07 kcal/mol. The dash line is used to detect outliers. In the right figure, the continuous line is the linear regression after discarding 4 outliers with an equation y=1.16x-0.05, R=0.70, and σ=2.09 kcal/mol.

Using the optimal dielectric constants which are mentioned above, the changes of folding free energy upon all 150 single point mutations were then calculated using the MM/PBSA method. The results are consistent with the rescaling results, as shown in Figure 2.6. The overall correlation coefficient is 0.50±0.020, with a 95% confidence interval of (0.37, 0.61), and RMS Error is 3.1±0.2 kcal/mol. After deleting 4 outliers, the RMSE is 2.1±0.2 kcal/mol, and correlation coefficient R is 0.70±0.007 with the 95% confidence interval for the population correlation coefficient of the mutations of (0.61, 0.77).

2.4. Discussions

2.4.1. Justification of this model

2.4.1.1 MM/PBSA

The MM/PBSA method combines explicit solvent simulation with continuous solvent model. Since only the final states are considered (unfolded and folded) and explicit solvent is removed, one major advantage of this method over the pathway method is its calculation speed, which makes it a suitable method for a macromolecule system, such as protein. The MM/PBSA method is founded on the statistical thermodynamics basis[85]. All solute-solute interactions are considered in this calculation, because no cutoff is applied[86]. Due to the cancelation of variance of polar solvation free energy and the coulomb energy, the MM/PBSA method could produce relatively stable total energy. The main challenge of the MM/PBSA method is the difficulty in accurately estimating the solute entropy[87]. Despite of the limitations, the MM/PBSA method could still achieve satisfactory accuracy in many cases compared with the pathway method and the experimental results. Srinivasan et.al applied this continuum solvent model and obtained good qualitative agreement with the experimental results[86]. Brice and Dominy calculated the free energy difference between A-form and B-form of DNA through MM/PBSA calculation, and the results achieved close agreement with an umbrella sampling approach[88]. Similarly, through the comparison of MM/PBSA and free energy perturbation (FEP) or thermodynamic integration (TI) method[89, 90,91] , it is demonstrated that the accuracy of MM/PBSA method is comparable with that of the FEP in calculating bisadamantyl-phosphate complex association free

energy, and is also comparable with the accuracy of the TI method in predicting protein-protein binding free energy changes upon alanine mutations.

2.4.1.2 Protein dielectric constant

The interior of a protein is a hydrophobic core, and has fewer charged atoms than the outer region of a protein. Thus, the innermost of a protein is much less polarizable than the surface region[84]. Furthermore, the surface of a proteins is "liquid-like", while the interior of a protein is "solid-like"[92], and the surface residues have more mobility than the buried groups in the core. Due to the lower flexibility and the less polarizable property, the interior of a protein is less capable to respond to the local electrostatic force field than the outer region of a protein. Therefore, the interior of a protein is corresponding to lower dielectric constant than that of the surface of a protein. Simonson et.al[81] have developed methods to evaluate the protein dielectric. If the protein is assumed to be a three-medium case, the dielectric constant can be determined based on equations (2.9), (2.10) and (2.11).

$$\langle \Delta M^2 \rangle = \sum_{ij} q_i q_j \langle \delta \boldsymbol{u}_i \delta \boldsymbol{u}_j \rangle \tag{2.9}$$

$$\frac{\langle \Delta M_{lf}^2 \rangle}{kTr_1^3} = \frac{f(\varepsilon_1, \varepsilon_2, \varepsilon_3)(\varepsilon_1 - 1) - f\left(\varepsilon_1^{hf}, \varepsilon_2^{hf}, \varepsilon_3^{hf}\right)\left(\varepsilon_1^{hf} - 1\right)}{f\left(\varepsilon_1^{hf}, \varepsilon_2, \varepsilon_3\right)} \tag{2.10}$$

$$f(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \frac{9\varepsilon_2\varepsilon_3}{(\varepsilon_1 + 2\varepsilon_2)(\varepsilon_2 + 2\varepsilon_3) - 2\left(\frac{r_1}{r_2}\right)^3 (\varepsilon_3 - \varepsilon_2)(\varepsilon_1 - \varepsilon_2)} \tag{2.11}$$

The protein is assumed to be made of 3 different regions, with dielectric constant of $\varepsilon_1, \varepsilon_2, \varepsilon_3$, and r_i is the distance from the measured region to the center of the protein

($r_1 < r_2 < r_3$). Superscripts *"lf"* stands for low frequency and *"hf"* stands for high frequency. ΔM is the deviation of the dipole moment from its mean, and is determined by the correlations between all pairs of protein atoms, q_i is the partial charge of atom *I*, and $\delta \boldsymbol{u}_i$ is its instantaneous displacement from its mean position. Their results indicated that the innermost region always showed lower dielectric constant than the outer regions. In Simonson's calculation[81], if ferro- and ferricytochrom *c* was viewed as a homogeneous one, estimated value of dielectric constant vary from 16 to 37. However, if the charged portions of the charged side chains, which were mainly at the interface of protein and solvent, were considered as the solvent medium, the calculated dielectric constants were 4.7±1.0 and 3.4±1.0 for ferro- and ferricytochrom *c* respectively, which is consistent with the experimental measurement. Furthermore, the calculated results indicated that the innermost region, which was within the distance of 10-11Å away from protein's center, were corresponding to even lower dielectric constant with a value of 1.5-2.0. Therefore, Simonson's work claimed that the dielectric constant in the inner region of protein is lower than that of the surface of protein. Simonson and Brooks[82] further applied Frohilch-Kirkwood theory of dielectric to study the dielectric property of 4 different proteins, their results also showed that the core in general had a lower dielectric constant than the protein surface.

2.4.1.3 Entropy

A single point mutation in the wild type usually does not perturb the structure of native or denatured states. The three dimensional structure of mutants is identical to that of

the wild type protein[93,94]. Therefore, it is reasonable to not take into account the changes of entropy upon single point mutation. To understand in detail how the single mutation affects entropy of proteins, translational entropy, vibrational entropy and rotational entropy will be considered separately. The equation for calculating translational/rotational entropy is given by equation (2.12).

$$S_{tr}^{0} = S_{t}^{0} + S_{r}^{0} = \left[2.5R - Rln\frac{N_j}{V}\Lambda^3\right] + \left[1.5R + Rln\pi^{1/2}\left(\frac{8\pi^2 kT}{h^2}\right)^{2/3} det\left(A_j\right)^{1/2}\right] \quad (2.12)$$

The first part of Eq. (2.12) is the translational entropy, which depends on molecular weight, through Λ^3. $\frac{N_j}{V}$ is the protein solution concentration, which is a constant. R is the gas constant, and T is temperature. Therefore, S_t^0 only depends on molecular weight of protein. Since a single mutation has minimum influence on the molecular weight of protein, the change of translational entropy upon a single mutation can be ignored. The second part of Eq. (2.12) is the rotational entropy. The term S_r^0 depends on molecular structure, which is represented by $det\left(A_j\right)^{1/2}$. Because a single point mutation in the wild type protein couldn't perturb the structure of folded and unfolded protein, rotational entropy also has negligible contribution to the changes of protein folding free energy upon single point mutation. Doig et al.[95] suggested that 25% of protein side chain entropy is lost upon protein folding, but the vibrational entropy is retained during the protein folding process. Therefore, changes in vibrational entropy upon protein folding are often negligible, and it is reasonable to ignore changes of vibrational entropy upon single point mutation.

2.4.1.4 Conformations generated from CONCOORD and MD simulation

The ensemble of conformations generated from CONCOORD sampling and MD simulation is determined here, and whether this ensemble is based on statistical thermodynamic is one important problem that we concern. Here, we studied the distribution of $G_{MM/PBSA}$ over the trajectories in the folded state for Staphylococcal nuclease (PDB ID: 1STN). Figure 2.7 is the histogram for the free energy distribution for ensemble trajectories. This result indicated that the free energy distribution of snapshots followed Gaussian distributions, and the sampling level here is well converged.

Peter V. Coveney and coworkers[96,97,98] have demonstrated that the ensemble of simulations are more efficient in sampling configurational space than a single long trajectory. Their research reveals that the binding free energy of each snapshot, which is calculated from MM/PBSA, fits quite well to the Gaussian distribution, and the sampling from the ensemble simulations is better converged than that from a single long trajectory. This Gaussian distribution indicates that particular free energy falling into this mathematical category satisfies the central limit theorem, which means that conformations being sampled are not correlated and the sampling is adequate.

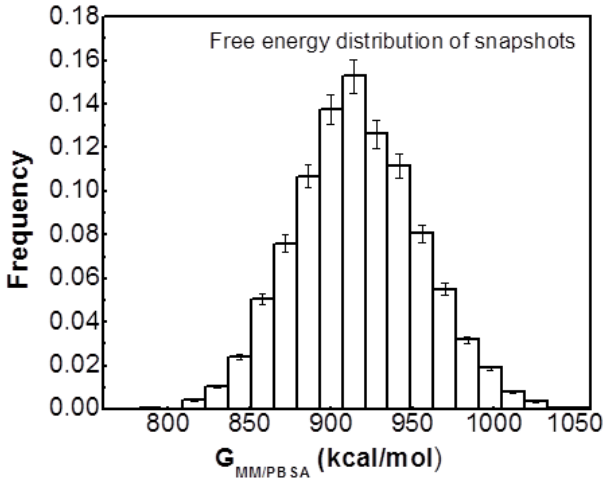

Figure 2.7: Histogram for free energy of snapshots for folded proteins in ensemble simulations

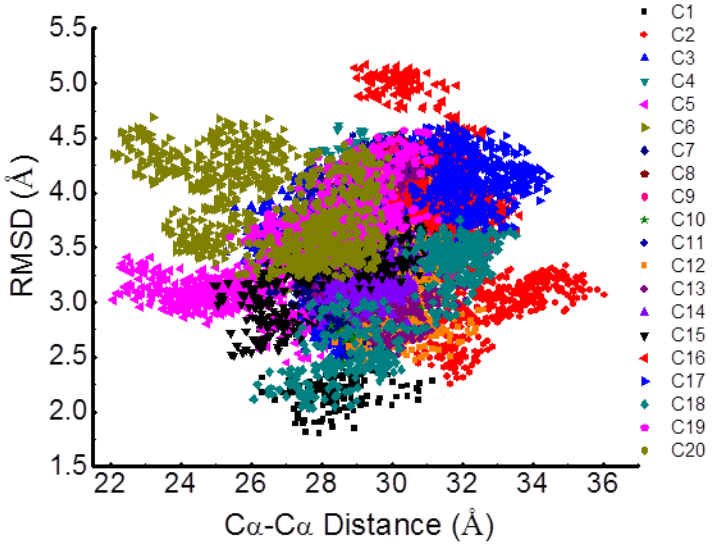

Figure 2.8: RMSD plotted against Cα-Cα distance for 20 individual MD trajectories

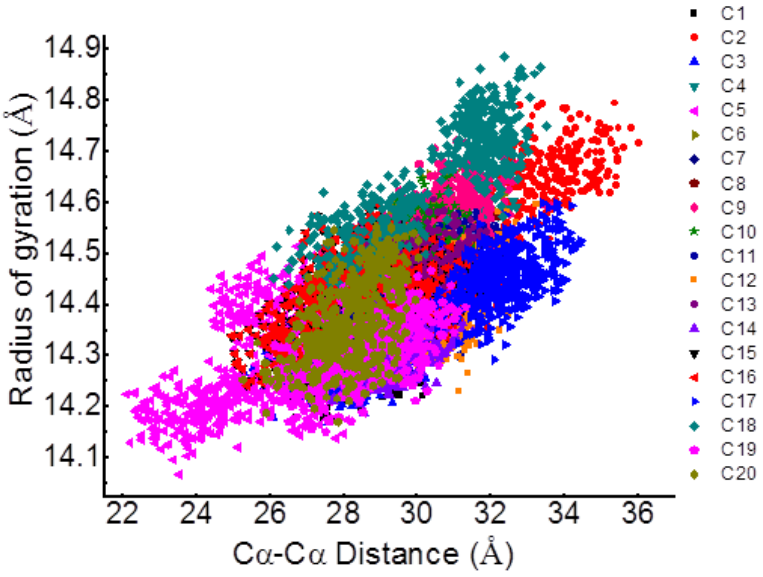

Figure 2.9: RMSD plotted against Cα-Cα distance for 20 individual MD trajectories

To further identify whether the conformational ensemble is canonical ensemble or not, RMSD relative to the minimized structure, radius of gyration of the protein, and the distance between two alpha carbon atoms for the staphylococcal nuclease (pdb:1STN) were calculated here. One pair of C_α atoms from Pro47 and Lys116 in the staphylococcal nuclease, where the distance between these two Cα atoms changes the most, is chosen for this study. The conformations are determined by the plot of RMSD or radius of gyration as a function of Cα-Cα distance along all ensemble trajectories for all the 20 CONCOORD structures of the staphylococcal nuclease as shown in Figure 2.8 and 2.9.

It appears that not only each trajectory is distinguished from other trajectories, but there are also obvious overlaps among ensemble trajectories, which indicated that the trajectories generated from different CONCOORD structures shared common conformations, thus the sampling level here is quite well converged, and the free energy barrier between two adjacent initial conformation is small enough to be overcome by two 1.6 ns simulations. Therefore, we can conclude that all conformations follow a canonical ensemble, and this method is rational from the perspective of statistical thermodynamics. In general the overlapped conformations among ensemble trajectories are corresponding to lower free energy than that of the rest conformations, which is also one of the properties of conformational Boltzmann distributions.

2.4.2. Prediction accuracy on uncharged or charged residue mutations

Using the three dielectric constants model, we calculated the RMSE of prediction results compared against experimental data. As shown in Table 2.5, the result indicated that overall, this model could provide better accuracy in predicting the change of protein stability upon uncharged residue related mutations than upon charged residues related mutations. More than 58% of the uncharged to uncharged amino acid mutations in our dataset are in buried regions of proteins, while over 85% of the charged residue related mutations occur in partial exposed or exposed region. The dielectric property of the outer region of proteins is even more heterogeneous than the inner region of proteins, thus ε=5, or ε=6 in outer region may not be suitable to all residues in this region. Therefore, the relative inaccuracy in predicting the effect of charged site mutations are related to the more heterogeneous dielectric property in outer regions of proteins. Based on the outlier definition we used before, there is one outlier in the charged to uncharged mutation group, and the RMSE drops to 1.4±0.2 kcal/mol after discarding this outlier. Using the same outlier definition, in the uncharged to uncharged mutation category, there are 7 outliers, and this model produced a RMSE of 1.7±0.2 kcal/mol and a correlation of 0.73 for the rest of 89 mutations. These outliers are due to the overestimation of Coulombic energy or the difficulty in predicting structural reorganization.

Table 2.5: Predicted RMSE for different types of mutations

class of mutations	number of mutations	RMSE (kcal/mol)
uncharged to uncharged	96	2.5±0.2
charged to uncharged	32	1.8±0.3
uncharged to charged	14	6.6±1.8
charged to charged	8	4.0±1.4
total	150	3.1±0.2

2.4.3 Prediction accuracy based on position of mutations

Since our models are built through the characterization of different dielectric regions, we analyzed the prediction accuracy of the mutations that are buried, exposed or partially exposed. With the exclusion of the two outliers in the buried mutation category and the two outliers in the exposed mutation category, the RMSE of the buried mutations decreases from 2.5 to 1.5, which is smaller than the RMSE of partial exposed or exposed region by over 0.5 kcal/mol., The correlation between the calculated value and the experimental data in the buried region is the largest among all three categories, with R=0.79, as shown in Table 2.6 and Figure 2.10.

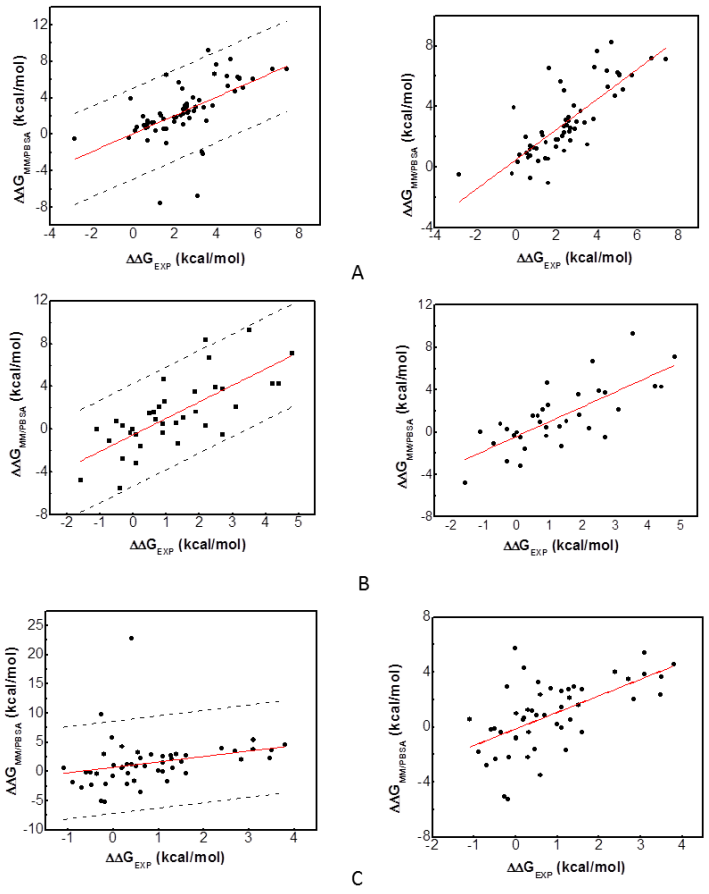

Figure 2.10 Calculated data compared against experimental data for mutations at different positions. The continuous line in every right figure is the linear regression after the outliers were discarded equation. The outliers in this figure is marked with cross. Figure 2.10A shows the results of buried mutations, the linear regression equation is y=1.00x+0.52, with R=0.79. Figure 2.10B shows the results of exposed mutations, the linear regression equation is y=1.20x-0.13, with R=0.57. Figure 2.10C shows the results of partially exposed mutations, the linear regression equation for figure c is y=1.39x-0.37, with R=0.73.

