Development and Application of the Free Energy Based Screening Methods

A Thesis

Presented to the Department of Functional Molecular Science

School of Mathematical and Physical Science

The Graduate University for Advanced Studies

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Science

by

Shinichi Banba

September 2002

Acknowledgements

Directory or indirectly, many have contributed to the completion of this thesis and it gives me great pleasure to mention them here.

My most heartfelt thanks go to my thesis advisors, Professors Charles L. Brooks III and Yuko Okamoto, for their patient guidance and constant encouragement. I believe that this thesis would never have been completed without their advice and suggestions. Dr. Brooks has also brought up my comprehension and techniques for research in the field of computational chemistry.

I would like to thank Drs. Zhuyan Guo and Ryuichi Kiyama for valuable and helpful discussions and suggestions. I am greatly indebted to Dr. Komath V. Damodaram for valuable and helpful discussions and suggestions. It is notable that he patiently removed my grammatical errors from my reports. I wish to express my gratitude to all the member of Brooks Group for stimulating discussions and encouragement. I would like to thank Drs. Tom Cleveland, Hervé Minoux, and Ryan Morton for making the laboratory to be a pleasant place to work in. I am also deeply indebted to Dr. Tom Cleveland for instructing me how to use CHARMM. I am grateful to Ms. JoAnn Meyer for guiding my family towards a better life in San Diego.

I wish to express my gratitude to Mitsui Chemical, Inc., especially to Drs. Kazushi Ohshima, Kunio Sannohe, and Eishi Tanaka, for providing me with an opportunity to visit The Scripps Research Institute and Institute for Molecular Science.

Finally, I give very special thanks to my parents Masakazu and Shigeko Banba, my wife Kazuyo Banba, and my children Hiroki, Nagomi, and Sota Banba for their constant encouragement and support.

Contents

Chapter 1

General Introduction

The design of molecules that bind tightly and specifically to a target protein or host is an important goal. Since three-dimensional structures of new proteins or protein-ligand complexes are becoming available at a dramatically increasing rate, computer modeling methods should play a key role in structure-based drug design. The structures developed from structural genomic efforts will also provide potential targets for drug design and thus provide an even greater impetus for computational approach to rational drug design. Recently, several examples have been documented where structure-based molecular design has led to the development of drug candidates.[1-3] Despite recent advances in both methodologies and computational power, identification of ligands with high binding affinity, or low binding free energy, using structure based approaches remains a challenge. Several methods, with differing levels of accuracy and computational cost, have been developed and applied to biological systems in recent years.[4-10] Empirical methods such as DOCK[9,10] rank ligands based on shape and chemical complementarity. The simple scoring scheme used in DOCK enables this approach to rank ligands rapidly. Therefore it has become a routine tool for rapid 3D-database searches in drug discovery. In a DOCK search, a large 'free energy' cutoff value is normally used in order not to miss any promising compounds, and consequently a large number of DOCK hits are generated. To funnel down those hits to a manageable number is still a challenging task in drug discovery, and a more accurate but rapid method is needed to this end. Semi-empirical methods, such as the Linear Interaction Approximation (LIA) method,[11] give more accurate results but at higher computational cost. LIA requires the calculation of the difference in average interaction energy between the ligand in the unbound state and in the bound state. The scaling coefficient used in this method varies from system to system. Therefore, a large number of ligands with known binding affinity have to be used to obtain the coefficients. It is well known that the thermodynamic-cycle methods, such as free energy perturbation (FEP)[12] or thermodynamic integration (TI),[4] ideally give an accurate measure of the binding free energy (assuming the force field used is accurate and the sampling is complete). However they are too computationally intensive to be routinely used in drug design. Recently, "free energy based screening methods" as known by the λ-dynamics method[6,13,14] and Chemical Monte Carlo / molecular dynamics (CMC/MD) method[15,16] have been developed for rapid evaluation of the binding free energies of a large number of ligands. These methods are more efficient than FEP or TI, especially when multiple ligands are investigated. It has been shown previously that the λ-dynamics method successfully discriminates the good binders from the bad ones for benzamidine-based inhibitors complexed with trypsin.[6] In that application, modification of the ligands was localized to the para-position and then the perturbation was relatively small. The results from that study compared well with FEP calculations. Encouraged by these promising findings, in this thesis, we further attempt to make free energy based methods more efficient and practical tools for computer-aided drug and protein design. These enhancements consist of the use of the umbrella sampling techniques for efficient sampling of the chemical coordinates, a restraining potential for the multiple topology representation, the incorporation of a continuum solvation model into these methods, λ-dependent partial charge models for the hybrid topology representation, and an efficient sampling on a free energy surface.

In this thesis, I will describe our development and application of "free energy based screening methods." In **Chapter 2**, I first briefly explain the background and basic formulation of the conventional free energy calculation methods and then present an overview of the recently developed computational methods for ligand screening. The details of "free energy based screening methods" are given in **Section 2.3** together with several enhancements to these methods.

Next, the applications of “free energy based screening methods” will follow. In **Chapter 3**, the λ-dynamics method using a multiple topology model is applied to a set of 10 five-member ring heterocycle derivatives interacting with cytochrome c peroxidase, highlighting the effectiveness of a newly introduced restraining potential and its ability to explore binding orientations or conformations on a free energy basis. In **Chapter 4**, I describe the λ-dynamics and CMC/MD methods with the generalized Born (GB) implicit solvation model to calculate the relative binding free energies for four benzamidine derivatives binding to trypsin. In **Chapter 5**, a β-cyclodextrin-benzene derivative system is studied to compare the GB model with the use of an explicit water model, as well as to evaluate a newly introduced λ-dependent partial charge model for hybrid topology λ-dynamics simulation. In **Chapter 6**, both λ-dynamics and CMC/MD methods are applied to the stability analysis of the DNA-binding domain of the c-Myb transcriptional regulator. In **Chapter 7**, the results of the application of a modified version of MC/MD to explore the binding orientations of toluene in β-cyclodextrin are presented. Finally, a Summary and outlook is given. In Appendices, problems in the analytical GB implicit solvation model are described and their temporary solutions are presented.

Chapter 2

Methodologies

2.1 Conventional free energy methods

2.1.1 Background

All thermodynamic properties can be obtained from knowledge of free energy and its derivatives. Thus, one should focus on the free energy of the molecular system when aiming to quantitatively predict or rationalize the interactions between the putative inhibitors and a receptor. In this thesis, I frame my discussion of computational approaches to calculating free energies in the canonical ensemble. In this ensemble the Helmholtz free energy, which I denote as *G* throughout this thesis, is the appropriate thermodynamic potential and is given by the following configurational integral of the Boltzmann factor.

$$G = -k_B T \ln \int \exp(-\beta V(X)) dX \ , \qquad (1)$$

where *V(X)* is the potential energy, k_B is the Boltzmann's constant, and *T* is the absolute temperature. The exponential dependence of the Boltzmann factor on energy makes the configurational integral notoriously slow to converge. To see the problem more clearly, we re-write **Eqn. 1** as follows:[17]

$$G = k_B T \ln \langle \exp(\beta V) \rangle \ . \qquad (2)$$

The angular brackets in this expression symbolize the configurational integral of the canonical ensemble. In principle, **Eqn. 2** provides a means of calculating (excess) free energy from a single conventional simulation. However, conventional simulations predominately sample the lower energy regions of conformational space, i.e. in accordance with the Boltzmann factor given in

Eqn. 1, and never adequately sample the higher energy states that contribute most significantly to the ensemble average of the free energy (as given by the “inverse-Boltzmann” factor in **Eqn. 2**). Therefore, the calculation of the free energy using a conventional simulation leads to poorly converged, and consequently inaccurate free energy estimates. Fortunately, it is the free energy difference that is generally of greatest interest, and this can be calculated using a coupling-parameter approach when the states are similar, i.e. the energy difference is small for all important configurations.