Section 2.4.2 shows that our current model could predict the uncharged site mutations more accurately than the charged residue related mutations. In our current data set, almost 90% of the buried region mutations are from uncharged to uncharged amino acids. Therefore, the majority of charged sites mutations are in partially exposed and exposed regions. Since most of the mutations in the buried region are uncharged residues, the RMSE in this region is relatively small. The buried region includes two outliers, 2lzmI58Y and 1stnV39T. One of the outliers (2lzmI58Y) is small to large mutation. The structural reorganization may be one main reason for the difficulty in predicting. For the mutation 1stnV39T, there might be an underestimation of the solvation penalty associated with the polar atoms in the buried region.

Table 2.6: RMSE of prediction for different categories of mutations

categories of mutations	number of mutations	RMSE (kcal/mol)	number of outliers	RMSE without outliers (kcal/mol)	Correlation without outliers
buried	63	2.5±0.3	5	1.5±0.2	0.79
partially exposed	36	2.4±0.4	1	2.2±0.4	0.73
exposed	51	4.0±0.6	2	2.1±0.3	0.57

This three dielectric constant model could provide more accuracy in predicting changes of folding free energy upon buried mutations or uncharged mutations. Compared with mutations in buried and partial exposed regions, the predicting accuracy for exposed mutations is relatively low, with a correlation coefficient of 0.57, which is much lower than the correlation coefficient of 0.79 for the buried mutations. Even though the overall optimal dielectric constant for the exposed region is 6 in this study, the actual dielectric

environment within each category is heterogeneous. It is possible that the polarizability of different amino acids at different positions response differently to the external electric field, thus lead to different dielectric constants. The outer region, where more charged residue side chains are included, is much more heterogeneous than the inner most region of proteins[81]. We further classified the mutations in exposed regions based on mutation types. If ε=8 is applied to the mutations from lysine to uncharged amino acid, the correlation between prediction and experimental result will be improved from 0.57 to 0.62, with a 95% confidence interval for correlation coefficient of (0.42, 0.76) Thus, due to the variety of dielectric environments in this region, it is difficult to predict accurately currently.

2.4.4. The relationship between the energy decomposition and the RMSE

The energy decomposition enables us to learn which energy term dominates the changes of protein folding free energy upon each single mutation. In the three dielectric constants model, these 150 mutants in our dataset were further classified into 5 groups based on the dominating energy terms. The RMSE between predicted results and experimental results was calculated for each category, and the details are shown in Table 2.7. The group with changes of folding free energy ($\Delta\Delta G$) being dominated by Coulomb energy results in the highest RMSE, with a value of 4.75 kcal/mol, which is more than 2 kcal/mol higher than that of the other four groups.

Table 2.7: RMSE vs. Dominated energy terms

Dominated energy terms	number of mutants	RMSE (kcal/mol)
Coulomb energy	39	4.75±0.76
Polar solvation energy	11	2.75±0.55
vdw energy	57	2.65±0.33
Internal energy	6	1.27±0.34
Equally dominated	37	0.99±0.16

Since our three dielectric constant model predicts less accurately in the group dominated by electrostatic energy (Coulombic energy and polar solvation free energy),, some factors related to the electrostatic energy were studied. The 11 mutations which are dominated by polar solvation free energy were chosen for the further electrostatic interaction studies. First of all, the polar solvation energy was recalculated at different resolutions of the grid for the PB solver, and the results suggested that decreasing the grid space from 0.6 Å to 0.25 Å could not affect the correlation between the calculation results and the experimental data with R=0.84.

The effect of salt concentration on prediction accuracy is also investigated. Even though the concentrations of mobile ions may affect the solvation free energy, it was found here that the salt concentration could barely improve the prediction discrepancy of the outliers (data is not presented here). This result is consistent with our expectations. In the system with existence of salts, since there are many ions between two charges, the electrostatic interactions between them is strongly screened. The solvent dielectric constant is screened by a factor of $e^{\kappa r_{ij}}$ when calculating the polar solvation free energy[19, 99]. The coulomb energy is screened by a factor of $\left(\frac{e^{\kappa\alpha}}{1+\kappa\alpha}\right)^2 e^{-\kappa r_{ij}}$, where κ^{-1} is the Debye length,

α is the particle radius, r_{ij} is the center to center distance between two particles. Because of the high anti-correlation between the coulomb energy and the polar solvation free energy, the salt screening effects for both energy terms can be partially canceled out. Therefore, the large changes of coulomb energy, which lead these mutants to be outliers, still dominate the overall changes of energy upon point mutations.

2.4.5. Comparison with other methods

The three dielectric constants model provides a method to predict the changes of folding free energy upon single point mutations, and also demonstrates the the importance of considering the heterogeneous dielectric properties of protein in predicting ΔΔG. Even though the three dielectric constant model is not good at predict the effect of point mutations in the exposed region of proteins, it could provide good accuracy in calculating the changes of folding free energy upon buried site mutations, with the RMSE of 1.51 kcal/mol, and a correlation coefficient of 0.79 (with a confidence interval of (0.68, 0.87)) for the buried site mutations (exclusion of 5 outliers in this region). In the CC/PBSA method, the correlation to experimental results for the buried mutations is 0.70 with the confidence interval of (0.63, 0.76), which is lower than what we presented here. The overall correlation for 150 mutations is 0.50 with RMSE=3.07 kcal/mol, but after discarding 4 outliers, the correlation increased to 0.70, and RMSE=2.09 kcal/mol. This overall correlation is lower than the CC/PBSA method, where R=0.75, σ=1.04 kcal/mol.

The low prediction accuracy of exposed site mutation is one reason that leads to the lower overall correlation coefficient in our model. The heterogeneous protein dielectric

properties depend on the position of amino acids, and the prediction accuracy is affected by the various dielectric environments of the exposed region. To improve the prediction accuracy of the exposed mutations, more knowledge of the variety of dielectric property in the exposed region is needed. The three dielectric constants model produces lower correlation coefficient than Fold-X (R=0.73, σ=1.02 kcal/mol) and Eris[46] (R=0.75, σ=2.60 kcal/mol), which may be related to the number of weighing factors in the models. There are 5 weighting factors in Fold-X[45], and 20 weighing factors in Eris, but the dielectric constant is the only one flexible parameter in our current model, thus, our model is considered to be physically rational.

2.5 Conclusion

Molecular dynamics does a good job in accurately predicting the conformational energy landscape of proteins during short timescale. Simulation results of the short timescale could not represent the protein motion in laboratory experiment, since it is difficult for proteins to overcome the high energy barriers between two conformational state during a short timescale simulation under room temperature. On the other hand, a long timescale MD simulation is also unrealistic due to the limitation of computational efficiency. Here, in order to predict protein behavior in long timescale and reduce computational expense, we worked on increasing the conformational diversity and applying several nanosecond time scale simulations to simulate the behavior of proteins in long timescale. We then applied the MM/PBSA method to calculate the changes of protein folding free energy upon single mutations.

This study also revealed the diverse dielectric properties of proteins. In general, the dielectric constant is lower in the innermost region, while higher on the surface of proteins. The value of 4, 5, and 6 were applied for the interior, partially exposed, and exposed region, respectively. This trend was consistent with the calculated results of the protein dielectric constants by Simonson et al[81]. This three dielectric constants model provided more accurate prediction of the aliphatic group mutations and buried mutations than the other mutations.

CHAPTER THREE

EFFECT OF ACCUMULATED MUTATIONS ON PLASMODIUM FALCIPARUM DIHYDROFOLATE REDUCTASE DRUG RESISTANCE

ABSTRACT

Dihydrofolate reductase-thymidylate synthase (DHFR-TS) in plasmodium falciparum (pf) is a bifunctional protein. The pfDHFR domain plays an essential role in the folate pathway, reducing dihydrofolate (DHF) to tetrahydrofolate (THF), which is crucial in the production of purine, pyrimidine, and amino acid. Therefore, DHFR-TS usually acts as the therapeutic target for malaria. Pyrimethamin (Pyr) is one of the antimalarial drugs. However, during the course of Pyr treatment, mutationsoccurred, such as S108N, C59R/S108N, and N51I/C59R/S108N/I164L/, thus leads to antimalarial resistance. To gain more insight into the drug resistance mechanisms, we applied the molecular dynamics simulation to study the wild type, C59R/S108N, and N51I/C59R/S108N/I164L/ mutant pfDHFR-TS, which are all complexed with Pyr and NADPH. The calculation results indicate that the interaction between pfDHFR and Pyr is decreasing as mutations accumulating, and is the weakest in the N51I/C59R/S108N/I164L mutant. This result is consistent with the experimental study.

Mutations in this study cause significant structural and dynamic changes in the Pyr binding pocket and in the Leu46 loop (residue 42~50). The hydrogen bond strength between D54 and Pyr, the ring-ring stacking interaction between Pyr and NADPH, and the interaction between Pyr and Leu46 loop are all getting weaker in mutants than in the wild type pfDHFR, while they are the weakest in the quadruple mutant. In the simulation of the N51I/C59R/S108N/I164L quadruple mutant, the hydrogen bonds between L46 and K49 and between I51 and D54 break, which is an important reason for weak binding between this mutant and Pyr. The structural and dynamic changes caused by mutations also increased the number of communities in both double and quadruple mutants. Significantly weakened communications among key residues, which contribute to pfDHFR-Pyr association in quadruple mutant, was observed in this study. The weakened communications cause the failure of conformational reorganization upon the binding Pyr, and lead to weak binding between pfDHFR and Pyr.

3.1 Introduction

Malaria is caused by the infection of plasmodium parasites, and is transmitted to human by female Anopheles mosquitoes[100]. Among all five species of plasmodium parasites, plasmodium falciparum (pf) is the deadliest. Currently, around 400 million people are suffering from this disease with 130 million new cases occurring each year. The treatment of malaria is achieved through the inhibition of dihydrofolate reductase (DHFR) in plasmodium falciparum[101-103]. DHFR in plasmodium falciparum forms a bifunctional protein with thymidylate synthase (TS). The DHFR-TS of plasmodium falciparum consists

of 608 amino acids. The first 231 residues are in DHFR domain, the last 288 residues form TS domain, and these two domains are connected through a junction region comprising of 89 residues[26].

DHFR plays an essential role in the folate pathway, which reduces dihydrofolate to tetrahydrofolate with the help of a cofactor nicotinamide adenine dinucleotide phosphate (NADPH). Because tetrahydrofolate is crucial in the production of purine, pyrimidine, and amino acid, the deficiency of tetrahydrofolate can lead to the failure of cell division. Thus, inhibiting the activity of DHFR can reduce the level of tetrahydrofolate, as a result, cells growth and proliferation can be impeded.

To inhibit pfDHFR-TS, antimalarial drugs, such as pyrimethamine(Pyr), have long been used. However, during the course of Pyr treatment, mutations occurred, and have led to antimalarial resistance. The development of antifolate resistance in pfDHFR is widely studied in experiment[104-106]. It is found that SER108 is critical for drug resistance and catalytic activity of pfDHFR. Substitution of SER108 with most other amino acids can cause great decrease or absence of pfDHFR activity, but the mutation S108N can retain catalytic function of the enzyme[107]. Furthermore, mutation S108N is the origin of the subsequent multiple sites mutants with higher level of antimalarial resistance. The steric interaction between the bulky side chain of ASN108 and the p-Cl atom of the 5-p-chlorophenyl group in Pyr is one key reason associated with the decreased binding affinity of pfDHFR towards Pyr[108]. Other frequent mutations appear as A16V, N51I, C59R, S108T and I164L[105].

The double mutants C59R/S108N is one of the key mutants in pfDHFR responsible for high levels of resistance against Pyr. As reported previously, these double mutations at position 59 and 108 can result in about 48 fold higher inhibition constant[105]. Similar to the single mutant S108N, ASN108 in the double mutant also has close contacts with the nicotinamide ring of NADPH and the p-chlorophenyl group of Pyr.

The quadruple mutant N51I/C59R/S108N/I164L shows the maximum global resistance towards Pyr. It exhibits approximately 570-fold increase in the inhibition constant. Among all these four mutations, only S108N and I164L occur in the Pyr-pfDHFR interaction domain. Brown and his coworkers[109] studied the possible mutational pathway from wild type to this mutant. They found the ten most frequent mutational pathways from wild type to this mutant, and only these four mutations are involved in the top ten pathways.

Apart from the experimental studies on pfDHFR, only a few computational studies were performed to gain some insight on the mechanism of changing binding affinity between Pyr and pfDHFR upon mutations[30, 110-113]. Homology modeling of wild type pfDHFR and the mutants suggested that the mutation C59R could cause repulsion between the Pyr and the positive charge of Arg[30]. The key residues contributing to tight binding in Pyr were identified through molecular dynamics simulation, and these residues are I14, D54 and I164[113], which directly interact with Pyr through hydrogen bond. Mutations occurring at these positions might change the hydrogen bond network, which lead to weak binding affinity between pfDHFR and Pyr.

Despite abundant studies of pfDHFR, there are still several problems to be solved. First of all, the cause of developing resistance of pfDHFR towards Pyr upon C59R/S108N

and N51I/C59R/S108N/I164L mutations is not well understood. Perturbations of protein conformations and dynamic properties by mutations might be one of the causes, but they are not well explored. The changes of these properties may directly relate to the changes of binding affinity of pfDHFR towards antifolate drugs. Elaborating this cause will be important for developing new antifolate drugs. Secondly, C59R/S108N and N51I/C59R/S108N/I164L mutants will not only affect the Pyr-pfDHFR binding, but will also change the catalytic function of the pfDHFR protein. The catalytic function is facilitated through the Leu46 loop (residue 42-50). However, the influence of mutations on the Leu46 loop is not identified yet. Thus, more efforts are needed in addressing these problems.

Here we performed the molecular dynamics (MD) simulation for wild type, C59R/S108N, N51I/C59R/S108N/I164L mutant pfDHFR-TS, with the aim of understanding how these mutations lead to drug resistance through analyzing the changes of the structure and dynamics of Pyr ligand binding pocket and the Leu46 loop upon mutations. The calculation results indicate that the binding affinity between pfDHFR and Pyr is decreasing as the accumulation of mutations, and the binding is the weakest in the N51I/C59R/S108N/I164L mutant, which is consistent with the experimental study. Our work provides an insight into the mechanism of pfDHFR resistance to Pyr, where the changes of hydrogen network and the ring-ring stacking interactions in the binding pocket upon mutations are associated with weak binding affinity Both enthalpic and entropic changes caused by mutations in pfDHFR lead to weaker communication among key residues that contribute to protein-ligand binding. The binding pocket of

N51I/C59R/S108N/I164L mutant is more rigid than that of the wild type pfDHFR, which also contributes to the loss of binding affinity. Besides the effect on interactions in the binding pocket, mutations also disturb the interactions between Pyr and Leu46 loop, and the conformation of this loop is significant changed upon quadruple mutations. Our results indicate the importance of developing antimalarial drugs with higher flexibility in future.

3.2 Methods

3.2.1 Simulation System preparation

The initial coordinates for wild type, C59R/S108N, and N51I/C59R/S108N/I164L were obtained from the protein data bank, corresponding to PDB ID 3QGT[114], 1J3J[26], and 3QG2[114] respectively. All three DHFR variants are bound to an inhibitor Pyr, and a cofactor NADPH. N1 atom of Pyr is modeled as protonated state, as validated in the previous NMR studies[115, 116]. Topology files and parameter files of Pyr and NADPH are obtained through the CHARMM General Force Field (CGenFF) program[117]. The initial version of CGenFF was based on CHARMM biomolecular force field, and parameters of Pyr and NADPH were examined and found to be consistent with those parameters for corresponding chemical groups within the CHARMM36 force field. The initial structures were built in CHARMM[13], using the package c35b6 with CHARMM36 force field, and the hbuild command in CHARMM was applied to build missing hydrogen coordinates. The energy of initial crystal structures was then minimized by CHARMM. In order to hold the atoms near the desired positions, a harmonic restraint was applied during energy minimization over 26 cycles in vacuum with a restraint force constant reduced from 30 kcal/mol/$Å^2$ to 5

kcal/mol/Å^2 at a decrement of 1 kcal/mol/Å^2 in each cycle. During each cycle, the structures were first minimized for 3000 steps using the steepest descent (SD) algorithm, followed by another 5000 minimization steps applying the Adopted Basis Newton-Raphson (ABNR) algorithm. A cutoff of 14 Å was applied, and the dielectric constant was set to 1, so that the permittivity is the same as that in the vacuum, and proteins are able to sample all conformations.

Due to the feature of CHARMM force field, the TIP3P water model is applied in the simulation. Each protein-ligand complex was then placed in a periodic cubic box containing TIP3P water with a density of 0.0555 mol/mL (or approximately 1 g/ml). Though there is not a hard rule for the ratio of solute and solvent atoms, sufficient TIP3P solvent molecules are required to allow the solute to interact with solvent, as well as to prevent the solute from interacting with its image while applying periodic boundary conditions. On the other hand, too many solvent molecules could make the simulation very expensive. Thus, the ratio of atoms in complex and atoms in TIP3P water is around 1:10 in our systems. TIP3P water molecules overlapping with complex were removed. Na^+ cations and Cl^- anions were added to the system to maintain a neutralized system and an ion concentration of 0.15 M, and the neutralized environment can ensure the application of the PME method for the electrostatic energy calculation.

3.2.2 Molecular dynamics simulation

After the process of solvating and adding ions to the system, the energy minimization of the system was then performed in NAMD 2.10[118] by two steps. In the first

step, the coordinates of protein-ligand complex and all crystal water were fixed, and the solvent molecules were minimized for 5000 steps applying the method of conjugate gradients (CG) to avoid any nonphysical contacts from the water atoms. In the second step, 8000 steps of CG energy minimization were performed for the whole system to remove any bad contacts.

NAMD 2.10-GPU was applied to perform all simulations using CHARMM36 force field[119]. For each protein, the MD simulations were repeated for three times under the same conditions of temperature, pressure, initial coordinates and simulation procedure, but different random seeds. During each simulation, the system was first heated from 100 K to 300 K within 1.6 ns with a temperature increment of 0.25 K every 2000 steps. The system was then equilibrated for another 6.4 ns, and followed by another 60 ns production run. The time step is 1 fs for heating and equilibration, and 2 fs for production run. Since the bonds involving hydrogen tend to vibrate at very high frequency, which are impossible to be simulated in a large time step MD simulation, all bonds between heavy atoms and hydrogen atoms were constrained through the SHAKE algorithm during the production runs. Long range Coulombic interactions were treated using the particle Mesh Ewald (PME) method[120] with a grid spacing of 1.0 Å. A smooth switching function was applied to truncate the van der Waals potential energy smoothly between 10.5 Å and 12.0 Å. Each MD simulation was performed in an NPT ensemble. The constant temperature is maintained at 300K through Langevin dynamics[121], while the constant pressure is controlled to 1 atm using the Langevin piston Nosé-Hoover method[122].

3.2.3 Binding free energy calculation

The binding affinity between protein and ligand was obtained by the molecular mechanics/generalized Born surface area (MM/GBSA)[123, 124] method and was calculated in CHARMM using equation (3.1) and (3.2):

$$\Delta G_{binding} = G_{complex} - G_{protein} - G_{ligand} \tag{3.1}$$

$$G_{MM/GBSA} = \langle E_{VDW} \rangle + \langle E_{ELEC} \rangle + \langle E_{INT} \rangle + \langle G_{Solv} \rangle - T\langle S \rangle \tag{3.2}$$

MM/GBSA calculation was applied to single structures, which were extracted from the MD simulation trajectory every 200 ps, and the final MM/GBSA free energy was estimated based on ensemble average of the energy terms. Averaging over time along the molecular dynamics simulation trajectory is denoted by "< >" in equation (3.2). For each single structure, the binding free energy calculated by MM/GBSA includes four terms, which are gas phase energies, generalized Born polar solvation energy, non-polar solvation energy, and entropy of solute. The gas phase energy is the sum of Van der Waals energy (ΔE_{VDW}), Coulombic energy (ΔE_{ELEC}) and other energy terms (ΔE_{INT}), such as bond energy, angle energy and dihedral angle energy, where no non-bond cutoff was applied to these energy terms calculations. The generalized Born solvation energy term (ΔG_{Solv}^{polar}) is calculated through the GBSW module in CHARMM. The dielectric constant for GB calculation is set to 4, the salt concentration was set to 0.05 M, which is based on the experiment environment, and the temperature is 300 K in this calculation. The non-polar solvation energy ($\Delta G_{SA}^{non-polar}$) was calculated through evaluating the solvent accessible

surface area (SASA) with a radius probe of 1.4 Å, and this energy term was calculated by the following equation:

$$\Delta G_{SA}^{non-polar} = \gamma SASA + b \tag{3.3}$$

Where the value of 0.00542 kcal/mol/ Å^2 was used for the surface tension γ, and the value of 0.92 kcal/mol was used for the constant b[125]. We assume that the influence of mutations on the change of solute entropy (ΔS) upon protein-ligand binding is small and can be neglected[7, 126], which is discussed in the results section

3.2.4 Trajectory analysis

Distance between certain atoms, and the root mean square fluctuation (RMSF) (equation (3.4)) around the average structure from MD simulation trajectory for each alpha-C atom were calculated using the CHARMM c35b6 package with CHARMM36 force field. Equation (3.4) is the calculation of the RMSF of a single atom, the summation runs over a specified set of *N* Cartesian coordinate of this atom along the MD trajectory, $x_i^{MD}$is the position of the atom at the frame *i* in the trajectory, and $\langle x_i^{MD} \rangle$ denotes the average position of the atom along the MD trajectory.

$$RMS_{fluc.} = \left[\frac{1}{N} \sum_{i=1}^{N} (x_i^{MD} - \langle x_i^{MD} \rangle)^2 \right]^{1/2} \tag{3.4}$$

Protein network and community analysis[127] was performed in VMD[128]. In a protein network, each single node denotes an amino acid residue, and a pair of in contact nodes are connected through an edge. A pair of nodes are considered to be in contact if the distance between two alpha carbon atoms of the corresponding residues is within 4.5 Å for more

than 75% of frames in a trajectory. Our choice of above network definition is supported by past studies[127, 129] on network robustness and parameters defining the network, as well as our calculation results. These studies reveal the variation of cutoffs used to define contacts and changes in the parameters (75% of frames and 4.5 Angstroms cutoff between any pair of heavy atoms in residues) defining the network contacts led to minimal changes in the community distribution of the network. From our calculations on wild type pfDHFR, we also learn that changing the cutoff value to 4.0 Å instead of 4.5 Å for wild type protein could not disturb the community distribution with a community repartition difference of 0.21. Each edge is weighted using equation (3.5), where C_{ij} is the correlation value of correlated motion for the two end residues.

$$w_{ij} = -log(|C_{ij}|) \tag{3.5}$$

The betweenness of a node is defined as the number of shortest paths between pairs of nodes that pass through that node. The community structure is obtained by using the Girvan-Newman algorithm[130], which iteratively removes the node with the largest betweenness and recalculates the betweenness of the nodes affected by the removal. Modularity score is recorded to identify the clusters that result in an optimal community network. Community analysis helps to identify the overall communication among residues. Nodes belonging to the same community are strongly interconnected and can communicate with one another more efficient through a large number of routes, while connections between nodes in different communities are weaker.

3.3 Results

3.3.1. Calculation of binding affinity between pyrimethamine and pfDHFR

To obtain a quantitative view of the effect of mutations on pfDHFR – Pyr binding affinity, MM/GBSA calculations were performed. Table 3.1 presents the binding free energy between Pyr and the rest of the complex for wild type, C59R/S108N, and N51I/C59R/S108N/I164L. The calculated binding free energies were obtained from the average of three 68 ns trajectories. The interaction energy is defined as the sum of coulombic potential energy, van der Waals potential energy and polar solvation free energy.

In the wild type, the computed binding free energy is -49.68 kcal/mol, which is exaggerated by around 37 kcal/mol from the experimental measurement. The cause of the exaggerated absolute MM/GBSA free energy estimation is due to the neglection of entropy during calculation[7, 87]. Even though some studies aims to calculate absolute binding free energy ΔG_{bind}. Here, we are only interested in evaluating the change of binding affinity between Pyr and pfDHFR upon mutations. There is a connection between the MM/GB(PB)SA calculation method and statistical thermodynamics. It is reasonable to assume that the change of vibrational motion due to the loss of translational and rotational freedom upon association is minimal, and the energy landscape can be determined from a sufficiently long MD simulation[87]. It has been found that MM/GB(PB)SA performs well at identifying the effect of mutations on association process[131],[132]. Through computational alanine scanning, Kollman and coworkers[132] suggested that MM/GB(PB)SA calculation could achieve good agreement between calculation and experimental results. Therefore,

MM/GB(PB)SA is precise enough to calculate the difference of binding free energy between the wild type pfDHFR and the mutant pfDHFR.

The van der Waals potential energy term contributes more to the difference of binding free energy than the electrostatic energy. This result also indicates that the interaction between Pyr and pfDHFR is the most favorable in the wild type complex, and is the least favorable in the quadruple mutant complex. The result is consistent with the experimental data, where the quadruple mutant (N51I/C59R/S108N/I164L) is the most Pyr resistant.

Table 3.1: Energies between PYR and the rest of the complex

Proteins	ΔEcoul (Kcal/mol)	ΔEvdw (Kcal/mol)	$\Delta G_{solv\text{-}polar}$ (Kcal/mol)	$*\Delta G_{INTER}$ (Kcal/mol)	G_{EXP} (Kcal/mol)
Wild type	1.4±0.2	-29.7±0.2	-21.4±0.2	-49.7±0.2	-12.0
Double MU	3.2±0.1	-26.2±0.1	-21.9±0.1	-45.0±0.1	-9.7
Quadruple MU	1.2±0.1	-24.8±0.1	-19.8±0.01	-43.0±0.1	-8.2

$*\Delta G_{INTER} = \Delta E_{coul} + \Delta E_{vdw} + \Delta G_{solv-polar}$

The MM/GBSA calculation is also able to evaluate the interaction between Pyr and individual amino acid. Figure 3.1 presents the changes of interactions between individual residue and Pyr upon double or quadruple mutations. It is indicated that mutations mainly significantly perturb interactions between Pyr and a few key residues in pfDHFR, including L46, D54, F58, and I164. This result is as expected, because the above four residues are all important amino acids in protein active sites as reported in experiment[26]. Except for L46, all other key residues are located in the Pyr binding pocket. Interaction between D54 and antifolate drug is of crucial importance in maintaining binding affinity between pfDHFR and antifolates. As reported by Sirawaraporn and coworkers, mutations occurring at D54 can lead to detrimental effect of enzyme activity and antifolate binding affinity[133]. F58 and

I164 both have direct interaction with Pyr[26]. L46 is in the Leu46 loop (residue 42-50), which is important in facilitating the catalytic function of pfDHFR[26], and it offers multiple binding site for the cofactor NADPH[26]. Pyr makes more favorable interactions with L46, D54, and F58, but less favorable interaction with residue 164 in wild type than in C59R/S108N and N51I/C59R/S108N/I164L mutant. The interactions between Pyr and residue 164 is dominated by van der Waals energy. The changes of the paired interactions are related to the local conformation changes, and are described below.

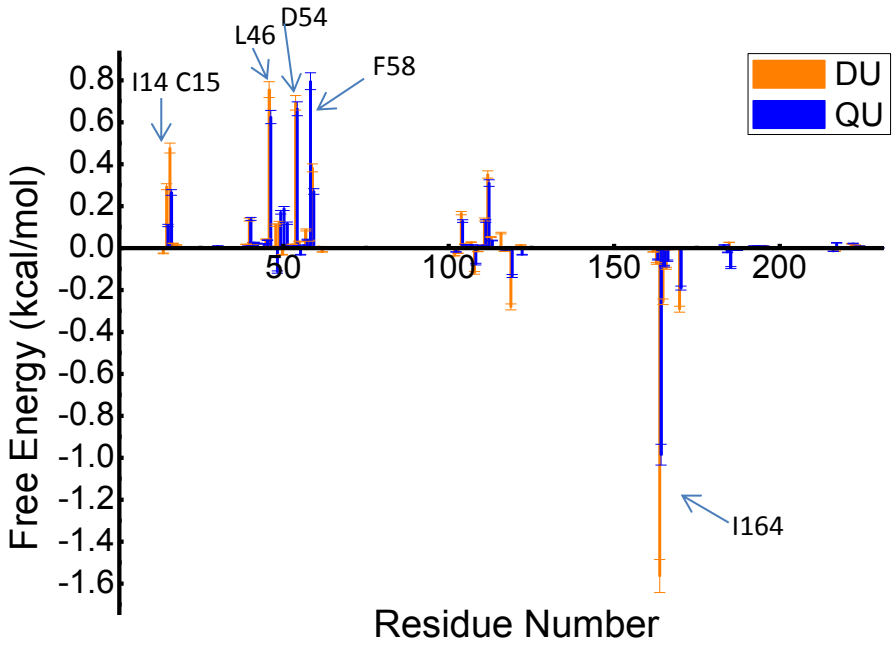

Figure 3.1 Changes of free energies between each residue and PYR upon mutations. The free energy is the sum of Coulombic energy, van der Waals energy and polar solvation free energy. The free energy of the double (quadruple) mutant minus that of the wild type gives the value of the orange (blue) color.

3.3.2 Interactions in binding pocket

To understand the reason of weak binding affinity in double mutations (C59R/S108N) and quadruple mutations (N51I/C59R/S108N/I164L), we further explored the interactions in the binding pocket. We analyzed the hydrogen bond network and the ring-ring stacking interactions inside the binding site with the 68 ns MD simulation trajectories for wild type and mutants.

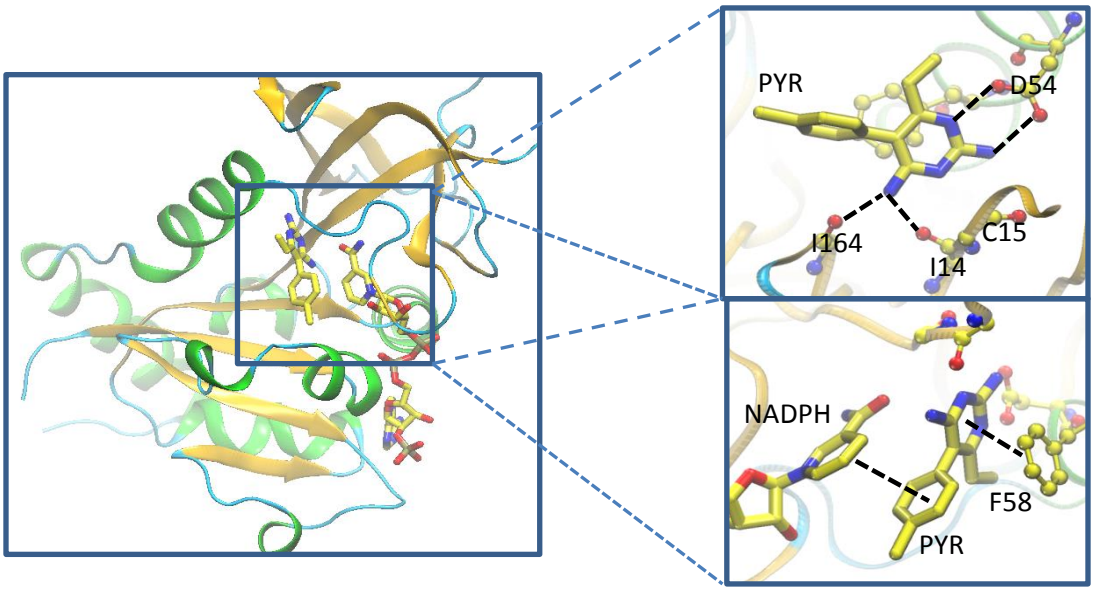

Figure 3.2 Binding pocket and two side views of the binding pocket in wild type protein

The hydrogen bond network and ring-ring interactions in wild type binding site are shown in Figure 3.2. The ligand Pyr can form hydrogen bond with residue I14, C15, D54, and I164, where the hydrogen bond between D54 and Pyr is the strongest. Other than hydrogen bond, Pyr is also stabilized in the binding pocket through the ring-ring stacking interactions with F58 and the cofactor NADPH.

We analyzed the probability distribution of center of mass distance between Pyr and residue 164 in the binding pocket, which is presented in Figure 3.3. In general, these mutants cause the movement of residue 164 towards Pyr. The center of mass distance between I164 and Pyr ligand is shortened by 0.82±0.71 Å in double and 0.42±0.65 Å in quadruple mutants compared with wild type. This observation is also consistent with the results of binding free energy calculation. Results of the binding free energy analysis indicate that the interaction between residue 164 and Pyr is dominated by the van der Waals energy. Figure 3.1 shows that this interaction is stronger in the double mutant and quadruple mutant than in wild type pfDHFR. Based on L-J potential equation, the decrease in distance can lead to more negative van der Waals potential energy in double mutant, compared against wild type.

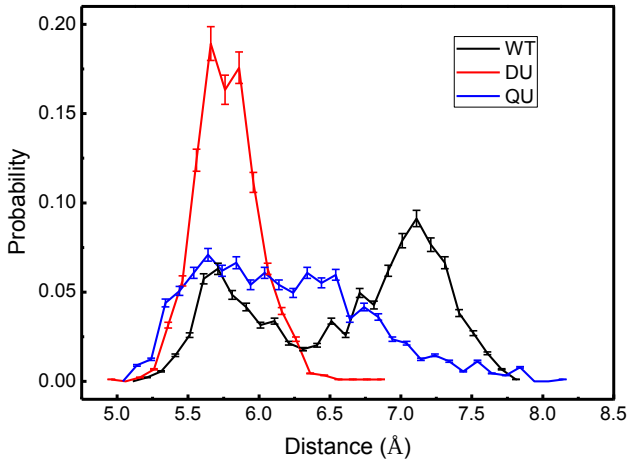

Figure 3.3 Probability distribution of center of mass distance between PYR and residue 164

The hydrogen bond forming between N14 atom of Pyr and OD1 atom of D54 was studied. The distribution of this hydrogen bond length is presented in figure 3.4. The length of this hydrogen bond gets larger as more mutations accumulating, where the average length is 2.81±0.22 Å in wild type, 2.92±0.22 Å in the double mutant, and 2.98±0.30 Å in the quadruple mutant. This figure indicates that comparing with wild type, the hydrogen bond interaction is weaker in C59R/S108N mutant and N51I/C59R/S108N/I164L mutant. Hydrogen bonds with donor-acceptor distance of 2.2-2.5 Å is defined as strong, 2.5-3.2 Å as moderate, and 3.2-4.0 Å as weak. The energy of a weak hydrogen bond is usually 0~14 kcal/mol higher than that of a moderate hydrogen bond[134]. Weaker hydrogen bonds are related to the higher interaction energies between pairs of D54 and PYR in C59R/S108N and N51I/C59R/S108N/I164L mutants, which are presented in Figure 3.1.