2.1.2 Basic formulation of conventional free energy methods

Conventional free energy calculation methods, such as the free energy perturbation (FEP)[12] and thermodynamic integration (TI) approaches,[4] can be utilized to evaluate the relative binding free energy between two ligands according to the thermodynamic cycle as shown in **Figure 1**.

$$\begin{tikzcd} L_0+R \arrow[r, "\Delta G_0"] \arrow[d, "\Delta G(solv)"'] & L_0R \arrow[d, "\Delta G(bind)"] \\ L_1+R \arrow[r, "\Delta G_1"'] & L_1R \end{tikzcd}$$

Figure 1. Thermodynamic cycle used for free energy calculations. L_0 and L_1 represent the free ligands in aqueous solution and L_0R and L_1R represent the corresponding ligands complexed with the protein receptor R.

The relative binding free energy of the two ligands, $\Delta\Delta G_{(bind)}=\Delta G_1-\Delta G_0=\Delta G_{(bind)}-\Delta G_{(solv)}$, is the difference between the relative free energy of the ligands in the complexed state and that of the free ligands. The free energy difference in each half of the thermodynamic cycle can be calculated using the potential energy of a hybrid system, written as a linear function of the two endpoint states and connected through a “chemical coordinate,” the coupling parameter λ:

$$\begin{aligned} V(x,\lambda) &= (1-\lambda)V_0(x) + \lambda V_1(x), \\ &= V_0(x) + \lambda \Delta V(x). \end{aligned} \tag{3}$$

The free energy difference between a state with the value of λ and the initial state (λ=0) is given by the FEP connection formula:[12]

$$\Delta G = G(\lambda) - G(0) = -k_B T \ln \left\langle \exp(-\beta \Delta V) \right\rangle_0 , \tag{4}$$

where the angular brackets stand for the configurational integral over the initial state. The conformational sampling indicated by **Eqn. 4** is generated according to the Boltzmann probability associated with the initial state potential. As discussed in **Section 2.1.1**, convergence of conformational sampling has been a major issue in free energy calculations, and the relationship shown in **Eqn. 4** doesn’t always lead to converged free energy estimates. To ensure adequate sampling of the important conformations, FEP calculations are generally limited to free energy differences of less than 2kcal/mol.[4,18] However, the free energy differences for many chemical and biological systems are larger than this. Therefore a multi-step approach is generally adopted. By summing over the intermediate states along the λ coordinate, the total free energy change is determined by

$$\Delta G = \sum_{i=1}^{n} \Delta G(\lambda_{i-1} \rightarrow \lambda_i) , \tag{5}$$

where the interval λ=(0,1) has been divided into *n* small increments *Δλ*.

An alternative approach to free energy calculations is the thermodynamic integration (TI) method,[4] which considers the ensemble average of the first derivative of the hybrid potential with respect to λ at various values of λ:

$$\Delta G = \int_0^1 \left(\frac{\partial A}{\partial \lambda}\right) d\lambda = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_\lambda d\lambda, \\ = \sum_{i=1}^{n} \left\langle {}^{\partial V}\!/_{\partial \lambda} \right\rangle_i \Delta\lambda_i . \tag{6}$$

Although these methods have been successfully applied to assess the relative binding free energy in a number of protein-ligand systems,[4,19-24] they are computationally expensive because of time-consuming sampling of nonphysical intermediate states. This computational expense has hindered the extensive application of conventional free energy based approaches to the drug design process.

2.1.3 Formulation of the umbrella sampling techniques

While FEP and TI are, in principle, umbrella sampling methods, the more conventional format in which we consider umbrella sampling is that used to obtain the free energy along a "reaction coordinate" ξ, typically a configurational coordinate. Here the "reactant" configuration and the "product" configuration are represented by $\xi = 0$ and $\xi = 1$, respectively. The Helmholtz free energy difference for a continuous coordinate, or the reversible work required to carry the system from the reactant configuration to the product configuration, is often referred to as the "potential of mean force", $W(\xi)$, and is derived from $\rho(\xi)$, the probability density of the system:[25]

$$W_i(\xi) = -k_B T \ln \rho(\xi) . \tag{7}$$

This leads to the free energy difference $\Delta G=W(\xi=1)-W(\xi=0)$. As discussed earlier, inadequate sampling may occur if $W(\xi)$, differs by more than a few kcal/mol over the range of ξ. To overcome this problem, the umbrella sampling technique[26,27] has frequently been used to enhance the sampling of conformational space along a reaction coordinate. In this approach, the original potential $V(x)$ is replaced by the modified potential $V(x)+ U(\xi)$. The auxiliary potential, or umbrella potential, $U(\xi)$ is used either to flatten out the energy barriers along the reaction coordinate ξ, or to restrict the sampling of the coordinate to a specific region of conformational space. In the former case, a more uniformly distributed density function $\rho^*(\xi)$, can be generated with a fixed amount of sampling because transitions between the reactant and product configurations are now more facile. In the latter case, the statistical sampling of important regions in the configurational space of the reaction coordinate can be better controlled. In both cases, the true probability density is recovered from the following equation:

$$\rho(\xi) = \frac{\rho^*(\xi)\exp\{\beta U^*(\xi)\}}{\left\langle \exp\{\beta U^*(\xi)\}\right\rangle^*}, \tag{8}$$

where the notation <…>* emphasizes that the ensemble average is being taken over conformations biased by the modified potential function.

In many applications, a single biasing potential is not sufficient to cover the whole range of ξ and simultaneously produce good sampling. Thus a set of restraining potentials, $U_i^*(\xi)$, are used to shift the local minima in the desired direction. In this "windowing" approach, the potential of mean force, $W_i(\xi)$, in each window takes the form

$$W_i(\xi) = -k_B T \ln \rho_i^*(\xi) - U_i^*(\xi) + C_i, \tag{9}$$

where the constant C_i is $C_i = \beta^{-1}\ln\left\langle \exp\{\beta U_i^*(\xi)\}\right\rangle^*$. In order to achieve a uniformly good estimate of the potential of mean force, the difference constants from successive simulation windows has to be perfectly matched so as to make $W_i(\xi)$ agree in the overlapping regions.[4,28] Optimal data combining methods such as the weighted histogram analysis method (WHAM) can be used to optimize links between simulations and produce the best possible estimation of free energies.[26,27,29-33] It is clear that umbrella sampling is a powerful technique and should be of use in sampling "chemical space" as well as configurational space. I discuss an extension of umbrella sampling to "chemical coordinates" in **Section 2.3**.

2.2 Overview of approximate approaches for multiple-ligand screening

When relative binding free energies of multiple ligands are to be evaluated, FEP and TI require considerable amounts of computational time. This is because of the multi-step approach necessary to compute incremental free energy changes and the many pairwise comparisons that must be done. Alternatively, methods based on favorable interaction energies have been developed to rapidly approximate the free energy.[2,9,10,34-38] Although such approaches are relatively rapid, they are inherently incomplete since the entropy contribution to the free energy is (at least partially) ignored. Because of the potential importance of such entropic effects in chemical and biological systems, the development of new methodology for the free energy calculations is an area of active research. Recently, statistical mechanical "*ab initio*" free energy-based computational methods have been developed to screen out the better binders from a group of candidate compounds.[6,13-16,39-47] These methods will be briefly reviewed in **Section 2.3**. In the following I present an overview of the interaction energy-based approaches.