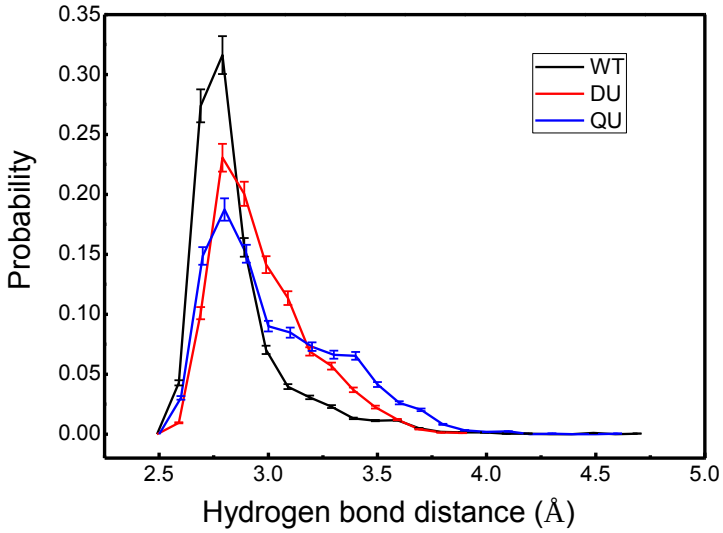

Figure 3.4 H-bond distance between N14 atom of PYR and OD1 atom of D54. The black color is for wild type, red is for double mutant, and blue is for quadruple mutant.

To understand the effect of mutations on the ring-ring interactions in the Pyr binding pocket, we examined the probability distribution of center of mass distance between Pyr and nicotinamide ring of NADPH, as shown in Figure 3.5B. The distance is most likely around 5.0 Å, 6.8 Å, and 8.9 Å in wild type, C59R/S108N and N51I/C59R/S108N/I164L mutant, respectively. The trend of distance is consistent with the trend of interaction energies between Pyr and NADPH, where the most favorable interaction between Pyr and NADPH occurs in wild type. Thus, the lack of ring-ring interactions in the binding pocket of C59R/S108N and N51I/C59R/S108N/I164L mutants are associate with the weaker interaction between Pyr and NADPH, and the ring-ring interactions are important for retaining Pyr in position. The ring-ring interaction between

Pyr and F58 does not vary much among wild type and mutants. The weaker ring-ring interactions in mutants are directly caused by changes of conformations of Pyr and NADPH in the binding site, as explained in Figure 3.5A.

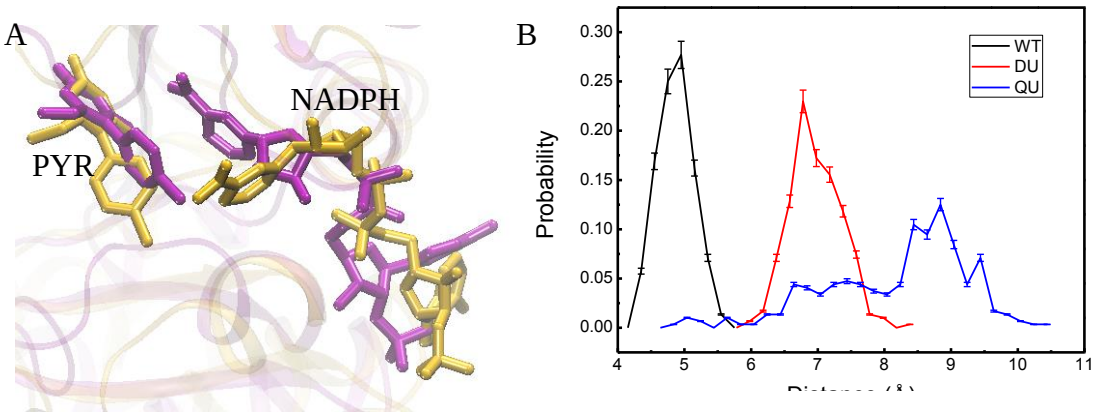

Figure 3.5 (A) Representative snapshot for the binding mode of PYR and NADP in binding pocket of wild type (purple) and quadruple mutant (yellow). (B) Center of mass distance between PYR and the nicotinamide ring of NADPH.

3.3.3 Conformation changes in binding site

To gain more insight into the effect of mutations on protein binding site conformations, we examined the shape of binding site for wild type and mutants. Here, the binding site is defined as all protein atoms within 3.5 Å of Pyr and the free volume among these atoms. Figure 3.6 is obtained from the representative snapshot from trajectories, figure 3.6A is the shape of binding site for wild type, and figure 3.6B is for N51I/C59R/S108N/I164L mutant. Comparing figure 3.6A with 3.6B, the shape of binding site is experiencing obvious changes caused by quadruple mutations.

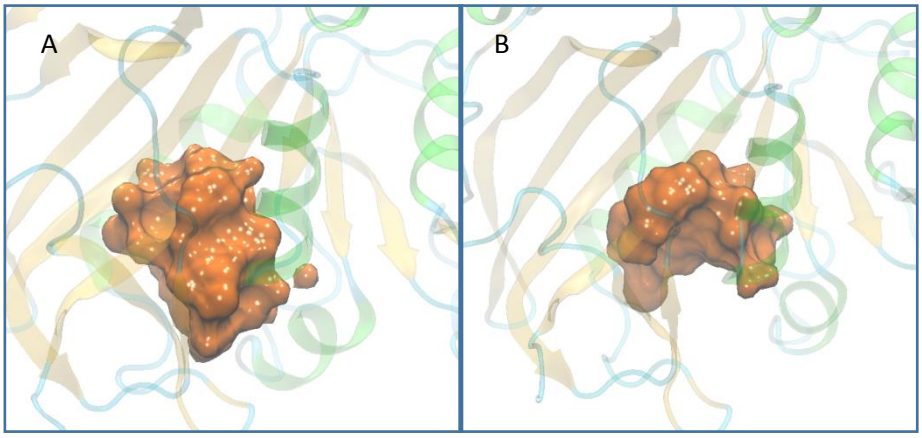

Figure 3.6 The shape of Pyr binding site in pfDHFR. (A) is for wild type, and (B) is for quadruple mutant protein.

The shape of the binding site is usually related to the binding mode of ligand. Thus, further analysis on the binding mode of Pyr is performed, so that the cause of shape change could be clear. Figure 3.7A shows the probability distribution of dihedral angle for C8, C7, C4, and C3 atoms of Pyr in wild type, C59R/S108N, and N51I/C59R/S108N/I164L. The dihedral angle in C59R/S108N and N51I/C59R/S108N/I164L is most likely to be around 60 degree, which is off by 60 degree from the 120 degree dihedral angle in wild type. As shown in Figure 3.5, the ring-ring stacking interaction between PYR and NADPH is weaker or even broken in the double mutant and quadruple mutant, therefore, the chlorophenyl ring of Pyr exhibits higher mobility in mutants than in the wild type. Furthermore, the longer side chain of N108 in mutants upon the mutation S108N has caused steric hindrance between chloride atom in Pyr and N108. The steric hindrance is another reason for the change of dihedral angle in C59R/S108N, and N51I/C59R/S108N/I164L. The change of binding mode of Pyr is associated with changes of the binding site shape and interactions in binding site, thus leads to the changes of binding affinity.

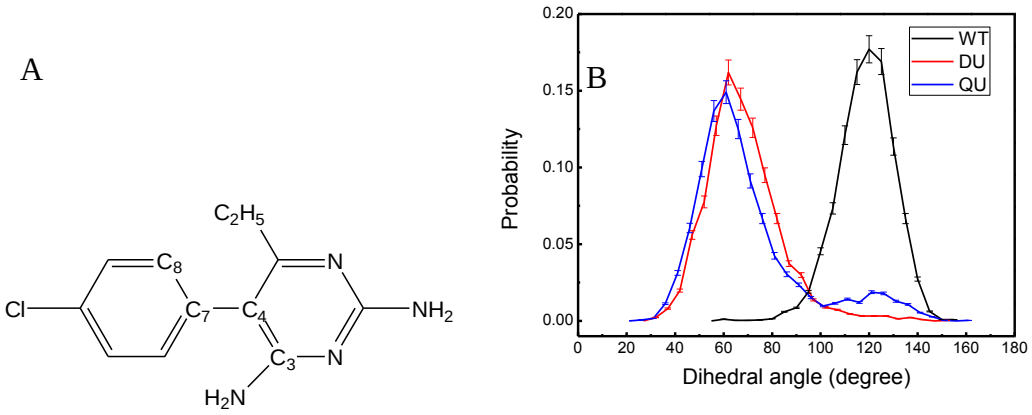

Figure 3.7 (A) The structure of PYR, with C3 C4 C7 C8 labeled. (B) Probability distribution of dihedral angle for C8 C7 C4 C3 atoms in PYR.

3.3.4 Effect of mutations on Leu46 loop conformation

Leu46 is one of the key residues that contribute significantly to the changes of binding affinity upon mutations based on the calculations described above. Rather than locating in the binding pocket, it is located in the Leu46 loop, which is important in facilitating the pfDHFR catalytic function[26]. In the wild type, the side chain of Leu46 interacts with Pyr through hydrophobic interactions. In order to further understand the effect of mutations on interactions between Pyr and Leu46, we calculated the probability distribution of center of mass distance between Pyr and Leu46, and the result is in Figure 3.8. The center of mass distance is much shorter in wild type than that in C59R/S108N and N51I/C59R/S108N/I164L, and the distance in C59R/S108N is shorter than that in

N51I/C59R/S108N/I164L. The changes of center of mass distance cause the change of interaction between Leu46 and Pyr, which is dominated by van der Waals potential energy.

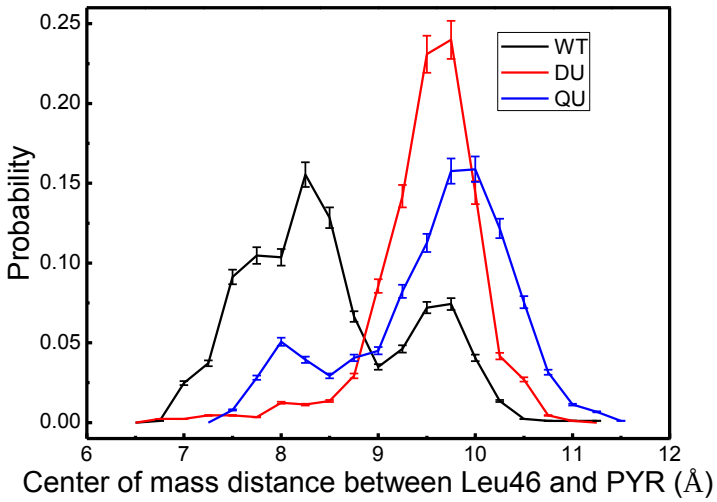

Figure 3.8 Probability distribution of the center of mass distance between Leu46 and Pyr. Wild type protein is presented in black, C59R/S108N is in red, and N51I/C59R/S108N/I164L is in blue.

To gain more insight into the effect of mutations on Leu46 loop, we explored the conformation of Leu46 loop in wild type and mutants. Snapshots of trajectories have shown that the conformation of Leu46 loop in N51I/C59R/S108N/I164L mutant is significantly different from that in wild type and C59R/S108N mutant, which is presented in Figure 3.9. The direct reason of conformation change is that the hydrogen bond between L46 and K49 is broken in N51I/C59R/S108N/I164L mutant, while it exists in wild type and C59R/S108N mutant. The change of this hydrogen bond is shown in Figure 3.10. The N51I

mutation leads to the shift of residues 48-51, and they move away from Pyr as shown in Figure 3.11. The movement is related to the conformation change of Leu46 loop in the quadruple mutant.

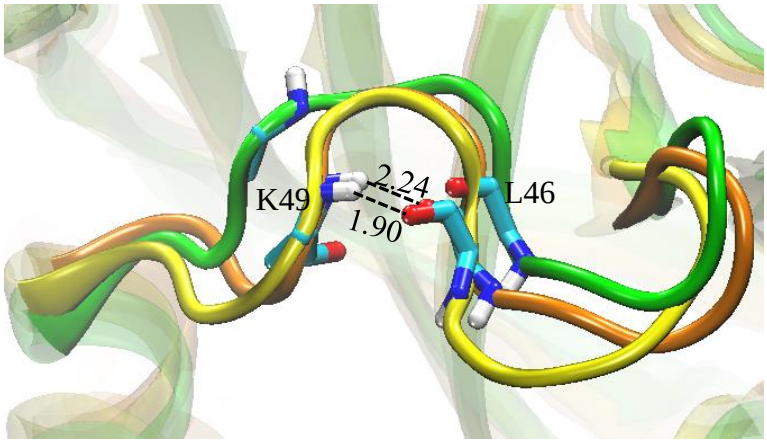

Figure 3.9 Representative snapshot for conformation changes of Leu46 loop in quadruple mutant (green). The wild type is shown in orange, and the double mutant is shown in yellow.

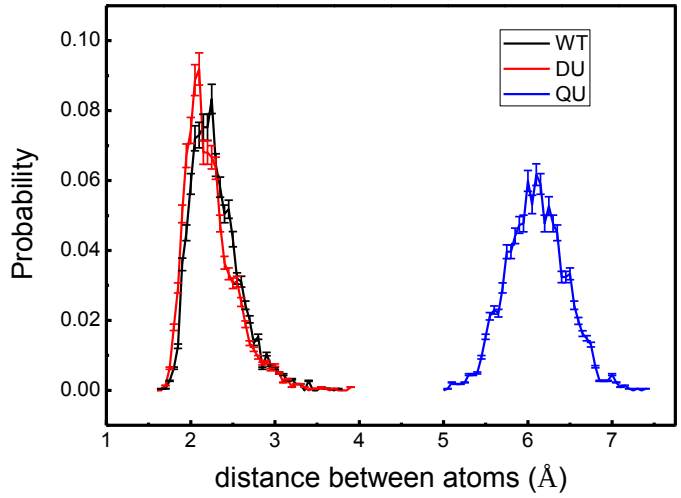

Figure 3.10 Distance between O atom of L46 and HN atom of K49

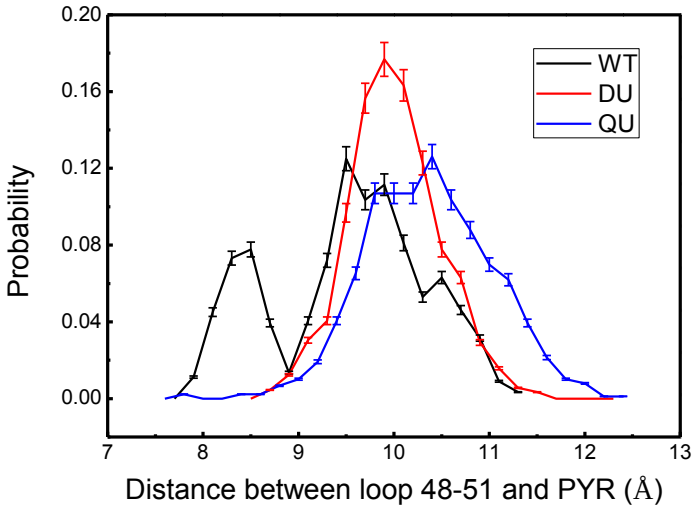

Figure 3.11 Probability distribution of the center of mass distance between residue 48-51 and PYR. Wild type protein is presented in black, C59R/S108N mutant is in red, and N51I/C59R/S108N/I164L mutant is in blue.

3.3.5 Flexibility of pfDHFR

Flexibility plays an important role in modulating protein-ligand binding affinity[135]. Process of both protein-ligand/protein interactions requires structural flexibility. To examine the effect of mutations on pfDHFR flexibility, we calculated root mean square fluctuation (RMSF) of wild type and mutants. Figure 3.12 depicts the RMSF of alpha-C of each residue in the wild type, C59R/S108N and N51I/C59R/S108N/I164L mutant. The result indicates that the quadruple mutant is less flexible than the wild type and double mutant. Since the binding pocket of quadruple mutant is the very rigid, and is not perfectly complementary with Pyr, thus, it could not tightly bind to a rigid ligand as Pyr. Furthermore, the low flexibility of quadruple mutant suggests that the binding pocket is not able to moderately modulate the conformation upon the binding of Pyr ligand. The trend in flexibility has good agreement with the trend in pfDHFR-Pyr binding affinity. N51I/C59R/S108N/I164L mutant is the most rigid among all three proteins mentioned above, and its binding affinity toward Pyr is also the weakest.

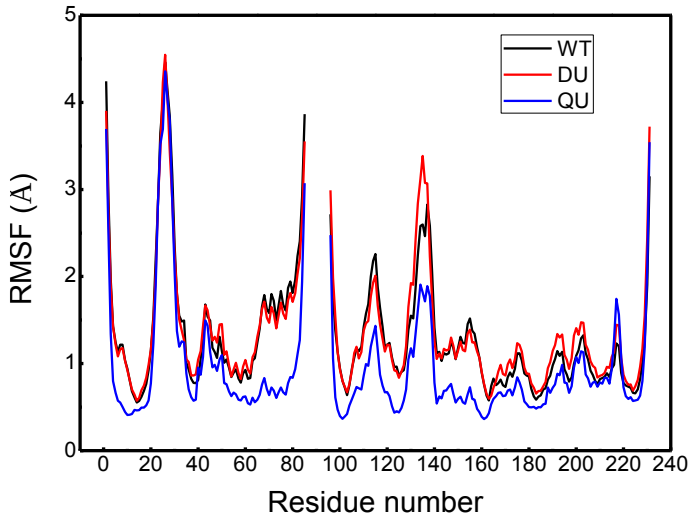

Figure 3.12 Root mean square fluctuations of wild type (Black), double (Red) and quadruple (Blue) mutant proteins.

3.3.6 Community network analysis

To gain more insight into the relationship between protein dynamic network and protein binding affinity, we performed the community network analysis. With the occurrence of quadruple mutations, the levels of network communication in pfDHFR are altered. The number of communities is 8 in the wild type protein. It increases to 9 and 11 upon the double and quadruple mutations, respectively. The community repartition difference between the quadruple mutant and the wild type pfDHFR is0.51 The above results indicate that 49% of the node pairs remain in the same community, while 51% of the node pairs are broken into different communities. This result suggests that quadruple mutations have caused weaker connection between nodes which are broken into different

communities. A significant change in community repartition can be attributed to the weakened interaction of the quadruple mutant's residues with its neighbors. Weak interaction edges in the quadruple mutant are removed early in the Girvan–Newman algorithm and, as a result, have larger overall effect on the community node assignment. Furthermore, the weaker interactions between nodes indicates that the correlation of motion between these two residues is weaker. Weaker correlation between two residues may suggest this interaction is not instrumental in defining information flow in protein. The weaker correlation between two residues may also suggest that these interactions are not contributing to determine changes in topology of protein due to events like ligand binding.