2.2.1 Linear interaction approximation

The linear interaction approximation (LIA) was introduced by Åqvist and co-workers[11,48-51] to calculate absolute binding free energies via molecular dynamics (MD) simulations. A version of the LIA equation takes the following form

$$\Delta G_{(bind)} = \beta\langle \Delta E_{Coulomb}\rangle + \alpha\langle \Delta E_{L-J}\rangle + \gamma\langle SASA\rangle . \qquad (10)$$

In this expression, $<\Delta E>$ is the energy difference between average contributions from ligand-solvent and ligand-protein interactions for the bound and unbound states. The scaling parameter for electrostatic interactions, β, is taken to be 0.5, from theories of ion solvation in which there is a linear response of the solvent to the electrostatic field of the ion. The scaling coefficient on the Lennard-Jones terms, α, varies from system to system. Therefore, a large number of ligands with known binding affinity have to be used to obtain the proper coefficients. Carlson and Jorgensen[52] also added a penalty for cavitation, which is linearly related to the change in solvent-accessible surface area (SASA) upon binding. This term was added to obtain positive free energies of hydration for molecules such as hydrocarbons. LIA has been successfully applied to a number of protein-ligands systems.[11,48-50,52-54] However, the values of β and α seem to depend on both the system and the force field.

2.2.2 Extrapolation from a single reference

An alternative approach, which is rooted in the FEP methodology described above, involves the extrapolation of free energy differences from a single reference. The method was introduced by Liu, Mark, and van Gunsteren for the estimation of free energies of related compounds.[45,55] In implementing this approach, the incorporation of a soft-core potential leads to an expansion in the sampling of configurations for related ligands and consequently the range of correctly estimated free energy differences.[43] Mordasini and McCammon demonstrated the usefulness of the extrapolation method for similar sized molecules and the difficulties of obtaining reliable results for different sized molecules.[56]

Radmer and Kollman have introduced an approach they call PROFEC (Pictorial Representation of Free Energy Components),[46] as a tool for optimizing ligand affinity based on extrapolations from a single dynamics simulation. The PROFEC contour maps can be used to visualize how the free energy changes when additional particles are added to a residue of the protein or to the ligand. The contour map is generated by evaluating the insertion free energy of a test particle at various grid points near the residue of interest, using coordinates from a molecular dynamics simulation:

$$\Delta G(g) = -k_B T \ln\left\langle \exp(-\beta \Delta V(g)) \right\rangle_0 , \tag{11}$$

where *g* is the coordinate of a grid point, *ΔG* is the free energy cost to insert the test particle at that point, *ΔV(g)* is the interaction energy between the test particle and the system, and the angular brackets indicate an average with respect to the reference state. PROFEC has been used to modify a ligand to improve its binding affinity[16,46] and selectivity,[57] as well as to increase protein stability.[58] Recently, Pearlman developed a variant of this approach, the floating independent reference frame (FIRF), which may be applicable to flexible ligands.[47]

2.2.3 Linear interaction approximation with continuum solvent

The molecular mechanics with Poisson-Boltzmann /surface area approach (MM/PBSA)[44] is a semi-empirical method to calculate free energy differences between protein-ligand complexes[41], protein-protein complexes,[42] and different forms of DNA and RNA.[44,59] The basic approach used in MM/PBSA follows the procedures used to analyze peptide and protein conformations as outlined by Yang and Honig[60] and Osapay et al.[61] Shen and co-workers demonstrated that the use of a single conformation together with the PB electrostatic and surface-area-dependent terms in MM/PBSA can lead to successful estimation of the binding free energies for a number of ligands.[62,63] In the MM/PBSA method, the binding free energies are estimated from

$$\Delta G_{(bind)} = \Delta G_{(bound)} - (\Delta G^{ligand}_{(solv)} + \Delta G^{receptor}_{(solv)}) , \tag{12}$$

and

$$\Delta G \approx \left\langle \Delta E_{gas} \right\rangle + \left\langle \Delta G_{PB} \right\rangle + \left\langle \Delta G_{SA} \right\rangle - T\left\langle S \right\rangle . \tag{13}$$

The solute configurations are sampled as “snap-shots” from a molecular dynamics simulation calculated using explicit solvent. For each solute configuration, the gas-phase energy, E_{gas}, is calculated without any solvent. Free energies of solvation are then re-introduced by using a PB calculation for the electrostatic term (G_{PB}) and a surface-area-dependent term (G_{SA}) for non-electrostatic contributions. Solute entropy contributions are estimated from (quasi-) harmonic analysis. The differences (ΔE_{gas}, ΔG_{PB}, and ΔG_{SA}) are calculated between the bound state and unbound states as shown in **Eqn. 12**. Variants of the MM/PBSA approach such as “computational alanine scanning”[42] or “computational fluorine scanning”[41] were also introduced and shown to be useful techniques to explore sensitivity of a given receptor site (or amino acid site in a protein) to changes in composition.

2.3 Free energy based screening methods: *λ-dynamics & Chemical Monte Carlo / molecular dynamics*

In two sections (**2.3.1** and **2.3.2**), the basic ideas of the free energy based screening methods are introduced. In the following sections, I describe our methodological development of free energy based screening methods, which is the main methodology of the present thesis.

2.3.1 Basic formulation

“The free energy based screening methods” are an extension of the coupling parameter approach used in the thermodynamic cycles. They differ, however, in the following aspects: (1) In the coupling-parameter approach, a single λ is used to transform one ligand into another, whereas in the free energy based screening methods, multiple λs (each corresponding to a given ligand) are used. Because of this feature, the binding free energies of multiple ligands are evaluated simultaneously. (2) In FEP and TI, λ is fixed during the simulation. In the free energy based screening methods, the λs evolve according to “equations of motion” via molecular dynamic or Monte Carlo methods. In this section, I will briefly elaborate on the main ideas and formulas associated with these approaches.

For a protein and a total of L ligands, a hybrid potential energy function is constructed as follows

$$V_0(X,\{x\},\{\lambda\}) = V_{env}(X) + \sum_{i=1}^{L} \lambda_i^2 (V_i(X,x_i) - F_i)\,, \tag{14}$$

where $\sum_{i=1}^{L} \lambda_i^2 = 1\,.$

In **Eqn. 14**, i indicates the ith ligand, X and x_i denote the coordinates of environment atoms and ligand i respectively, V_{env} is the potential energy involving the environment atoms only (i.e., those atoms which are common to all protein-ligand pairs), V_i is the interaction energy involving ligand i in the protein-ligand complex state and λ_i is the coupling parameter associated with ligand i, and F_i is a pre-calculated biasing potential, which may correspond to the relative free energy of ligand i in the unbound state (relative solvation free energy). The coupling parameter λ is replaced by λ^2 in order to avoid non-physical negative values in the λ-dynamics simulations.[14]

By properly coupling the system to a heatbath, the configurational partition function of the hybrid potential is canonical:

$$Z_0(X,\{x\},\{\lambda\}) = \int \exp\left[-\beta\left\{V_{\text{env}}(X) + \sum_{i=1}^{L} \lambda_i^2 \left(V_i(X,x_i) - F_i\right)\right\}\right] dXd\{x\}d\{\lambda\}\,. \tag{15}$$

In many cases, F_i may be rapidly evaluated using continuum solvation models such as the PB or generalized Born (GB) methods.[63-69] Note that F_i can also serve as a biasing potential to achieve better sampling of the phase space of interest, and faster convergence of the calculations as shown below.