Apart from the overall community network, N51I/C59R/S108N/I164L mutations also perturb the communications among key residues as mentioned in Figure 3.1. These key residues are in 3 separate communities in the wild type, while they are in 5 different communities in N51I/C59R/S108N/I164L pfDHFR. Nodes in the same community are strongly interconnected and can communicate with one another more efficiently through multiple routes, while connections between nodes in different communities are weaker.

The betweenness of a node is defined as the number of shortest paths between pairs of nodes that pass through that node. The betweenness is used to measure the importance of the edge for communication within the network. As shown in Figure 3.13, Asp54 and Phe58 are in the same community in wild type, with high edge betweenness value, however, they are split into two communities in C59R/S108N and N51I/C59R/S108N/I164L mutant, where the betweenness is the lowest in the quadruple mutant. Even though Ile14 and

residue 164 remains in the same community after C59R/S108N mutations occurring, the communication strength between them is largely decreased. However, the strength of communication between residue 14 and 164 is restored through N51I/I164L mutations. The edge betweenness between Phe58 and Ile164 is much smaller in C59R/S108N mutation than in wild type, and this value becomes zero upon N51I/C59R/S108N/I164L mutations.

Thus, in general, key residues are connected through strong communication in wild type, but are loosely connected in N51I/C59R/S108N/I164L mutant due to the repartitioning among the community network. The strong communication among these key residues in wild type means that the information flow between these residues is more efficient, and the correlated/anti-correlated motion between these residues is more significant. Changes of correlated motions upon mutations may also affect the protein function[136]. Mutations can lead to changes of protein conformations, as well as the modification of protein internal motions, which influence the height of activation free energy barrier[137].

The Leu46 loop in wild type is in one common community. However, the breaking of the hydrogen bond between Leu46 and Lys49 in the quadruple mutant leads to the two separate communities in the Leu46 loop. Figure 3.9 has shown that the Leu46 loop in N51I/C59R/S108N/I164L mutant experiences conformation changes. Beside the broken of hydrogen bond between Leu46 and Lys49, the reorganization of communities in this loop is also another contributing factor for the conformational change. The mutation N51I leads to the split of community which contains residue Cys15 and the Leu46 loop. Cys 15 and

residues in Leu46 loop are located in three different communities in the N51I/C59R/S108N/I164L mutant, which indicate that the communications among these residues are weakened by the quadruple mutations.

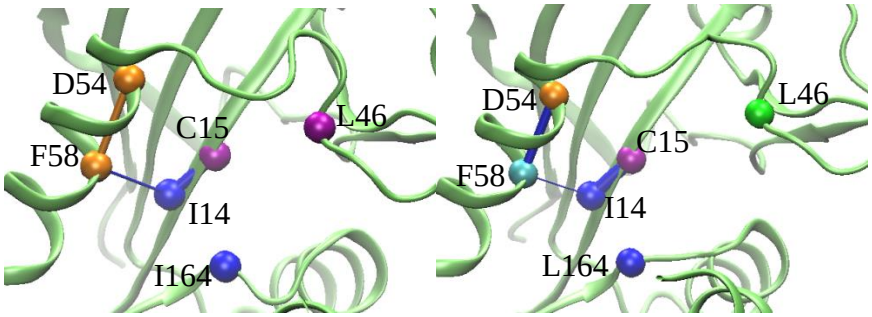

Figure 13 Communities for the key residues in wild type (A) and quadruple mutant (B). Nodes of the same color are in the same community. The nodes in the figure represent for the six key residues contributing for Pyr binding.

3.4 Discussion

3.4.1 The effect of accumulated mutations in N51I/C59R/S108N/I164L

S108N is the first mutation occurring in the quadruple mutant. The mutation S108N leads to a steric clash between the bulky side chain of N108 and *p*-Cl atom of the 5-*p*-chlorophenyl group of Pyr, which causes the change of the binding mode of Pyr in the binding pocket of pfDHFR.

The single point mutation in C59R could in fact slightly enhance the binding between Pyr and pfDHFR[105]. However, upon mutations C59R/S108N, the binding affinity between Pyr and pfDHFR is even weaker than that in S108N mutant as reported in experiment. Therefore, the overall effect of C59R/S108N is not an accumulated effect of each single point mutation. The reason may be that the interactions between the local environment of the residue 59 and residue 108 is also affected by the mutaions.

The mutation N51I in N51I/C59R/S108N/I164L mutant, even though is outside of the Pyr binding pocket of pfDHFR, not only breaks the ring-ring stacking interactions between PYR and NADPH, but also perturbs the binding pocket conformation. The local conformation rearrangement in respond to N51I leads to the breaking of the H-bonds between L46 and K49, between I51 and D54, and a weaker H-bond between W48 and C50. Hydrogen bonds between L46 and K49, as well as between W48 and C50 are crucial in maintaining the shape of the Leu46 loop in wild type pfDHFR. According to our calculation, residues 48-51 move away from Pyr in the quadruple mutant, this is consistent with the experimental result that the mutation at residue 51 caused a main chain movement of residues 48-51 with respect to the wild type[26]. The Leu46 loop offers binding sites for the second ligand NADPH. Thus, the conformation change of the Leu46 loop causes the binding mode change of NADPH, which may further lead to the loss of the ring-ring stacking interactions between PYR and NADPH. The local conformation of D54 is changed due to the broken of H-bond between I51 and D54. Since D54 tightly binds to Pyr in wild type pfDHFR through H-bond, conformation changes around D54 caused by

mutation N51 lead to a weaker hydrogen bond between D54 and Pyr, thus the interaction between D54 and Pyr is less favorable in the quadruple mutant than that in the wild type.

Residue 164 is closer to Pyr in the mutants than in the wild type, which is consistent with a previous finding[30]. The methyl group in Leu164 is closer to Pyr in the quadruple mutant, thus, Leu164 contributes to the favorable interaction between residue 164 and Pyr.

3.4.2 Reduced communication strength in Pyr binding pocket

The decreased binding affinity between Pyr and pfDHFR in quadruple mutant is associated with its weak communication among residues in the binding site. The breaking of hydrogen bond between I51 and D54, which is caused by the mutation N51I in N51I/C59R/S108N/I164L , leads to rearrangement of the mean conformation in and around D54 (enthalpic change). This change also results in the reorganization of communities, and residues in the binding pocket split into more communities. As a consequence, the communication among residues in the Pyr binding pocket is much weaker in the double or quadruple mutant than that in the wild type. Our results imply that communications among residues are affected by the changes of enthalpy. The result is also consistent with precious research that protein conformational dynamics can mediate the protein-ligand binding process[138]. In other words, the structural change caused by mutations can lead to changes of binding affinity, which can further disturb communications among residues in the pfDHFR binding pocket. These communications suggest that these residues are central to information flow as the result of pfDHFR-Pyr binding. The weaker communication among

the residues in the binding pocket of quadruple mutants further indicates the functional significance of these conformational changes in protein.

Communications among residues not only occurred from enthalpy contribution, but also from entropic contribution[139-141]. The fluctuation of residues about the mean position can propagate residue communications. Weaker correlations among the key residues in the binding pocket of quadruple mutants relate to a decrease in information flow as compared with the binding event in wild type, thus the key residues in the binding site of quadruple mutant are not capable of responding to the fluctuation of other residues. Because the structural fluctuation plays an important role in modulating protein-protein or protein-ligand interaction[142], the low flexibility of quadruple mutant suggests that the binding pocket is not able to moderately modulate the conformation upon the binding of Pyr ligand. Consequently, the binding affinity between Pyr and quadruple mutant is quite low. As presented in Figure 3.7, the quadruple mutations lead to the change of binding mode for Pyr in response to the binding pocket conformation change. However, Pyr is too rigid to achieve a tight binding in the relatively rigid binding pocket of quadruple mutant.

The flexibility of both the protein and the ligand plays significant roles in regulating the binding interactions in the protein-ligand complex[143]. Therefore, it is necessary to design drugs with higher flexibility to accommodate to the rigid binding pocket of N51I/C59R/S108N/I164L mutant. The flexible ligand may be able to bind to the rigid protein with multiple binding mode, which can increase the possibility of the ligand binding to multiple binding sites of the protein. Thus, designing a flexible ligand for

N51I/C59R/S108N/I164L mutant may result in favorable enthalpy of binding upon association.

3.5 Conclusion

In summary, pfDHFR is an important target for antimalarial drugs, such as Pyr. The highest Pyr resistance occurs in N51I/C59R/S108N/I164L mutant. However, it's not clear of how this high resistance occurs. In this study, we applied the MD simulations for the wild type, double mutant and quadruple mutant pfDHFR, all of which are complexed with Pyr and NADPH. The structure and dynamics of the binding pocket and the Leu46 loop, which are affected by mutations, can impact the protein-ligand binding affinity, and are crucial for the development of drug resistance. The Leu46 loop, even is outside of the Pyr binding pocket, can perturb Pyr binding upon quadruple mutations. The breaking of Leu46-Lys49 hydrogen bond causes the weak ring-ring stacking interaction between Pyr and NADPH through the conformation change of the Leu46 loop. Communications among key residues are much weaker in quadruple mutant that in wild type or double mutant. The weaker communication can result in less efficient information flow among these key residues upon pfDHFR-Pyr binding. Furthermore, among all three proteins studied above, the binding pocket of quadruple mutant is the most rigid, and is not perfectly complementary with Pyr, thus, it could not tightly bound to a rigid ligand as Pyr. Therefore, antimalarial drugs with higer flexibility need to be designed in future in order to achieve tight binding with pfDHFR.

CHAPTER FOUR

PROBING THE ROLE OF N-TERMINAL TAIL ON ACTIVITY AND DOMAIN-DOMAIN COMMUNICATION IN PLASMODIUM FALCIPARUM DIHYDROFOLATE REDUCTASE-THYMIDYLATE SYNTHASE

ABSTRACT

Dihydrofolate reductase-thymidylate synthase (DHFR-TS) in plasmodium falciparum (pf) is a bifunctional protein. The DHFR domain and the TS domain in plasmodium parasites are encoded by a single gene and are expressed as one protein. Even though the DHFR function is conserved in plasmodium parasites, comparing with DHFR in bacteria or other eukaryotes, there are several unique structural features in plasmodium falciparum. One of these unique features is the N-terminal tail in pfDHFR. The N-terminal tail is essential in modulating the interactions between DHFR and TS, as well as maintaining the pfDHFR activity. Since the N-terminal tail is remote from the pfDHFR active site, it is not clear how this distant tail could perturb the protein activity and the domain-domain interaction. In this chapter, the role of the N-terminal tail in domain-domain communication and DHFR activity in plasmodium falciparum is examined through molecular dynamics simulations, correlated motions, and principal component analysis. It

is found that the deletion of the N-terminal tail can disturb the pfDHFR activity through indirectly changing the local conformation of SER108, which is a key residue in maintaining the pfDHFR activity. While involving the interactions between DHFR and TS, the N-terminal tail also has an impact on the strong anti-correlated motions between the two DHFR domains in pfDHFR-TS dimer.

4.1 Introduction

Dihydrofolate reductase – thymidylate synthase (DHFR-TS) in plasmodium falciparum (pf) is an important bifunctional protein involving in the process of DNA production[144, 145]. Unlike eukaryotes, where DHFR and TS are expressed as two distinct enzymes, they are encoded by a single gene and are expressed as one bifunctional protein in plasmodium parasite[146]. The DHFR and TS domain is connected through a 54-residue junction region[26]. The bifunctional pfDHFR-TS is a dimeric protein (Figure 4.1)[147, 148] and the dimerization of the bifunctional protein is formed through broad contacts between the two TS domains. Even though some structural features are conserved in the DHFR family, there are some unique features that are only found in pfDHFR[26]. PfDHFR has two extra inserts which are not found in eukaryotes. The first insert comprise residues from 20 to 36, and the second one from residue 64 to 99. Besides the extra inserts, there is an N-terminal tail existing in pfDFHR, which has found to be structurally and functionally important for pfDHFR-TS[31].

The N-terminal tail consists of the first five amino acids in the DHFR domain of pfDHFR-TS. Such N-terminal tail does not exist in bacteria or other eukaryotes DHFR,

and it is also absent in Cryptosporidium hominis bifunctional DFHR-TS, but it could be found in other bifunctional DHFR-TS proteins. For instance, in Leishmania major, there is a 22-residue tail, wrapping around the enzyme surface[149]. The 5 amino acids long N-terminal tail in pfDHFR is located on the surface of pfDHFR-TS, and a hydrogen bond network is formed among residues GLN4, VAL5, ASP7, VAL8, and PHE9. Thus, the N-terminal tail plays an important role in maintaining the stability of the protein structure. The N-terminus is remote from the active region of the DHFR domain, and the exact functional role of this tail is unknown yet, but it has been reported that the N-terminal tail is crucial in modulating the activity of pfDFHR-TS protein[31, 32, 150].

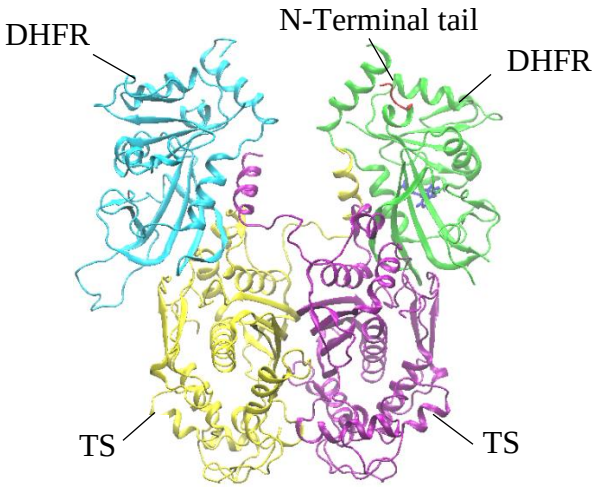

Fig 4.1 The dimeric structure of the bifunctional pfDHFR-TS protein. The tail in red color is the N-terminus in pfDHFR domain

To obtain a deeper understanding of the role of N-terminal tail in pfDHFR, the truncation of N-terminal tail in the pfDHFR-TS monomer has been studied previously[150]. The result showed that deleting MET2 or further deleting GLU3 in the N-terminus did not perturb the activity of pfDHFR at all. Thus, it is concluded that MET2 and GLU3 might not be important for maintaining the pfDHFR activity. However, poorer activity was observed after the further deletion of residue GLN4 and VAL5 in the N-terminal tail. The activity of pfDHFR is even completely prohibited after further deleting CYS6 and more subsequent residues. Besides the study of N-terminal tail in pfDFHR-TS monomer, Dasgupta and Anderson also examined the effect of N-terminus on the activity of DHFR in the full-length dimeric pfDFHR-TS protein through single-turnover experiments[31]. After deleting the residues from 2 to 5, the mutant DHFR rate is only about half of that in the wild type. However, different from the results in pfDHFR-TS, deleting the 22 residue N-terminal tail in L. major DHFT-TS can surprisingly increase the DHFR rate. Therefore, the mechanism of N-terminal regulating DHFR activity is different for different species.

In spite of the findings in the role of N-terminal tail in pfDFHR, unsolved problems still exist. Up to now, there is not a solid explanation for the decreased DHFR activity causing by deleting the N-terminal tail. One possible explanation is that the deletion of the N-terminal tail disturbs the interactions of N-terminus with Insert II or αβ loop (residue 141-184), which may further perturb the geometry of Insert II or αβ loop[31]. Even though Insert II and αβ loop are not in the active sites, they play an important role in keeping the conformation of active sites. The changes of active site conformation may alter the DHFR activity. However, this statement is lack of support in either experimental or computational

study. Thus, further study is needed to verify this statement. A reasonable explanation for decreasing pfDHFR activity by deleting the N-terminal tail will provide insight into novel strategies in inhibiting malaria parasites DHFR-TS.

The physical interactions between DHFR and TS are essential for maintaining the catalytic function of pfTS[32]. Furthermore, it is also demonstrated that only the right conformation of pfDHFR can retain an active pfTS. Since the N-terminal tail may play an important role in keeping the pfDHFR in a correct conformation, it is possible that the deletion of N-terminal tail may change the conformation of DHFR domain, and thus disturb the communication between DHFR and TS domains. However, the function of N-terminal in maintaining the domain-domain communication has not been verified, and it is not clear how such communication affect the protein function.

In present study, we performed molecular dynamics (MD) simulations for the full-length dimeric wild type pfDHFR-TS and the mutant with deletion of the N-terminal tail (residue from 1 to 5), with the aim to understand: 1) the domain-domain interaction in bifunctional DHFR-TS, 2) the role of N-terminal tail in modulating the domain-domain interactions, 3) the role of N-terminal tail in maintaining the conformation of DHF binding pocket conformation, and 4) the role of N-terminal tail in maintaining the activity of pfDHFR-TS. In this study, covariance matrix analysis and principal component analysis was utilized to demonstrate the importance of the N-terminal tail in the domain-domain communication of the dimeric bifunctional protein. This study also proposed that the decreased activity by the deletion of the N-terminal tail is due to the disturbance of the binding site of DHF, especially SER108. Therefore, the N-terminal tail, even though is

remote form the active site, is important in preserving the active site conformation. This is the first computational study that focused on the full-length dimeric pfDFHR-TS with the aims to understand the domain-domain interaction and the effect of the distant mutations on the conformation of the active site.

4.2 Methods

To assess the effect of the N-terminal tail on the communication and catalysis of pfDHFR-TS, we simulated the wild type enzyme as well as the mutant bound to DHF and NADPH cofactor. The wild type structure and the mutant structure is obtained from the crystal structure (PDB ID: 4DPD[151]). In the following steps of our study, we performed molecular dynamics simulations, covariance matrix analysis, principle component analysis, and molecular mechanics/generalized Born surface area (MM/GBSA)[19] free energy calculations. The detailed description of the methods used in this study is presented below.

4.2.1 System preparation

The initial coordinates for wild type pfDHFR-TS crystal structure are obtained from the protein data bank, with PDB ID 4DPD. The mutant is obtained by manually deleting the N-terminal tail in the wild type protein. Both the wild type and the mutant are in the form of dimer. In the crystal structure, each dimer is bounded to one DHF, one NADPH, and two UMP ligands. Parameter files of small ligands (DHF, NADPH, UMP) are obtained from ParamChem through the CHARMM General Force Field (CGenFF) program[117]. The initial structures were built using c35b6 CHARMM[13] package with CHARMM36 force

field. The missing hydrogen coordinates were added by applying the hbuild command in CHARMM. The energy minimization of the crystal structure were performed by CHARMM to remove the nonphysical contacts. During energy minimization, a harmonic restraint with a restraint force constant reducing from 30 kcal/mol/$Å^2$ to 5 kcal/mol/$Å^2$ was applied over 26 cycles in vacuum. In each cycle, the structures were first minimized for 3000 steps using the steepest decent (SD) algorithm, followed by another 5000 minimization steps applying the Adopted Basis Newton-Raphson (ABNR) algorithm. A cutoff of 14 Å was applied.

Each system was solvated in a periodic cubic box containing TIP3P water with a density of 1 g/ml. The system was neutralized and maintained at a salt concentration of 0.15 M by adding Na^+ cations and Cl^- anions.

4.2.2 Molecular dynamics simulation

Before running the MD simulation, the energy of the system was minimized in NAMD 2.10[118]. In the process of energy minimization, , 7000 steps of conjugate gradients (CG) energy minimization were performed for the solvent molecules with the coordinates of protein-ligand complex and all crystal water at fixed position, and then the whole system were minimized for 9000 steps applying the method of CG.

All simulations were preformed using NAMD 2.10-GPU and the CHARMM36 force field[119]. During each simulation, the system was first heated from 100 K to 300 K within 1.6 ns with a temperature increment of 0.25 K every 2000 steps. After equilibrating for another 6.4 ns, a production run of 172 ns was performed for the system. The time step

is 1 fs for heating and equilibration, and 2 fs for production run. All bonds between heavy atoms and hydrogen atoms were constrained through the SHAKE algorithm all through the simulations to avoid imprecise movement of the hydrogens in proteins. The particle Mesh Ewald (PME) method[120] with a grid spacing of 1.0 Å was applied to calculate the long range Coulombic interactions in the neutralized system. The van der Waals potential energy was smoothly truncated between 10.5 Å and 12.0 Å through a smooth switching function. For each MD simulation, the NPT ensemble was applied to the system. The constant temperature is maintained at 300K through Langevin dynamics[121], while the constant pressure is controlled to 1 atm using the Langevin piston Nosé-Hoover method[122].

4.2.3 Correlation and Principle Component Analysis

Dynamic cross-correlation matrix[152] between C-alpha atoms of all residues were calculated for the last 152 ns of the 180 ns MD simulation trajectory using the CHARMM c35b6 package. The equation for calculating the normalized covariance is:

$$C_{ij} = \frac{\langle \Delta\vec{r_i}(t) \cdot \Delta\vec{r_j}(t) \rangle}{\left(\langle \Delta\vec{r_i}(t)^2 \rangle \, \langle \Delta\vec{r_j}(t)^2 \rangle \right)^{1/2}} \qquad (4.1)$$

where $\Delta\vec{r_i}(t) = \vec{r_i}(t) - \langle \vec{r_i}(t) \rangle . \vec{r_i}(t)$ is the position vector of the C-alpha atom of the *i*th residue at time *t*. The bracket "⟨ ⟩" denote the time average of the value within the bracket. The correlations between C-alpha atoms are in the range from -1 to 1. The higher the absolute value, the stronger the correlation or anti-correlation is. Positive value (correlation) indicates that both residues move towards the same direction, and negative value (anti-

correlation) implies motions in opposite direction for most of the time. If the correlation between two residues is around zero, then they are uncorrelated.

Examining the distribution of the protein conformational space can help to understand the relationship between different structures. Due to the high dimension of the protein conformational space, it is impractical to directly examine the conformational space. Therefore, it is necessary to produce a lower dimensional representation of the structural dataset. Usually, 3-5 dimensions are sufficient to capture more than 70% of the total variance in a given family of structures. Principle component analysis (PCA)[153, 154] is a useful analysis that enable the projection of the high dimensional conformational space into the lower dimensional subspace. PCA provides a statistical analysis of molecular dynamics simulation trajectories, and has been applied for extracting the collective modes of displacement from MD trajectories. Through analyzing the mean-square displacement of all PCs, we are able to capture the non-harmonic motions of proteins.

The PCA calculation is performed with the Bio3D package. All C-alpha atoms of the protein is selected for calculation. A 3N dimensional covariance matrix, C-matrix, associating with the positional deviation of the selected set of atoms is constructed. The elements of C-matrix are defined as:

$$c_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \qquad (4.2)$$

where x_i and x_j are coordinates of atoms, and the brackets denotes the ensemble average.

We applied PCA to 180 ns MD simulations with the aim to identify correlated motions in the pfDHFR-TS protein.

4.3 Results and discussions

4.3.1 The domain-domain interaction in bifunctional DHFR-TS

To understand the domain-domain interactions in bifunctional DHFR-TS, correlated motions between C-alpha atoms of all residue were calculated by normalizing the cross correlation matrix of atomic fluctuation over the MD trajectory. The plot of the correlation matrix is shown in the left graph of Fig. 4.2.

There are strong correlations among residues in the DHFR domain. The Leu46 loop, which is important in facilitating the enzyme catalysis, is strongly correlated with residues in active site. The residues in active site are also correlated with each other. The N-terminal tail, including residues 1-5, is correlated with part of Insert 2 (residue 61~84), part of active site, and the αβ loop (residue 141~184). The above coupling indicates that except for local correlation, correlations also exit between distant residues.

Besides correlations within one single domain, there are cross domain correlated or anti-correlated motions. The two DHFR domains are strongly anti-correlated, while the two TS domains are mildly correlated. There are mainly three regions which anti-correlate with one another in the two DHFR domains. The three major regions include the residues 5~17, which is connecting N-terminal tail of DHFR, the Insert II residues, and the αβ loop. In each single DHFR domain, the three regions also correlated with each other as mentioned above. The results suggest that the motions of Insert II and the αβ loop have great dependence on the motion of residue 5. If residue 5 is deleted, motions of these residues may be disturbed, and changes in DHFR conformation may occur. It is still not clear about the implication of the communications between the two DHFR domains here.

The highly anti-correlated motion between the two DHFR domains may be necessary for the stabilization of the protein conformation and the regulation of signal transduction[155].

The domain-domain communication does not only exist in the pfDHFR-TS dimer, this behavior has been broadly studied previously[156]. In the HIV-1 protease dimer, the domain communications are suggested to be involved in functional energy transfer of the enzyme.

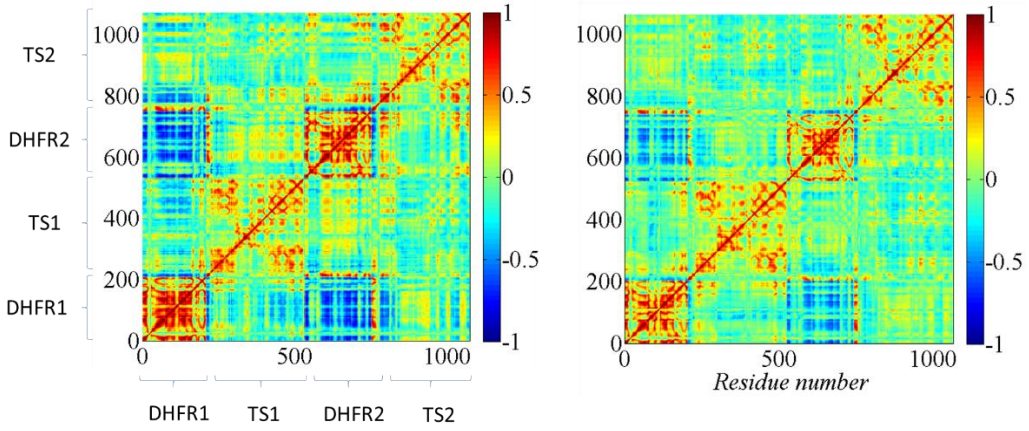

Fig 4.2. The domain-domain interaction in bifunctional protein DHFR-TS. The left figure is for the wild type. The plot on the right is for the mutant with the deletion of N-terminal tail in DHFR domains.

4.3.2. The role of N-terminal tail in modulating the domain-domain interactions

The deletion of N-terminal tail affects conformation of the dimer, which is important in modulating the domain-domain interaction. To understand the importance of N-terminal tail on the dimer conformation, the conformations before and after deleting

the N-terminal tail were studied. The distance between C-alpha atoms of asp54 residues and between C-alpha atoms of asp54 residues in both DHFR domains determines one conformation. As shown in the left graph of Fig. 4.3, the distance between C-alpha atoms of asp54 residues in both DHFR domains is highly correlated with the distance between C-alpha atoms of Ile164 residues in both DHFR domains. This result indicates that both distance increases or decreases at the same time, thus it implies that the two DHFR domains are moving towards opposite directions. This is also consistent with the results shown in figure 4.2, which shows that the motion of the two DHFR domains in DHFR-TS dimer is highly anti-correlated.

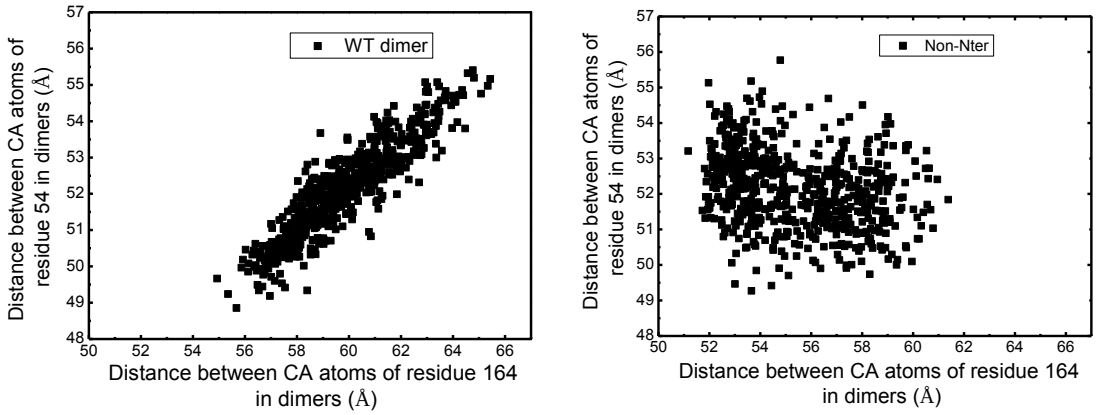

Fig 4.3. These two plots represent the relationship between the distance of the asp54 and the distance of ile164 in two DHFR domains. The left plot is for the wild type, and the right plot is for the mutant with the deletion of the N-terminal tail in pfDHFR.

Upon deleting the N-terminal tail, as shown in the right plot of Fig. 4.3, the distribution of the DHFR-DHFR domain conformation is changed, and there is no correlation between the two distance mentioned above, which indicates that the motion of neither the two asp54 residues nor the two ile164 residues in both DHFR domains are correlated or anti-correlated. These results are consistent with what are presented in figure 4.2. The anti-correlated motions between residue asp54 or between residue ile164 in both DHFR domains are much weaker or even disappear upon deleting the N-terminal tail in the DHFR domain. After the deletion of N-terminal tail, the motions of two DHFR domains are in general much less correlated.

N-Terminal tail is important in the interaction between the DHFR domain and the TS domain. Upon deleting the N-terminal tail, besides the decreased anti-correlated motions between the two DHFR domains in pfDHFR-TS dimer , the motion of the DHFR and TS domain in one monomer is also less anti-correlated. According to a previous experiment study[31], the binding of the ligand UMP in TS domain can affect the activity of DHFR. If the TS domain is bound with UMP ligand, the DHFR rate is almost doubled. This result implies that domain-domain communication may occur through active sites communications in both domains. As shown in the left plot of figure 4.2, within one monomer, the motion of the part of DHFR and TS domain are mildly anti-correlated with each other. Residue 470 in TS active site is anti-correlated with residue 108 in DHFR active site. However, upon the deletion of the N-terminal tail, the anti-correlated motion between DHFR and TS domain is significantly decreased, and there is no anti-correlation between residue 108 and residue 407. This result indicates that the N-terminal tail of DHFR domain

can affect the domain-domain communication, which may further affect activity in one single domain.

N-Terminal tail is important for the correlated motion within the DHFR domain. The experiment report[31] shows that the DHFR rate is decreased by 2-fold with the deletion of the N-terminal tail in DHFR. Since the Leu46 loop is important in facilitating the catalytic function of pfDHFR[26], it is necessary to examine the effect of the N-terminal tail on the motion of the Leu46 loop. As shown in figure 4.2, the correlated motion between the Leu46 loop and the active site residues is highly decreased in the mutant. For instance, the residue SER108 is strongly correlated with the entire Leu46 loop, as well as all other residues within the active site in the wild type, however, these correlations are all weakened upon the deletion of the N-terminal tail.

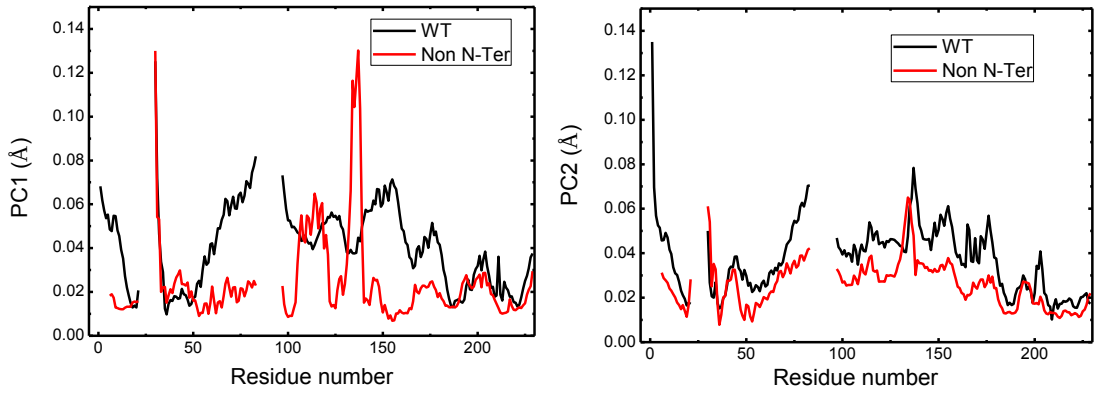

Fig 4.4. PCA of the wild type and the mutant. The left plot is for the first PC, and the right plot is for the second PC. The black line represents the wild type, while the red line represents the mutant with the deletion of the N-terminus.

We then applied PCA to further identify the correlated motions within the DHFR domain, and the result is shown in figure 4.4. According the analysis of the first principal component, the N-terminal tail, the Insert 2, the αβ loop, and the loop connecting residue 107 and 129 are highly correlated with each other, and all these correlated motions are greatly weakened when the N-terminal tail is deleted. Therefore, the N-terminal tail is crucial in maintaining the dynamic behavior of the DHFR domain. The deletion of N-terminal tail can lead to the decrease of DHFR activity through changing the dynamic motion of the Leu46 loop.

4.3.3 The role of N-terminal tail in maintaining the conformation of DHF binding pocket conformation

The conformation of the binding pocket is very important in maintaining the catalytic function of enzyme. Understanding the influence of N-terminal tail on conformation of the DHFR binding site can help to explain the effect of N-terminal tail on enzyme activity. To gain more insight into the effect of N-terminal tail on the DHFR active site, we examined the interactions in the DHF binding pocket. Figure 4.5 shows part of the DHF binding site, which presents the major change of this binding site upon the deletion of N-terminal tail. The left graph is for the wild type, while the right graph represents the

binding pocket of the mutant. In the wild type, both SER108 and LYS56 form stable hydrogen bond with the DHF ligand. However, upon deleting the N-terminal tail, the hydrogen bond between SER108 and DHF is broken, and LYS56 no longer has close contact with DHF.

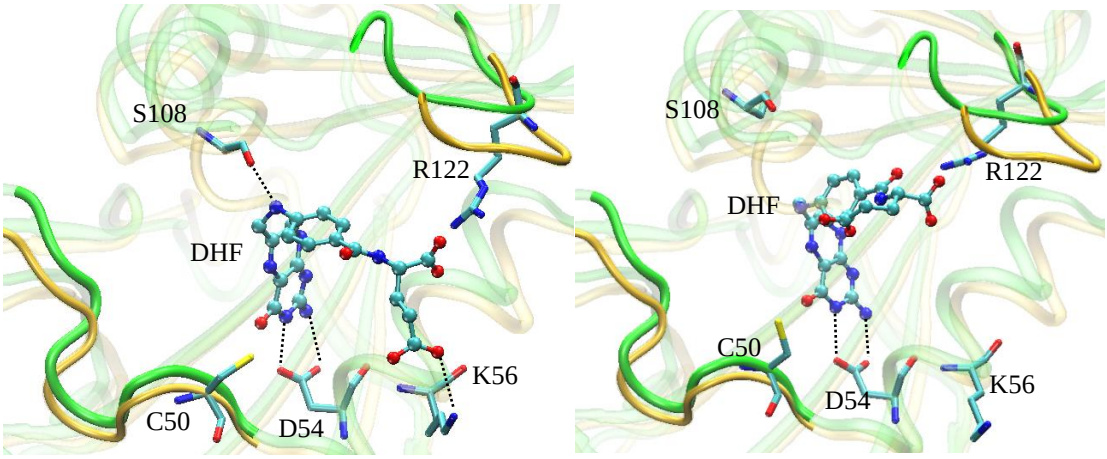

Fig 4.5. Part of the DHF binding site. The left figure is the binding site conformation for the wild type, and the right figure is for the mutant with the deletion of the N-terminal tail. The wild type is colored in yellow, and the mutant is in green.