The difference in binding free energy ($\Delta\Delta G_{j\to i}$) between arbitrarily chosen ligands i and j

can be obtained from $P_0(\lambda_i^2 = 1, \{\lambda_{m\neq i}^2 = 0\})$,[14,70] which corresponds to the amount of time ligand *i* occupies $\lambda_i^2=1$ during the simulation, is an indicator of the binding free energy of that ligand to the protein receptor. Because the reference free energy appears in the hybrid potential of the protein-ligand complexed state, the resulting free energy from **Eqn. 16** directly corresponds to the binding free energy difference. Therefore, these methods tend to provide better sampling for ligands that have more favorable binding free energies. Furthermore, such calculations often result in smaller statistical errors for the most favored compounds. There exists an analogy between this formalism and competitive binding experiments carried out in the laboratory. In fact, a competitive binding experiment usually consists of different ligands and a single receptor in solution and the best ligands are determined by the probability that a ligand is bound to the receptor. We have

$$\begin{aligned}
\Delta\Delta G_{j\to i} &= \Delta G_{(bound)} - \Delta G_{(solv)}, \\
&= -k_B T \ln \frac{\int \exp\{-\beta(V_{env}(X) + V_i(X, x_i))\}}{\int \exp\{-\beta(V_{env}(X) + V_j(X, x_j))\}} - (F_i - F_j), \\
&= -k_B T \ln \frac{\int \exp\{-\beta(V_{env}(X) + V_i(X, x_i) - F_i)\}}{\int \exp\{-\beta(V_{env}(X) + V_j(X, x_j) - F_j)\}}, \\
&= -k_B T \ln \frac{Z_0(\lambda_i^2 = 1, \{\lambda_{m\neq i}^2 = 0\})}{Z_0(\lambda_j^2 = 1, \{\lambda_{m\neq j}^2 = 0\})}, \\
&= -k_B T \ln \frac{P_0(\lambda_i^2 = 1, \{\lambda_{m\neq i}^2 = 0\})}{P_0(\lambda_j^2 = 1, \{\lambda_{m\neq j}^2 = 0\})}.
\end{aligned} \qquad (16)$$

In the λ-dynamics method,[13] both λ variables (coupling parameters) and the atomic coordinates are propagated using molecular dynamics (MD). The dynamics of the system is generated from an extended Hamiltonian:[71,72]

$$\begin{aligned}
H_0(X,\{x\},\{\lambda\}) &= T_x + T_\lambda + V_0(X,\{x\},\{\lambda\}), \\
T_\lambda &= \sum_{k=1}^{L} \frac{m_\lambda}{2} \dot{\lambda}_k^2,
\end{aligned} \qquad (17)$$

where T_x and T_λ are the kinetic energies of the atomic coordinates and λ variables, respectively.

The λs are treated as volumeless particles with mass m_λ. Since the λ variables are associated with the "chemical reaction coordinates", the λ-dynamics method can utilize the power of specific biasing potentials in the umbrella sampling method to overcome sampling problems that require conventional FEP calculations to be performed in multiple steps.

Instead of using MD, the λ variables may also be sampled stochastically. In the CMC/MD approach, Metropolis Monte Carlo method[73] is used to evolve the λ-variables and molecular dynamics is used to evolve the atomic coordinates. The Metropolis Monte Carlo criteria leads to the generation of a canonical ensemble of the ligands when the following transition probability from ligand *i* to *j* is used

$$A_{i\rightarrow j} = \min\left(1, \exp(-\beta \Delta V_{i\rightarrow j})\right), \tag{18}$$

where $\Delta V_{i\rightarrow j} = V_j - V_i$. Both the λ-dynamics method and the CMC/MD method give the same configurational partition functions (**Eqn. 15**). Therefore, **Eqn. 16** can be applied to CMC/MD as well. The hybrid Monte Carlo / molecular dynamics method was originally presented by Bennett[74] and Tidor.[70,75] The straightforward extension of this approach to multiple ligands, which is called CMC/MD, was carried out by Pitera & Kollman.[15]

2.3.2 Iterative techniques using Weighted Histogram Analysis Method

An iterative procedure using the weighted histogram analysis method (WHAM)[26,27,29-33] was developed to improve sampling of the chemical space, and therefore to make free energy calculations converge more rapidly. The use of this method in conjunction with λ-dynamics is described below.[14] Its extension to CMC/MD is also straightforward.

Since {λ} is treated as a dynamical variable, just as the atomic coordinates, we use X_{Tot} to denote the phase space that encompasses *X*, *{λ}*, and *{x}*. Thus the hybrid potential in **Eqn. 14** can be rewritten as

$$V_0(X_{Tot}) = V_{ref}(X_{Tot}) - \sum_{i=1}^{L} F_i \lambda_i^2 \ , \tag{19}$$

where

$$V_{ref}(X_{Tot}) = V_{\text{env}}(X) + \sum_{i=1}^{L} \lambda_i^2 V_i(X, x_i) \ . \tag{20}$$

When this potential is utilized in a series of λ-dynamics, or CMC/MD, the WHAM equations for multiple reaction coordinates and at constant temperature can be readily applied to obtain the best estimate of free energy using all of the data from *n* previous simulations

$$P^{n}_{\{F^n\}}\left(\{\lambda^2\}\right) = \frac{\sum_{j=1}^{n} N_j(\{\lambda^2\}) \exp\left(-\beta \sum_{i=1}^{L} (-F_i^{\,n}) \lambda_i^2\right)}{\sum_{k=1}^{n} n_k \exp\left(f_k - \beta \sum_{i=1}^{L} (-F_i^{\,k}) \lambda_i^2\right)} , \tag{21}$$

$$\exp(-f_m) = \sum_{\{\lambda^2\}} P^{m}_{\{F^m\}}\left(\{\lambda^2\}\right), \tag{22}$$

where **Eqn. 21** and **22** are solved self-consistently. After the *m*-th iteration, the estimated free energy relative to the reference free energy $\{F^0\}$ is

$$G_{i\{F\}} = -k_B T \ln P^{m}_{\{F^0\}}(\lambda_i^2 = 1, \{\lambda_{k \neq i}^2\} = 0) \ , \tag{23}$$

and a new biasing potential for the next iteration is estimated as

$$F_i^{\,m+1} = G_{i,\{F=0\}} \ . \tag{24}$$

The above procedure can be used to extract the free energy, G_i, of each ligand.

Since this approach biases the sampling of different ligands in the receptor by successively better estimates of their relative binding free energy, the bound conformations of all the ligands are expected to be sampled equally well after some number of cycles of simulation. Sometimes an additional term Δ_i may also be added to **Eqn. 24** to either enhance or reduce sampling of a state dominated by ligand *i*. As in all iterative procedures, an initial trial value of F_i must be given. If a poor initial free energy is used, then the states with F_i<G_i will be sampled less frequently than they would with F_i=G_i. Similarly, states with F_i>G_i will be sampled more frequently. The approach was also applied with the CMC/MD method by Kollman and co-workers, which they renamed the adapted CMC/MD method.[16]

In the iterative approach using a constant bias {*F*}, the free energy barrier is reduced each successive iteration and therefore produces complete sampling of important configurations along the coupling parameter (i.e., reaction coordinates). Furthermore, more complicated umbrella potentials (e.g., see **Eqn. 27**) may also be applied with the iterative procedure.

The applications presented in **Chapters 4 - 6** demonstrate the robustness of the iterative procedure using WHAM.

2.3.3 Efficient sampling of the chemical coordinates

In the CMC/MD method, the stochastic sampling by MC steps permits one to restrict the sampling of chemical space, i.e., the space of {λ}. For example, Kollman and co-workers[15,16] limit their chemical sampling only to transitions between the end points in their CMC/MD simulations as follows:

$$\sum_{i=1}^{L}\lambda_i^2 = 1 \text{ and } \lambda_i^2 = \{0,\ 1\}\,. \tag{25}$$

This condition allows sampling of the end states of interest exclusively. However, inefficient sampling of the chemical states, such as trapping in one end state, may occur. This is prevalent when there is a large free energy gap between the ligands. Trapping may, however, be partially avoided by the addition of a few chosen intermediate states to bridge the end points. In the λ-dynamics method, the λ-variables are treated as continuous variables, so smooth transitions between the end points are expected, and generally observed.

As shown in **Eqn. 5**, the conventional free energy calculation methods such as FEP and TI require the introduction of intermediate states to obtain converged free energy differences between the end points. In the free energy based screening methods, some ligands can serve as the intermediate states connecting otherwise dissimilar end points. Nevertheless, there are potential problems. For example, if the relative free energy of the intermediate states is lower than that of the end points, most of the computational time will be spent exploring unphysical intermediate states. Therefore, the relative free energy of the end points will be less well determined. Conversely, higher energy intermediate states result in rare transitions in the chemical coordinates and thus slow convergence.

The umbrella sampling technique[26,27] can be utilized to overcome these difficulties. An umbrella potential along the λ coordinates can be expressed as

$$V_{um} = V_0(X,\{x\},\{\lambda\}) + \sum_{i=1}^{L} B_i(\lambda_i), \tag{26}$$

where the λ-dependent potential term, $B_i(\lambda_i)$, will serve as an umbrella (or a biasing) potential to limit the range of {λ} and to increase the rate of transitions among potential wells separated by high-energy barriers. A harmonic potential is commonly used to flatten the energy surface and enhance sampling along the chemical coordinate λ:

$$B_i(\lambda_i) = k_i\left(\lambda_i^2 - B_i^0\right)^2, \tag{27}$$

where $0 < B_i^0 < 1$. The condition ($k_i > 0$) can be used to increase the transition between the end points, while $k_i < 0$ tends to increase sampling of the end points. The unbiased probability of the bound states can be calculated by using the umbrella sampling formalism as follows (see **Eqn. 8**):

$$
\begin{aligned}
P_0(\lambda_i^2=1,\{\lambda_{m\neq i}^2=0\}) &= \frac{\int \exp(-\beta V_0)\theta(\lambda_i^2-1)}{\int \exp(-\beta V_0)}, \\
&= \frac{\int \frac{\exp\{-\beta(V_0+\sum B_k)\}}{\int \exp\{-\beta(V_0+\sum B_k)\}} \exp(\beta\sum B_k)\theta(\lambda_i^2-1)}{\int \frac{\exp\{-\beta(V_0+\sum B_k)\}}{\int \exp\{-\beta(V_0+\sum B_k)\}} \exp(\beta\sum B_k)}, \\
&= \frac{\left\langle \exp(\beta\sum_{k=1}^{L} B_k)\theta(\lambda_i^2-1)\right\rangle_{um}}{\left\langle \exp(\beta\sum_{k=1}^{L} B_k)\right\rangle_{um}},
\end{aligned} \tag{28}
$$

where the angular brackets denote the ensemble average over the biased distribution and *$\theta(x)$* is a step function, which is unity when its argument is greater than zero but is otherwise zero. By using the probability of the bound states, *$\Delta\Delta G$* is obtained from

$$
\begin{aligned}
\Delta\Delta G_{j\to i} &= -k_B T \ln \frac{P_0(\lambda_i^2=1,\{\lambda_{m\neq i}^2=0\})}{P_0(\lambda_j^2=1,\{\lambda_{m\neq j}^2=0\})}, \\
&= -k_B T \ln \frac{\left\langle \exp(\beta\sum_{k=1}^{L} B_k)\theta(\lambda_i^2-1)\right\rangle_{um}}{\left\langle \exp(\beta\sum_{k=1}^{L} B_k)\theta(\lambda_j^2-1)\right\rangle_{um}}, \\
&= -k_B T \ln \frac{\left\langle \theta(\lambda_i^2-1)\right\rangle_{um} \exp\left\{\beta\sum_{k\neq i}^{L} B_k(\lambda_k=0)\right\} \exp\{\beta B_i(\lambda_i^2=1)\}}{\left\langle \theta(\lambda_j^2-1)\right\rangle_{um} \exp\left\{\beta\sum_{m\neq j}^{L} B_m(\lambda_m=0)\right\} \exp\{\beta B_j(\lambda_j^2=1)\}}, \\
&= -k_B T \ln \frac{P_{um}(\lambda_i^2=1,\{\lambda_{m\neq i}^2=0\})}{P_{um}(\lambda_j^2=1,\{\lambda_{m\neq j}^2=0\})} - (\Delta B_i - \Delta B_j),
\end{aligned} \tag{29}
$$

where $\Delta B_i = B_i(\lambda_i^2=1) - B_i(\lambda_i=0)$ and P_{um} is the probability function of the hybrid potential with the umbrella potential. If *ΔB_i=ΔB_j*, the effect of the umbrella potential will be canceled completely. An iterative procedure is sometimes required to produce complete sampling of important configurations along the chemical coordinates. In such cases, WHAM, as discussed in the former section, can be used to process the sampling data in an efficient and general way.

Another approach for efficient sampling along the chemical coordinates was suggested by Tidor.[70] In his work, simulated annealing was used to sample the chemical variables on the free energy surface of the system (e.g., high temperature for the chemical variables and low temperature for Cartesian variables). In a demonstration calculation, the method was applied to simple molecules to select the one with the most favorable solvation energy. Agreement between the observed and calculated trends was obtained.

In **Chapter 6**, I will show the application of a biasing potential (**Eqn. 27**) that successfully enhances the sampling of the end states.

2.3.4 Sampling of the unselected ligands in the multiple topology model

When a single topology representation of the ligand is used, i.e., one in which atoms that change are all connected to a common framework, the configuration of the unbound ligands is determined automatically. No ambiguity exists regarding the choice of the configuration of the unbound ligands with λ equal to zero, making the choice of proper MC steps straightforward. To see this, consider the detailed balance condition of MC for moves between ligands *i* and *j* using the single topology

$$\begin{gathered} P_0(\lambda_i^2 = 1, \{\lambda_{m\neq i}^2 = 0\})\alpha_j A_{i\to j} = P_0(\lambda_j^2 = 1, \{\lambda_{m\neq j}^2 = 0\})\alpha_i A_{j\to i}, \\ \exp(-\beta V_i)\frac{1}{L-1}A_{i\to j} = \exp(-\beta V_j)\frac{1}{L-1}A_{j\to i}, \end{gathered} \tag{30}$$

$$\frac{A_{i\to j}}{A_{j\to i}} = \exp\left\{-\beta(V_i - V_j)\right\}, \tag{31}$$

where α_i is the probability of selecting the ligand *i* and $A_{i\to j}$ is the acceptance probability of a move from ligand *i* to *j*. It is straightforward to demonstrate that the basic Metropolis scheme shown in **Eqn. 18** obeys this condition.[73] However, the assignment of a single topology model for multiple ligands is a complicated problem, especially when the ligands have different shapes. Furthermore, inadequate assignment of the common features in a single topology leads to small overlap of the important configurations between the selected ligand and unselected ones. In general, multiple independent topologies have been used for the multiple-ligand screening to avoid these problems.[15,16,39,40]

In the multiple topology model, the unbound ligands tend to move significantly from their preferred binding orientations and explore high-energy regions of conformation space when they are only weakly coupled to their environment.[39] This results in inefficient sampling of chemical space. In CMC/MD simulations,[16,57] the problem was addressed by the addition of a harmonic potential between the centers of mass of all ligands, and the imposition of "ghost" forces on the unbound ligands. The ligand "ghost" forces are those exerted on the unbound ligands by the environment atoms. However, in the approach of Kollman and co-workers, the unbound ligands remain invisible to the environment atoms. The effect of the harmonic potential is canceled out in the calculation of $\Delta\Delta G$, and thus moves with only this additional potential satisfy the detailed balance condition. On the other hand, it is difficult to correct for the effect of the "ghost" forces since they do not have physical origin. To understand the effect of "ghost" forces on CMC/MD steps and statistical averages arising from such calculations, we consider the special condition in which all environment atoms are fixed. With this idealization, the "ghost" forces can be recognized as those coming from a restraining potential, R_i, arising from the fixed environment. For this situation, the probability, $\alpha_i(x_i)$, of selecting the ligand i with the coordinates x_i becomes

$$\alpha_i(x_i) = \frac{1}{L-1}\frac{\exp(-\beta R(x_i))}{\int \exp(-\beta R(x_i))dx_i}. \tag{32}$$

The condition of detailed balance for an MC step between the ligand i with coordinate of x_i and ligand j with coordinate of x_j can be written down as follows.