The structural change of DHF binding pocket is related to the conformation change of DHF. According to figure 4.6, the atoms, which initially interacts with lys56 in the wild type, rotate, and move further away from lys56 in the mutant, and thus have quite weak interactions with this residue.

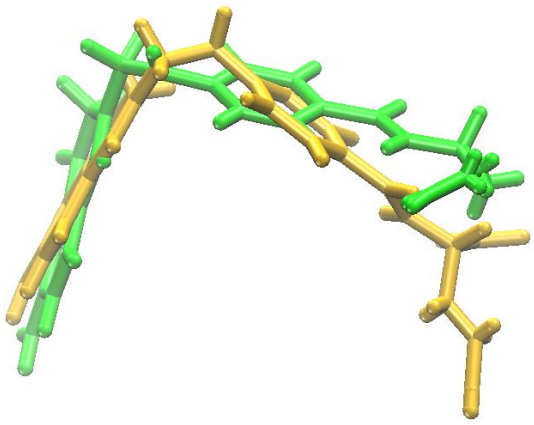

Fig 4.6. The averaged structure of DHF ligand obtained from MD simulation. The yellow structure represents the wild type, and the green one represents the mutant with the deletion of the N-terminal tail.

The conformation change of the ligand also cause significant change of the binding free energy between DHF and residue lys56, which increase by nearly 12 kcal/mol upon deleting the N-terminal tail, as shown in figure 4.7. The molecular mechanics/generalized Born surface area (MM/GBSA) calculation is applied to evaluate the interactions between the ligand DHF and each individual amino acid in the DHFR domain. Figure 4.7 shows the changes of the free energy between DHF and individual amino acid upon deleting the N-terminal tail. The result indicates that even the N-terminal tail is distant from the active site of DHFR, the deletion of this tail still cause significant change of binding free energy between DHF and individual residue in the binding pocket. We noticed that as the breaking of the hydrogen bond between DHF and the SER108 upon deleting the N-terminal tail, the binding free energy between them is increased. This result further supports that a remote

tail could perturb both the conformation and free energy of the active site in DHFR. SER108 plays a very important role in maintaining the activity of pfDHFR. Previous study has found that most mutations at position 108 could perturb the enzyme activity dramatically[107]. They substituted SER108 with all other 19 amino acid respectively, and found that only 9 mutants exhibited detectable activity. Even the active mutants, except for S108N, showed much poorer DHFR activity. Therefore, it is quite possible that any perturbation around SER108 could cause changes of DHFR activity.

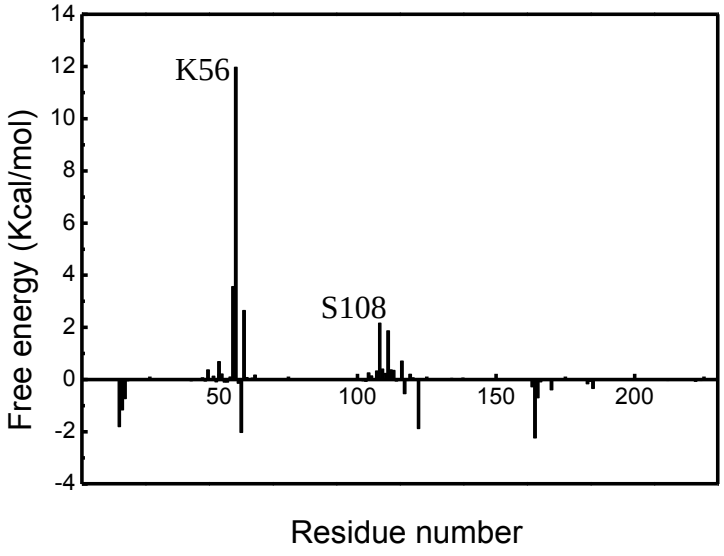

Figure 4.7 Changes of binding free energy between each residue and DHF upon deleting the N-terminal tail in the DHFR domain

4.4 Conclusion

This study provides evidence that the N-terminal tail plays an important role in modulating the domain-domain communications in pfDHFR-TS. The anti-correlated motions between the two DHFR domains, as well as the correlated motions between the DHFR domain and the TS domain are significantly weakened upon the deletion of the N-terminal tail. The communications between domains are likely to be important in stabilizing protein conformations.

The deletion of N-terminal tail also causes significant conformation changes in the DHFR active site. The local environment change around the residue SER108 may be an important reason for the decreased DHFR activity in the mutant, since SER108 is crucial for the functionally active DHFR enzyme. Also, the deletion of N-terminal tail can lead to changes in the dynamic motion of the Leu46 loop, which may further its motions in facilitating the DHFR catalytic function.

BIBLOGRAPHY

1. KA Dill and JL MacCallum. The Protein-Folding Problem, 50 Years On. Science 2012;338(6110):1042-1046.
2. V Daggett and A Fersht. The present view of the mechanism of protein folding. Nature Reviews Molecular Cell Biology 2003;4(6):497-502.
3. HA Scheraga, M Khalili and A Liwo. Protein-folding dynamics: Overview of molecular simulation techniques. Annu. Rev. Phys. Chem. 2007;58:57-83.
4. KA Dill. Dominant Forces in Protein Folding. Biochemistry 1990;29(31):7133-7155.
5. DS Marks, TA Hopf and C Sander. Protein structure prediction from sequence variation. Nat. Biotechnol. 2012;30(11):1072-1080.
6. JD Bryngelson, JN Onuchic, ND Socci and PG Wolynes. Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis. Proteins-Structure Function and Genetics 1995;21(3):167-195.
7. MK Gilson and HX Zhou. Calculation of protein-ligand binding affinities. Annu. Rev. Biophys. Biomol. Struct. 2007;36:21-42.
8. X Du, Y Li, YL Xia, SM Ai, J Liang, P Sang, XL Ji and SQ Liu. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci 2016;17(2).
9. E Gallicchio and RM Levy. Recent Theoretical and Computational Advances for Modeling Protein-Ligand Binding Affinities. Advances in Protein Chemistry and Structural Biology, Vol 85 2011;85:27-80.
10. DL Mobley and KA Dill. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". Structure 2009;17(4):489-498.
11. ATR Laurie and RM Jackson. Methods for the prediction of protein-ligand binding sites for Structure-Based Drug Design and virtual ligand screening. Curr. Protein Peptide Sci. 2006;7(5):395-406.
12. J Konc, S Lesnik and D Janezic. Modeling enzyme-ligand binding in drug discovery. Journal of Cheminformatics 2015;7.
13. BR Brooks, CL Brooks, AD Mackerell, L Nilsson, RJ Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, A Caflisch, L Caves, Q Cui, AR Dinner, M Feig, S Fischer, J Gao, M Hodoscek, W Im, K Kuczera, T Lazaridis, J Ma, V Ovchinnikov, E Paci, RW Pastor, CB Post, JZ Pu, M Schaefer, B Tidor, RM Venable, HL Woodcock, X Wu, W Yang, DM York and M Karplus. CHARMM: The Biomolecular Simulation Program. J. Comput. Chem. 2009;30(10):1545-1614.
14. DL Mobley, AE Barber, CJ Fennell and KA Dill. Charge asymmetries in hydration of polar solutes. J. Phys. Chem. B 2008;112(8):2405-2414.
15. J Chen and CL Brooks. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. PCCP 2008;10(4):471-481.
16. S Genheden, J Kongsted, P Soderhjelm and U Ryde. Nonpolar Solvation Free Energies of Protein-Ligand Complexes. Journal of Chemical Theory and Computation 2010;6(11):3558-3568.

17. KA Sharp and B Honig. Electrostatic Interactions in Macromolecules - Theory and Applications. Annu. Rev. Biophys. Biophys. Chem. 1990;19:301-332.
18. A Onufriev. The generalized Born model: its foundation, application, and limitation. 2010.
19. D Bashford and DA Case. Generalized born models of macromolecular solvation effects. Annu. Rev. Phys. Chem. 2000;51:129-152.
20. WC Still, A Tempczyk, RC Hawley and T Hendrickson. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. J. Am. Chem. Soc. 1990;112(16):6127-6129.
21. E Shakhnovich. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. Chem. Rev. 2006;106(5):1559-1588.
22. M Lorch, JM Mason, RB Sessions and AR Clarke. Effects of mutations on the thermodynamics of a protein folding reaction: Implications for the mechanism of formation of the intermediate and transition states. Biochemistry 2000;39(12):3480-3485.
23. SB Koukouritaki, MT Poch, MC Henderson, LK Siddens, SK Krueger, JE VanDyke, DE Williams, NM Pajewski, T Wang and RN Hines. Identification and functional analysis of common human flavin-containing monooxygenase 3 genetic variants. J. Pharmacol. Exp. Ther. 2007;320(1):266-273.
24. JE Shea and CL Brooks. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. Annu. Rev. Phys. Chem. 2001;52:499-535.
25. J Higo, J Ikebe, N Kamiya and H Nakamura. Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. Biophys Rev 2012;4(1):27-44.
26. J Yuvaniyama, P Chitnumsub, S Kamchonwongpaisan, J Vanichtanankul, W Sirawaraporn, P Taylor, MD Walkinshaw and Y Yuthavong. Insights into antifolate resistance from malarial DHFR-TS structures. Nat. Struct. Biol. 2003;10(5):357-365.
27. SJ Foote, D Galatis and AF Cowman. Amino-Acids in the Dihydrofolate Reductase-Thymidylate Synthase Gene of Plasmodium-Falciparum Involved in Cycloguanil Resistance Differ from Those Involved in Pyrimethamine Resistance. Proc. Natl. Acad. Sci. U. S. A. 1990;87(8):3014-3017.
28. GB Fogel, M Cheung, E Pittman and D Hecht. In silico screening against wild-type and mutant Plasmodium falciparum dihydrofolate reductase. J. Mol. Graph. Model. 2008;26(7):1145-1152.
29. P Maitarad, P Saparpakorn, S Hannongbua, S Kamchonwongpaisan, B Tarnchompoo and Y Yuthavong. Particular interaction between pyrimethamine derivatives and quadruple mutant type dihydrofolate reductase of Plasmodium falciparum: CoMFA and quantum chemical calculations studies. J. Enzyme Inhib. Med. Chem. 2009;24(2):471-479.
30. OA Santos-Filho, RB de Alencastro and JD Figueroa-Villar. Homology modeling of wild type and pyrimethamine/cycloguanil-cross resistant mutant type

Plasmodium falciparum dihydrofolate reductase. A model for antimalarial chemotherapy resistance. Biophys. Chem. 2001;91(3):305-317.
31. T Dasgupta and KS Anderson. Probing the role of parasite-specific, distant structural regions on communication and catalysis in the bifunctional thymidylate synthase-dihydrofolate reductase from Plasmodium falciparum. Biochemistry 2008;47(5):1336-1345.
32. S Shallom, K Zhang, L Jiang and PK Rathod. Essential protein-protein interactions between Plasmodium falciparum thymidylate synthase and dihydrofolate reductase domains. J. Biol. Chem. 1999;274(53):37781-37786.
33. PA BASH, UC SINGH, R LANGRIDGE and PA KOLLMAN. free energy calculations by computer simulation.pdf. Science 1987;236(4081):564-568.
34. BTaM Karplus. Simulation Analysis of the Stability Mutant R96H of T4 Lysozyme+.pd. Biochemistry 1991;30.
35. YN Vorobjev and J Hermans. ES/IS: Estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. Biophys. Chem. 1999;78(1-2):195-205.
36. JW Pitera and PA Kollman. Exhaustive mutagenesis in silico: Multicoordinate free energy calculations on proteins and peptides. Proteins: Structure Function and Genetics 2000;41(3):385-397.
37. S Piana, A Laio, F Marinelli, M Van Troys, D Bourry, C Ampe and JC Martins. Predicting the effect of a point mutation on a protein fold: the villin and advillin headpieces and their Pro62Ala mutants. J. Mol. Biol. 2008;375(2):460-470.
38. PD Thomas and KA Dill. Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol. 1996;257(2):457-469.
39. D Gilis and M Rooman. Predicting protein stability changes upon mutation using database-derived potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence. J. Mol. Biol. 1997;272(2):276-290.
40. C Hoppe and D Schomburg. Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. Protein Sci. 2005;14(10):2682-2692.
41. G Wainreb, L Wolf, H Ashkenazy, Y Dehouck and N Ben-Tal. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics 2011;27(23):3286-3292.
42. AJ Bordner and RA Abagyan. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 2004;57(2):400-413.
43. N Tokuriki, F Stricher, J Schymkowitz, L Serrano and DS Tawfik. The stability effects of protein mutations appear to be universally distributed. J. Mol. Biol. 2007;369(5):1318-1332.
44. L Wickstrom, E Gallicchio and RM Levy. The linear interaction energy method for the prediction of protein stability changes upon mutation. Proteins 2012;80(1):111-125.

45. R Guerois, JE Nielsen and L Serrano. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. J. Mol. Biol. 2002;320(2):369-387.
46. SY Yin, F Ding and NV Dokholyan. Eris: an automated estimator of protein stability. Nat. Methods 2007;4(6):466-467.
47. A Benedix, CM Becker, BL de Groot, A Caflisch and RA Bockmann. Predicting free energy changes using structural ensembles. Nat. Methods 2009;6(1):3-4.
48. I Getov, M Petukh and E Alexov. SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. Int J Mol Sci 2016;17(4).
49. JB Clarage, T Romo, BK Andrews, BM Pettitt and GN Phillips, Jr. A sampling problem in molecular dynamics simulations of macromolecules. Proc. Natl. Acad. Sci. U. S. A. 1995;92(8):3288-3292.
50. M Lei, MI Zavodszky, LA Kuhn and MF Thorpe. Sampling protein conformations and pathways. J. Comput. Chem. 2004;25(9):1133-1148.
51. EH Kellogg, A Leaver-Fay and D Baker. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins-Structure Function and Bioinformatics 2011;79(3):830-838.
52. BL deGroot, DMF vanAalten, RM Scheek, A Amadei, G Vriend and HJC Berendsen. Prediction of protein conformational freedom from distance constraints. Proteins-Structure Function and Genetics 1997;29(2):240-251.
53. LH Weaver and BW Matthews. Structure of Bacteriophage-T4 Lysozyme Refined at 1.7 a Resolution. J. Mol. Biol. 1987;193(1):189-199.
54. R Diamond. Real-space refinement of the structure of hen egg-white lysozyme. J. Mol. Biol. 1974;82(3):371-391.
55. TR Hynes and RO Fox. The Crystal-Structure of Staphylococcal Nuclease Refined at 1.7 a Resolution. Proteins 1991;10(2):92-105.
56. A Pastore, V Saudek, G Ramponi and RJ Williams. Three-dimensional structure of acylphosphatase. Refinement and structure analysis. J. Mol. Biol. 1992;224(2):427-440.
57. H Schindelin, MA Marahiel and U Heinemann. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. Nature 1993;364(6433):164-168.
58. T Gallagher, P Alexander, P Bryan and GL Gilliland. 2 Crystal-Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein-G and Comparison with Nmr. Biochemistry 1994;33(15):4721-4729.
59. Y Harpaz, N Elmasry, AR Fersht and K Henrick. Direct Observation of Better Hydration at the N-Terminus of an Alpha-Helix with Glycine Rather Than Alanine as the N-Cap Residue. Proc. Natl. Acad. Sci. U. S. A. 1994;91(1):311-315.
60. MM Skinner, H Zhang, DH Leschnitzer, Y Guan, H Bellamy, RM Sweet, CW Gray, RNH Konings, AHJ Wang and TC Terwilliger. Structure of the Gene-V Protein of Bacteriophage-F1 Determined by Multiwavelength X-Ray-Diffraction on the Selenomethionyl Protein. Proc. Natl. Acad. Sci. U. S. A. 1994;91(6):2071-2075.

61. CA Dennis, H Videler, RA Pauptit, R Wallis, R James, GR Moore and C Kleanthous. A structural comparison of the colicin immunity proteins Im7 and Im9 gives new insights into the molecular determinants of immunity-protein specificity. Biochem. J 1998;333:183-191.
62. N Eswar, B Webb, MA Marti-Renom, MS Madhusudhan, D Eramian, MY Shen, U Pieper and A Sali. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 2006;Chapter 5:Unit 5 6.
63. MD Kumar, KA Bava, MM Gromiha, P Prabakaran, K Kitajima, H Uedaira and A Sarai. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34(Database issue):D204-206.
64. D Seeliger and BL de Groot. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. Biophys. J. 2010;98(10):2309-2316.
65. DV Mikhailov, L Washington, AM Voloshin, VA Daragan and KH Mayo. Angular variances for internal bond rotations of side chains in GXG-based tripeptides derived from (13)C-NMR relaxation measurements: Implications to protein folding. Biopolymers 1999;49(5):373-383.
66. F Fogolari, A Brigo and H Molinari. Protocol for MM/PBSA molecular dynamics simulations of proteins. Biophys. J. 2003;85(1):159-166.
67. M Feig, J Karanicolas and CL Brooks. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model. 2004;22(5):377-395.
68. D Sitkoff, KA Sharp and B Honig. Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models. J. Phys. Chem. 1994;98(7):1978-1988.
69. T Alber, DP Sun, K Wilson, JA Wozniak, SP Cook and BW Matthews. Contributions of Hydrogen-Bonds of Thr-157 to the Thermodynamic Stability of Phage-T4 Lysozyme. Nature 1987;330(6143):41-46.
70. M Matsumura, WJ Becktel and BW Matthews. Hydrophobic Stabilization in T4 Lysozyme Determined Directly by Multiple Substitutions of Ile-3. Nature 1988;334(6181):406-410.
71. D Shortle, WE Stites and AK Meeker. Contributions of the Large Hydrophobic Amino-Acids to the Stability of Staphylococcal Nuclease. Biochemistry 1990;29(35):8033-8041.
72. A Tanaka, J Flanagan and JM Sturtevant. Thermal Unfolding of Staphylococcal Nuclease and Several Mutant Forms Thereof Studied by Differential Scanning Calorimetry. Protein Sci. 1993;2(4):567-576.
73. LS Itzhaki, DE Otzen and AR Fersht. The Structure of the Transition-State for Folding of Chymotrypsin Inhibitor-2 Analyzed by Protein Engineering Methods - Evidence for a Nucleation-Condensation Mechanism for Protein-Folding. J. Mol. Biol. 1995;254(2):260-288.
74. EL McCallister, E Alm and D Baker. Critical role of beta-hairpin formation in protein G folding. Nat. Struct. Biol. 2000;7(8):669-673.
75. AP Capaldi, C Kleanthous and SE Radford. Im7 folding mechanism: misfolding on a path to the native state. Nat. Struct. Biol. 2002;9(3):209-216.