$$P_0(\lambda_i^2=1,\{\lambda_{m\neq i}^2=0\},x_i)\alpha_j(x_j)A_{i\to j}(x_i,x_j)=P_0(\lambda_j^2=1,\{\lambda_{m\neq j}^2=0\},x_j)\alpha_i(x_i)A_{j\to i}(x_j,x_i),$$

$$\exp(-\beta V_i)\frac{1}{L-1}\frac{\exp(-\beta R_j(x_j))}{\int\exp(-\beta R_j(x_j))}A_{i\to j}=\exp(-\beta V_j)\frac{1}{L-1}\frac{\exp(-\beta R_i(x_i))}{\int\exp(-\beta R_i(x_i))}A_{j\to i}, \tag{33}$$

$$\frac{A_{i\to j}}{A_{j\to i}}=\left\{\exp\left(\beta\Delta R_{i\to j}\right)\frac{\int\exp(-\beta R_j)}{\int\exp(-\beta R_i)}\right\}\exp\left(-\beta\Delta V_{i\to j}\right), \tag{34}$$

where $\Delta R_{i\to j}=R_j(x_j)-R_i(x_i)$, $\Delta V_{i\to j}=V_j(x_j)-V_i(x_i)$ and $A_{i\to j}$ is the acceptance probability of a move from ligand *i* with x_i to ligand *j* with x_j. One acceptance rule that obeys the detailed balance condition and yields the canonical ensemble in this case is

$$A_{i\to j}=\min\left(1,\exp(\beta\Delta R_{i\to j})\frac{\int\exp(-\beta R_j)}{\int\exp(-\beta R_i)}\exp\left(-\beta\Delta V_{i\to j}\right)\right). \tag{35}$$

If, in fact, the environment atoms are allowed to move, the rigorous estimation of $A_{i\to j}$ is unclear. Kollman and co-workers have assumed that the effect of the “ghost” forces cancel for comparisons of similar ligands, and used an acceptance rule following **Eqn. 18** instead of **Eqn. 35** in their CMC/MD simulations.[16,57] This approximation was demonstrated for some systems to be at least qualitatively reasonable.[15,16]

In order to overcome the problems that can occur in sampling configurations of the unbound ligands when using a multiple topology model, we consider two types of restraining potentials. For simplicity, both restraining potentials are assumed to disappear at the bound states:

$$R_i'(\lambda_i^2=1)=R''(\lambda_i^2=1)=0\,. \tag{36}$$

The first type of the restraining potential for ligand *i*, R_i’, is defined as a function of *X*, x_i, and, λ_i, and we have

$$\mathrm{V}'(X,\{x\},\{\lambda\})=V_0(X,\{x\},\{\lambda\})+\sum_{i=1}^{L}R_i'(X,x_i,\lambda_i)\,. \tag{37}$$

With a straightforward application of the umbrella sampling formalism, *ΔΔG* is obtained from

$$\Delta\Delta G_{j\to i} = -k_B T \ln \frac{\left\langle \exp\left(\beta \sum_{k=1}^{L} R_k^{\prime} (X, x_k, \lambda_k) \right) \sigma(\lambda_i^2 - 1) \right\rangle_{V'}}{\left\langle \exp\left(\beta \sum_{l=1}^{L} R_l^{\prime} (X, x_l, \lambda_l) \right) \sigma(\lambda_j^2 - 1) \right\rangle_{V'}},$$

$$= -k_B T \ln \frac{\sum_{(\lambda_i^2=1)} \exp\left(\beta \sum_{k\neq i}^{L} R_k^{\prime} (X, x_k, \lambda_k = 0) \right)}{\sum_{(\lambda_j^2=1)} \exp\left(\beta \sum_{l\neq j}^{L} R_l^{\prime} (X, x_l, \lambda_l = 0) \right)}. \tag{38}$$

Here, the summation is taken at the bound state (λ^2=1) of the λ-dynamics trajectory, including the restraining potential. Unfortunately, with this biasing potential the effect of the restraint (*R_i’*) becomes too large to yield reasonable convergence as the number of the unbound ligands increases.

Another type of restraining potential for ligand *i*, *R_i”*, is defined as a function of x_i, and λ_i. In this case, the restraining potentials does not depend directly on environment atom coordinates, and we have

$$\mathrm{V}''(X,\{x\},\{\lambda\}) = V_0(X,\{x\},\{\lambda\}) + \sum_{i=1}^{L} R_i^{\prime\prime} (x_i, \lambda_i)\,. \tag{39}$$

Because *R_i”=0* when *$\lambda_i^{\ 2}$ =1* and none of the restraining potentials depend on the environment atoms, the partition function for the system when *$\lambda_i^{\ 2}$ =1* can be expressed as follows.

$$Z''_{\lambda_i^2=1} = \int \exp\{-\beta(\mathrm{V}_{env} + V_i - F_i)\} dX dx_i \prod_{k\neq i}^{L} \int \exp\left(-\beta R_k^{\prime\prime} (x_k, \lambda_k = 0) \right) dx_k\,. \tag{40}$$

Using this relationship, *ΔΔG* can be written as two terms. The first term involves the probability that a ligand is in the dominant states (*λ^2=1*) during the λ-dynamics simulation and in the presence of the restraining potential. The second term corresponds to the partition function of the restraining potentials (the umbrella correction):

$$\Delta\Delta G_{j\rightarrow i} = -kT\ln\frac{Z'_{\lambda_i^2=1}\prod_{k\neq j}^{L}\int\exp\left(-\beta R_k''(x_k,\lambda_k=0)\right)dx_k}{Z'_{\lambda_j^2=1}\prod_{l\neq i}^{L}\int\exp\left(-\beta R_l''(x_l,\lambda_l=0)\right)dx_l},$$

$$= -kT\ln\frac{P'_{\lambda_i^2=1}}{P'_{\lambda_j^2=1}} - kT\ln\frac{\int\exp\left(-\beta R_i(x_i,\lambda_i=0)\right)dx_i}{\int\exp\left(-\beta R_j(x_j,\lambda_j=0)\right)dx_j}. \qquad (41)$$

The second term in **Eqn. 41** is constant and may be estimated using free energy simulations or semi-empirical methods. When the same restraining potentials are added to the calculations of the solvation free energy difference for half of the cycle, the second terms are completely canceled in a closed thermodynamic cycle. In this case, only the first term yields the *ΔΔG* of the ligands from the λ-dynamics trajectory. However, the addition of these restraining potentials for the unbound states may require the complicated implementation of calculations. Furthermore, additional restraining potentials have a risk to hinder the complete sampling of the important conformations at the unbound states.

When the ligands are similar and the entropy terms associated with the restraining potential are expected to cancel, the second term can be approximated by an internal energy difference:

$$\begin{aligned} -kT\ln\frac{\int\exp\left(-\beta R_i''(x_i,\lambda_i=0)\right)dx_i}{\int\exp\left(-\beta R_j''(x_j,\lambda_j=0)\right)dx_j} &= (U_i''-U_j'')-(S_i''-S_j'')T, \\ \approx U_i''-U_j''. & \end{aligned} \qquad (42)$$

This internal energy can be estimated by using the trajectory of the free energy based screening simulations:

$$\begin{aligned} U_i'' &= \int R_i''(x_i,\lambda_i=0)\exp\left(-\beta R_i''(x_i,\lambda_i=0)\right)dx_i, \\ &\approx \frac{1}{n}\sum_{(\lambda_i=0)}^{m=1,n} R_m''. \end{aligned} \qquad (43)$$

The restraining potential should be chosen carefully since the important configurations for {R_i”}

should have large overlap with those for {V_i}. The interaction energy {V_i} for the average structure of the environment atoms is a reasonable choice for {R_i"}. In fact, it is an optimal choice to bias the ligands that do not have large λ-values because it restrains these ligands to the vicinity of the receptor. To represent fluctuations in this mean-receptor potential field, soft-core representations of the van der Waals or electrostatic interactions can be used, or the overall potential field can be scaled. In latter case, the restraining potential R_i" may be represented by

$$R_i''(x_i, \lambda) = \alpha V_i(x_i, X_0)\,, \tag{44}$$

where α is a scaling parameter. In this case, we have to choose the averaged coordinates of the environment atoms, X_0, properly. Since most protein atoms stay near the X-ray crystallographic structure throughout the MD simulation, the environment atoms may be assumed to be rigid or slowly varying as compared with ligand atoms. In this case, the restraining potential R_i*“(X_0,x_i)* may be replaced by R_i*”(X(t), x_i)*:

$$R_i''(X_0, x_i) \approx R_i''(X(t), x_i)\,. \tag{45}$$

The approximation shown in **Eqn. 45** saves considerable computational cost because interactions between X_0 and {x} are eliminated. This approximation is only true when all environment atoms are fixed at their average values during the simulation. But, if the environment atoms are relatively rigid and any bias from the ligand atoms can be cancelled out, this approximation (**Eqn. 45**) will be valid. Finally, with the inclusion of the restraining potential and using the approximations given above, the binding free energy difference between ligands *i* and *j* may be estimated as [39]

$$\Delta\Delta G_{j\to i} \approx -\frac{1}{\beta}\ln\left[\frac{P(\lambda_i^2 = 1.0, \{\lambda_{m\neq i}^2 = 0\})}{P(\lambda_j^2 = 1.0, \{\lambda_{l\neq j}^2 = 0\})}\right] + U_i'' - U_j''\,. \tag{46}$$

When iterative techniques using WHAM are carried out with the restraining potential, a set of new biasing offsets {F} for (n+1)-th simulation, where all ligands are expected to compete equally, can be calculated from

$$F_i^{n+1} = F_i^0 - \frac{\ln P_{\{F^0\}}^n\left(\left\{\lambda^2\right\}\right)}{\beta(1-\alpha)}. \tag{47}$$

The relative binding free energies may be estimated with the probability $P_{\{F^0\}}^n$, which is best estimated from WHAM equations (**Eqns. 21, 22**)

$$\Delta\Delta G_{j\to i} \approx -\frac{1}{\beta}\ln\left[\frac{P_{\{F^0\}}^n(\lambda_i^2 = 1.0, \{\lambda_{m\neq i}^2 = 0\})}{P_{\{F^0\}}^n(\lambda_j^2 = 1.0, \{\lambda_{l\neq j}^2 = 0\})}\right] + U_i^{"} - U_j^{"}. \tag{48}$$

I will present the effectiveness of this newly introduced restraining potential in **Chapters 3**, **5**, and **6**.

2.3.5 Incorporation of the generalized Born solvent model

The use of continuum solvent models will decrease the number of degrees of freedom in the system, and consequently accelerate the convergence of thermodynamic properties by eliminating the ensemble average of the solvent molecules. Moreover, the absence of the collisions between the unbound ligands and mobile solvent enhances the overlap of the important configurations. However, conventional numerical solutions to the Poisson-Boltzmann (PB) equations are too slow for practical applications and one must use approximate analytical representations such as the GB model originally proposed by Still and co-workers.[65,76] Since the electrostatic solvation energy and its derivative can be calculated analytically in the GB model, it may be applied to configurational sampling using molecular dynamics.[64,77-82] In the GB model, the solvent polarization energy, G_{pol}, is approximated by the following equation.

$$G_{pol} = -166\left(1-\frac{1}{\varepsilon}\right)\sum_i^N\sum_j^N \frac{q_i q_j}{\sqrt{\left(r_{ij}^2 + \alpha_i\alpha_j e^{-D_{ij}}\right)}}, \tag{49}$$

with $D_{ij} = \frac{r_{ij}^2}{4\alpha_i \alpha_j}$ where ε denotes the dielectric constant of solvent; q_i and q_j stand for the charges of atom *i* and *j*, respectively; r_{ij} represents the distance between atom *i* and *j*; and α_i is "generalized Born" radii of the atoms *i* in a specific molecular environment. The values of α_i can be estimated using

$$\alpha_i = -166 / G_{pol,i}\,, \tag{50}$$

and a linearized form of Still's original empirical formula to get $G_{pol,i}$:[83]

$$G_{pol,i} = \left[\frac{1}{\lambda_\alpha} \left(\frac{-166}{R_{vdW,i}} \right) + P_1 \left(\frac{166}{R_{vdW,i}^2} \right) + P_2 \sum_j^{bond} \frac{V_j}{r_{ij}^4} + P_3 \sum_j^{angle} \frac{V_j}{r_{ij}^4} + P_4 \sum_j^{nonbond} \frac{V_j}{r_{ij}^4} CCF \right]. \tag{51}$$

with

$$CCF = \begin{cases} 1.0, & \text{for} \left(\frac{r_{ij}}{R_{vdW,i} + R_{vdW,j}} \right)^2 > \frac{1}{P_5}\,, \\ \frac{1}{4} \left[1.0 - \cos\left\{ P_5 \pi \left(\frac{r_{ij}}{R_{vdW,i} + R_{vdW,j}} \right)^2 \right\} \right]^2, & \text{for} \left(\frac{r_{ij}}{R_{vdW,i} + R_{vdW,j}} \right)^2 \le \frac{1}{P_5}\,. \end{cases} \tag{52}$$

The values of λ_α and P_1 - P_5 can be determined by fits to PB solvation energies for a database of compounds.[83] The incorporation of the GB model into the free energy calculation methods using the FEP/TI and λ-dynamics was carried out.[84] The GB energy term should satisfy the following conditions at the intermediate states: (1) the GB solvation energy changes continuously among the end points, (2) when physically meaningful end points are sampled in **Eqn. 14** (i.e., one ligand is selected ($\lambda_i^2 = 1$) or a set of identical ligands adopt exactly the same coordinates), the GB energy of the multiple ligand system should be the same as that of the selected end point. From these conditions, we present two possible definitions for the GB energy of the intermediate states:

Type 1

$$G_{pol}=-166\left(1-\frac{1}{\varepsilon}\right)\left\{\sum_{i}^{i\in env}\sum_{j}^{j\in env}\frac{q_iq_j}{\sqrt{\left(r_{ij}^2+\alpha_i\alpha_je^{-D_{ij}}\right)}}+\sum_{k=1}^{L}\lambda_k^2\left(2\sum_{i}^{i\in env}\sum_{j}^{j\in k}\frac{q_iq_j}{\sqrt{\left(r_{ij}^2+\alpha_i\alpha_je^{-D_{ij}}\right)}}+\sum_{i}^{i\in k}\sum_{j}^{j\in k}\frac{q_iq_j}{\sqrt{\left(r_{ij}^2+\alpha_i\alpha_je^{-D_{ij}}\right)}}\right)\right\}, \tag{53}$$

where *env* represents the environment atoms, *k* denotes the ligand number, and *L* represents the total number of the ligands. The effective Born radii for the environment atoms can be calculated from **Eqns. 50** and **54**.

$$\begin{aligned}G_{pol,i}^{i\in env}=&\frac{1}{\lambda_\alpha}\left(\frac{-166}{R_{vdW,i}}\right)+P_1\left(\frac{166}{R_{vdW,i}^2}\right)+P_2\sum_{j\in env}^{bond}\frac{V_j}{r_{ij}^4}+P_3\sum_{j\in env}^{angle}\frac{V_j}{r_{ij}^4}+P_4\sum_{j\in env}^{nonbond}\frac{V_j}{r_{ij}^4}CCF\\&+\sum_{k=1}^{L}\lambda_k^2\left(P_2\sum_{j\in k}^{bond}\frac{V_j}{r_{ij}^4}+P_3\sum_{j\in k}^{angle}\frac{V_j}{r_{ij}^4}+P_4\sum_{j\in k}^{nonbond}\frac{V_j}{r_{ij}^4}CCF\right).\end{aligned} \tag{54}$$