76. R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. Science 1987;235(4786):318-321.
77. RH Austin, KW Beeson, L Eisenstein, H Frauenfelder and IC Gunsalus. Dynamics of Ligand-Binding to Myoglobin. Biochemistry 1975;14(24):5355-5373.
78. K Henzler-Wildman and D Kern. Dynamic personalities of proteins. Nature 2007;450(7172):964-972.
79. CN Schutz and A Warshel. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001;44(4):400-417.
80. E Demchuk and RC Wade. Improving the Continuum Dielectric Approach to Calculating pKas of Ionizable Groups in Proteins. The Journal of Physical Chemistry 1996;100(43):17373-17387.
81. T Simonson and D Perahia. Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. Proc. Natl. Acad. Sci. U. S. A. 1995;92(4):1082-1086.
82. T Simonson and CL Brooks. Charge screening and the dielectric constant of proteins: Insights from molecular dynamics. J. Am. Chem. Soc. 1996;118(35):8452-8458.
83. T Simonson and D Perahia. Microscopic Dielectric-Properties of Cytochrome-C from Molecular-Dynamics Simulations in Aqueous-Solution. J. Am. Chem. Soc. 1995;117(30):7987-8000.
84. T Simonson and D Perahia. Polar fluctuations in proteins: Molecular-dynamic studies of cytochrome c in aqueous solution. Faraday Discuss. 1996;103:71-90.
85. MK Gilson, JA Given, BL Bush and JA McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. Biophys. J. 1997;72(3):1047-1069.
86. J Srinivasan, TE Cheatham, P Cieplak, PA Kollman and DA Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. J. Am. Chem. Soc. 1998;120(37):9401-9409.
87. RHH Jessica M. J. Swanson, * and J. Andrew McCammon*. <Revisiting free energy calculations- a theoretical connection to MMPBSA and direct calculation of the association free energy.pdf>. Biophys. J. 2004;86.
88. AR Brice and BN Dominy. Analyzing the Robustness of the MM/PBSA Free Energy Calculation Method: Application to DNA Conformational Transitions. J. Comput. Chem. 2011;32(7):1431-1440.
89. H Tzoupis, G Leonis, T Mavromoustakos and MG Papadopoulos. A Comparative Molecular Dynamics, MM-PBSA and Thermodynamic Integration Study of Saquinavir Complexes with Wild-Type HIV-1 PR and L10I, G48V, L63P, A71V, G73S, V82A and I84V Single Mutants. Journal of Chemical Theory and Computation 2013;9(3):1754-1764.
90. SA Martins, MAS Perez, IS Moreira, SF Sousa, MJ Ramos and PA Fernandes. Computational Alanine Scanning Mutagenesis: MM-PBSA vs TI. Journal of Chemical Theory and Computation 2013;9(3):1311-1319.

91. I Bea, E Cervello, PA Kollman and C Jaime. Molecular recognition by beta-cyclodextrin derivatives: FEP vs MM/PBSA study. Combinatorial Chem. High Throughput Screening 2001;4(8):605-611.
92. YQ Zhou, D Vitkup and M Karplus. Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. J. Mol. Biol. 1999;285(4):1371-1375.
93. BW Matthews, H Nicholson and WJ Becktel. Enhanced Protein Thermostability from Site-Directed Mutations That Decrease the Entropy of Unfolding. Proc. Natl. Acad. Sci. U. S. A. 1987;84(19):6663-6667.
94. MJ Stone. NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. Acc. Chem. Res. 2001;34(5):379-388.
95. AJ Doig and MJE Sternberg. Side-Chain Conformational Entropy in Protein-Folding. Protein Sci. 1995;4(11):2247-2251.
96. SK Sadiq, DW Wright, OA Kenway and PV Coveney. Accurate Ensemble Molecular Dynamics Binding Free Energy Ranking of Multidrug-Resistant HIV-1 Proteases. Journal of Chemical Information and Modeling 2010;50(5):890-905.
97. SZ Wan and PV Coveney. Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. Journal of the Royal Society Interface 2011;8(61):1114-1127.
98. DW Wright, BA Hall, OA Kenway, S Jha and PV Coveney. Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. J Chem Theory Comput 2014;10(3):1228-1241.
99. Y Levin. Electrostatic correlations: from plasma to biology. Rep. Prog. Phys. 2002;65(11):1577-1632.
100. T Lemcke, IT Christensen and FS Jorgensen. Towards an understanding of drug resistance in malaria: Three-dimensional structure of Plasmodium falciparum dihydrofolate reductase by homology building. Biorg. Med. Chem. 1999;7(6):1003-1011.
101. G Rastelli, W Sirawaraporn, P Sompornpisut, T Vilaivan, S Kamchonwongpaisan, R Quarrell, G Lowe, Y Thebtaranonth and Y Yuthavong. Interaction of pyrimethamine, cycloguanil, WR99210 and their analogues with Plasmodium falciparum dihydrofolate reductase: Structural basis of antifolate resistance. Biorg. Med. Chem. 2000;8(5):1117-1128.
102. S Chusacultanachai, P Thiensathit, B Tarnchompoo, W Sirawaraporn and Y Yuthavong. Novel antifolate resistant mutations of Plasmodium falciparum dihydrofolate reductase selected in Escherichia coli. Mol. Biochem. Parasitol. 2002;120(1):61-72.
103. N Gonen and YG Assaraf. Antifolates in cancer therapy: Structure, activity and mechanisms of drug resistance. Drug Resistance Updates 2012;15(4):183-210.
104. JF Cortese and CV Plowe. Antifolate resistance due to new and known Plasmodium falciparum dihydrofolate reductase mutations expressed in yeast. Mol. Biochem. Parasitol. 1998;94(2):205-214.

105. W Sirawaraporn, T Sathitkul, R Sirawaraporn, Y Yuthavong and DV Santi. Antifolate-resistant mutants of Plasmodium falciparum dihydrofolate reductase. Proc. Natl. Acad. Sci. U. S. A. 1997;94(4):1124-1129.
106. Y Yuthavong, J Yuvaniyama, P Chitnumsub, J Vanichtanankul, S Chusacultanachai, B Tarnchompoo, T Vilaivan and S Kamchonwongpaisan. Malarial (Plasmodium falciparum) dihydrofolate reductase-thymidylate synthase: structural basis for antifolate resistance and development of effective inhibitors. Parasitology 2005;130:249-259.
107. W Sirawaraporn, S Yongkiettrakul, R Sirawaraporn, Y Yuthavong and DV Santi. Plasmodium falciparum: Asparagine mutant at residue 108 of dihydrofolate reductase is an optimal antifolate-resistant single mutant. Exp. Parasitol. 1997;87(3):245-252.
108. Y Yuthavong, T Vilaivan, N Chareonsethakul, S Kamchonwongpaisan, W Sirawaraporn, R Quarrell and G Lowe. Development of a lead inhibitor for the A16V+S108T mutant of dihydrofolate reductase from the cycloguanil-resistant strain (T9/94) of Plasmodium falciparum. J. Med. Chem. 2000;43(14):2738-2744.
109. KM Brown, MS Costanzo, WX Xu, S Roy, ER Lozovsky and DL Hartl. Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase. Mol. Biol. Evol. 2010;27(12):2682-2690.
110. RT Delfino, OA Santos and JD Figueroa-Villar. Molecular modeling of wild-type and antifolate resistant mutant Plasmodium falciparum DHFR. Biophys. Chem. 2002;98(3):287-300.
111. GB Fogel, M Cheung, E Pittman and D Hecht. Modeling the inhibition of quadruple mutant Plasmodium falciparum dihydrofolate reductase by pyrimethamine derivatives. J. Comput. Aided Mol. Des. 2008;22(1):29-38.
112. D Hecht and GB Fogel. Modeling the evolution of drug resistance in malaria. J. Comput. Aided Mol. Des. 2012;26(12):1343-1353.
113. W Mokmak, S Chunsrivirot, S Hannongbua, Y Yuthavong, S Tongsima and S Kamchonwongpaisan. Molecular Dynamics of Interactions between Rigid and Flexible Antifolates and Dihydrofolate Reductase from Pyrimethamine-Sensitive and Pyrimethamine-Resistant Plasmodium falciparum. Chem. Biol. Drug Des. 2014;84(4):450-461.
114. J Vanichtanankul, S Taweechai, J Yuvaniyama, T Vilaivan, P Chitnumsub, S Kamchonwongpaisan and Y Yuthavong. Trypanosomal Dihydrofolate Reductase Reveals Natural Antifolate Resistance. ACS Chem. Biol. 2011;6(9):905-911.
115. L Cocco, B Roth, C Temple, Jr., JA Montgomery, RE London and RL Blakley. Protonated state of methotrexate, trimethoprim, and pyrimethamine bound to dihydrofolate reductase. Arch. Biochem. Biophys. 1983;226(2):567-577.
116. EE Howell, JE Villafranca, MS Warren, SJ Oatley and J Kraut. Functional role of aspartic acid-27 in dihydrofolate reductase revealed by mutagenesis. Science 1986;231(4742):1123-1128.
117. K Vanommeslaeghe, E Hatcher, C Acharya, S Kundu, S Zhong, J Shim, E Darian, O Guvench, P Lopes, I Vorobyov and AD Mackerell, Jr. CHARMM general force

field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J. Comput. Chem. 2010;31(4):671-690.
118. JC Phillips, R Braun, W Wang, J Gumbart, E Tajkhorshid, E Villa, C Chipot, RD Skeel, L Kale and K Schulten. Scalable molecular dynamics with NAMD. J. Comput. Chem. 2005;26(16):1781-1802.
119. RB Best, X Zhu, J Shim, PEM Lopes, J Mittal, M Feig and AD MacKerell. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. Journal of Chemical Theory and Computation 2012;8(9):3257-3273.
120. T Darden, D York and L Pedersen. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J. Chem. Phys. 1993;98(12):10089-10092.
121. P Hunenberger. Thermostat algorithms for molecular dynamics simulations. Advanced Computer Simulation Approaches for Soft Matter Sciences I 2005;173:105-147.
122. SE Feller, YH Zhang, RW Pastor and BR Brooks. Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method. J. Chem. Phys. 1995;103(11):4613-4621.
123. W Im, M Feig and CL Brooks. An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. Biophys. J. 2003;85(5):2900-2918.
124. D Bashford and DA Case. Generalized born models of macromolecular solvation effects. Annu. Rev. Phys. Chem. 2000;51:129-152.
125. D Sitkoff, KA Sharp and B Honig. Correlating solvation free energies and surface tensions of hydrocarbon solutes. Biophys. Chem. 1994;51(2-3):397-403; discussion 404-399.
126. YB Yu, PL Privalov and RS Hodges. Contribution of translational and rotational motions to molecular association in aqueous solution. Biophys. J. 2001;81(3):1632-1642.
127. A Sethi, J Eargle, AA Black and Z Luthey-Schulten. Dynamical networks in tRNA: protein complexes. Proc. Natl. Acad. Sci. U. S. A. 2009;106(16):6620-6625.
128. W Humphrey, A Dalke and K Schulten. VMD: Visual molecular dynamics. J. Mol. Graph. Model. 1996;14(1):33-38.
129. A Sethi, J Tian, CA Derdeyn, B Korber and S Gnanakaran. A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein. PLoS Comput. Biol. 2013;9(5):e1003046.
130. M Girvan and MEJ Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. U. S. A. 2002;99(12):7821-7826.
131. S Huo, I Massova and PA Kollman. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J. Comput. Chem. 2002;23(1):15-27.
132. I Massova and PA Kollman. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. J. Am. Chem. Soc. 1999;121(36):8133-8143.

133. W Sirawaraporn, R Sirawaraporn, S Yongkiettrakul, A Anuwatwora, G Rastelli, S Kamchonwongpaisan and Y Yuthavong. Mutational analysis of Plasmodium falciparum dihydrofolate reductase: the role of aspartate 54 and phenylalanine 223 on catalytic activity and antifolate binding. Mol. Biochem. Parasitol. 2002;121(2):185-193.
134. GA Jeffrey. An Introduction to Hydrogen Bonding. 1997.
135. K Teilum, JG Olsen and BB Kragelund. Protein stability, flexibility and function. Biochim. Biophys. Acta 2011;1814(8):969-976.
136. JL Radkiewicz and CL Brooks. Protein dynamics in enzymatic catalysis: Exploration of dihydrofolate reductase. J. Am. Chem. Soc. 2000;122(2):225-231.
137. A Tousignant and JN Pelletier. Protein motions promote catalysis. Chem. Biol. 2004;11(8):1037-1042.
138. N Popovych, SJ Sun, RH Ebright and CG Kalodimos. Dynamically driven protein allostery. Nat. Struct. Mol. Biol. 2006;13(9):831-838.
139. H Pan, JC Lee and VJ Hilser. Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. Proc. Natl. Acad. Sci. U. S. A. 2000;97(22):12020-12025.
140. AJ Wand. Dynamic activation of protein function: A view emerging from NMR spectroscopy. Nat. Struct. Biol. 2001;8(11):926-931.
141. SW Homans. Probing the binding entropy of ligand-protein interactions by NMR. ChemBioChem 2005;6(9):1585-+.
142. R Grünberg, M Nilges and J Leckner. Flexibility and Conformational Entropy in Protein-Protein Binding. Structure 2006;14(4):683-693.
143. C Diehl, O Engstrom, T Delaine, M Hakansson, S Genheden, K Modig, H Leffler, U Ryde, UJ Nilsson and M Akke. Protein flexibility and conformational entropy in ligand design targeting the carbohydrate recognition domain of galectin-3. J. Am. Chem. Soc. 2010;132(41):14577-14589.
144. KS Crider, TP Yang, RJ Berry and LB Bailey. Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role. Advances in Nutrition 2012;3(1):21-38.
145. PJ Houghton, GS Germain, BJ Hazelton, JW Pennington and JA Houghton. Mutants of human colon adenocarcinoma, selected for thymidylate synthase deficiency. Proc. Natl. Acad. Sci. U. S. A. 1989;86(4):1377-1381.
146. DJ Bzik, WB Li, T Horii and J Inselburg. Molecular-Cloning and Sequence-Analysis of the Plasmodium-Falciparum Dihydrofolate-Reductase Thymidylate Synthase Gene. Proc. Natl. Acad. Sci. U. S. A. 1987;84(23):8360-8364.
147. D Japrung, S Chusacultanachai, J Yuvaniyama, P Wilairat and Y Yuthavong. A simple dual selection for functionally active mutants of Plasmodium falciparum dihydrofolate reductase with improved solubility. Protein Engineering Design & Selection 2005;18(10):457-464.
148. M Chanama, P Chitnumsub and Y Yuthavong. Subunit complementation of thymidylate synthase in Plasmodium falciparum bifunctional dihydrofolate reductase-thymidylate synthase. Mol. Biochem. Parasitol. 2005;139(1):83-90.

149. DR Knighton, CC Kan, E Howland, CA Janson, Z Hostomska, KM Welsch and DA Matthews. Structure of and Kinetic Channeling in Bifunctional Dihydrofolate Reductase-Thymidylate Synthase. Nat. Struct. Biol. 1994;1(3):186-194.
150. J Wattanarangsan, S Chusacultanachai, J Yuvaniyama, S Kamchonwongpaisan and Y Yuthavong. Effect of N-terminal truncation of Plasmodium falciparum dihydrofolate reductase on dihydrofolate reductase and thymidylate synthase activity. Mol. Biochem. Parasitol. 2003;126(1):97-102.
151. Y Yuthavong, B Tarnchompoo, T Vilaivan, P Chitnumsub, S Kamchonwongpaisan, SA Charman, DN McLennan, KL White, L Vivas, E Bongard, C Thongphanchang, S Taweechai, J Vanichtanankul, R Rattanajak, U Arwon, P Fantauzzi, J Yuvaniyama, WN Charman and D Matthews. Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. Proc. Natl. Acad. Sci. U. S. A. 2012;109(42):16823-16828.
152. T Ichiye and M Karplus. Collective Motions in Proteins - a Covariance Analysis of Atomic Fluctuations in Molecular-Dynamics and Normal Mode Simulations. Proteins-Structure Function and Genetics 1991;11(3):205-217.
153. MA Balsera, W Wriggers, Y Oono and K Schulten. Principal component analysis and long time protein dynamics. J. Phys. Chem. 1996;100(7):2567-2572.
154. CC David and DJ Jacobs. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. Protein Dynamics: Methods and Protocols 2014;1084:193-226.
155. JD Klemm, SL Schreiber and GR Crabtree. Dimerization as a regulatory mechanism in signal transduction. Annu. Rev. Immunol. 1998;16:569-592.
156. WE Harte, S Swaminathan, MM Mansuri, JC Martin, IE Rosenberg and DL Beveridge. Domain Communication in the Dynamic Structure of Human Immunodeficiency Virus-1 Protease. Proc. Natl. Acad. Sci. U. S. A. 1990;87(22):8864-8868.