Effective Born radii for the atoms belonging to ligand *k* are calculated from **Eqns. 50** and **55**, and we have

$$G_{pol,i}^{i\in k}=\frac{1}{\lambda_\alpha}\left(\frac{-166}{R_{vdW,i}}\right)+P_1\left(\frac{166}{R_{vdW,i}^2}\right)+P_2\sum_{j\in\left\{\begin{smallmatrix}env\\k\end{smallmatrix}\right.}^{bond}\frac{V_j}{r_{ij}^4}+P_3\sum_{j\in\left\{\begin{smallmatrix}env\\k\end{smallmatrix}\right.}^{angle}\frac{V_j}{r_{ij}^4}+P_4\sum_{j\in\left\{\begin{smallmatrix}env\\k\end{smallmatrix}\right.}^{nonbond}\frac{V_j}{r_{ij}^4}CCF\ . \tag{55}$$

A physical interpretation of “Type 1” coupling is that the environment atoms recognize the weighted ligands, but each ligand sees environment atoms and itself with no weighting for the calculation of the effective Born radius. From **Eqn. 54**, the effective Born radii of the environment atoms directly depend on {λ}.

Type 2

A second coupling scheme to be considered is given in **Eqn. 56**.

$$G_{pol}=-166\left(1-\frac{1}{\varepsilon}\right)\left\{\sum_{k=1}^{L}\lambda_k^2\sum_{i}^{i\in\left\{\begin{smallmatrix}env\\k\end{smallmatrix}\right.}\sum_{j}^{j\in\left\{\begin{smallmatrix}env\\k\end{smallmatrix}\right.}\frac{q_iq_j}{\sqrt{\left(r_{ij}^2+\alpha_i^k\alpha_j^ke^{-D_{ij}^k}\right)}}\right\} \tag{56}$$

where α_i^k represents the effective Born radius of environment atom *i* interacting with only the *k*-

th ligand. Every environment atom has *L* effective Born radii (α_i^m, *m=1, L*), in contrast, each atom belonging to the ligand has one effective Born radius. The effective Born radius is given by

$$G_{pol,i}^{k} = \frac{1}{\lambda_\alpha}\left(\frac{-166}{R_{vdW,i}}\right) + P_1\left(\frac{166}{R_{vdW,i}^2}\right) + P_2 \sum_{j\in\left\{\begin{matrix}env\\k\end{matrix}\right.}^{bond} \frac{V_j}{r_{ij}^4} + P_3 \sum_{j\in\left\{\begin{matrix}env\\k\end{matrix}\right.}^{angle} \frac{V_j}{r_{ij}^4} + P_4 \sum_{j\in\left\{\begin{matrix}env\\k\end{matrix}\right.}^{nonbond} \frac{V_j}{r_{ij}^4} CCF \ , \qquad (57)$$

$$\alpha_i^k = -166 / G_{pol,i}^k \ . \qquad (58)$$

The effective Born radii of the environment atoms are independent of {λ} in the “Type 2” definition so that we need not re-calculate them for each trial chemical state used in the MC procedure of CMC/MD simulations. On the other hand, “Type 1” is computationally more efficient than “Type 2” especially when the number of ligands increases because “Type 2” requires *L* separate calculations of the Born radius for all environment atoms. Both of these generalized Born coupling models have been incorporated into the program CHARMM for use with FEP, λ-dynamics and CMC/MD.

I will show the results of the free energy based screening methods combined with the GB implicit solvent in **Chapters 4** and **5**.

2.3.6 Inclusion of the non-electrostatic terms

In general, the total solvation free energy is given as the sum of electrostatic terms, based on the GB approach and non-electrostatic terms. Traditionally, non-electrostatic terms, consisting of the solvent-solvent cavity term (V_{cav}) and the solute-solvent van der Waals term (V_{vdW}), are linearly related to solvent-accessible surface area (SASA):[65,67]

$$V_{SA,i}(X, x_i) = V_{cav} + V_{vdW} \approx \sum_{i\in X, x_i} \sigma_i SASA_i \ , \qquad (59)$$

where $SASA_i$ is the total solvent-accessible surface area of atom *i* and σ_i is an empirical atomic solvation parameter for atom *i*. $V_{SA,i}$ represents the non-electrostatic terms only when ligand *i* and environment atoms are considered. Since calculating the *SASA* and its first derivative at every MD step is inhibitive, it is more efficient for the non-electrostatic terms (V_{SA}) to be ignored during the simulations and then their influence evaluated as a post-processing step. The total Hamiltonian (*H'*), including non-electrostatic terms, may be written as

$$H^{'}(X,\{x\},\{\lambda\}) = H_0(X,\{x\},\{\lambda\}) + \sum_{i=1}^{L} \lambda_i^2 V_{SA,i}(X,x_i) \tag{60}$$

where H_0 (see **Eqn. 17**) is the reference Hamiltonian excluding the non-electrostatic potential, in which the actual simulations are carried out. If *H'* and H_0 are regarded as the unbiased state and biased state respectively, the non-electrostatic terms ($-V_{SA}$) can be recognized as an umbrella potential and their effect may be calculated using the umbrella sampling technique[26,27] in the λ-dynamics or CMC/MD trajectories. The probability for the unbiased state *P'* can be written as follows using umbrella sampling formalism,

$$P^{'}(\lambda_i^2 = 1, \{\lambda_{m\neq i} = 0\}) = \frac{\int \exp\left\{-\beta\left(V + \sum_{m=1}^{L} \lambda_m^2 V_{SA,m}\right)\right\}\sigma(\lambda_i^2 - 1)}{\int \exp\left\{-\beta\left(V + \sum_{n=1}^{L} \lambda_n^2 V_{SA,n}\right)\right\}},$$

$$= \frac{\left\langle \exp(-\beta V_{SA,i})\sigma(\lambda_i^2 - 1)\right\rangle_{H_0}}{\left\langle \exp(-\beta \sum_{n=1}^{L} \lambda_n^2 V_{SA,n})\right\rangle_{H_0}}, \tag{61}$$

where the angular brackets represent the ensemble average over the biased state (H_0), *V* is the total potential energy that excludes the non-electrostatic terms, and the function $\theta(\lambda_i^2\text{-}1)$ is a step function which is one when its argument is zero but otherwise zero. The binding free energy difference ($\Delta\Delta G'$) between ligand *i* and *j* in the unbiased state (*H'*) can be obtained from

$$\begin{aligned}\Delta\Delta G'_{j\to i} &= -k_B T \ln \frac{P'(\lambda_i^2 = 1, \{\lambda_{m\neq i}^2 = 0\})}{P'(\lambda_j^2 = 1, \{\lambda_{m\neq j}^2 = 0\})} = -k_B T \ln \frac{\left\langle \exp(-\beta V_{SA,i})\theta(\lambda_i^2 - 1)\right\rangle_{H_0}}{\left\langle \exp(-\beta V_{SA,j})\theta(\lambda_j^2 - 1)\right\rangle_{H_0}}, \\ &= -k_B T \ln \frac{\sum\limits_{(\lambda_i^2=1)}^{(H_0)} \exp(-\beta V_{SA,i})}{\sum\limits_{(\lambda_j^2=1)}^{(H_0)} \exp(-\beta V_{SA,j})},\end{aligned} \tag{62}$$

where the summation of *exp(-βV$_{SA,i}$)* is carried out where ligand *i* takes the dominant state in the simulations of the biased state (*H$_0$*). By using **Eqn. 62**, the effect of non-electrostatic terms can be estimated from the λ-dynamics or CMC/MD simulations without their direct inclusion. The effect of non-electrostatic terms can be simply included in the FEP method without the umbrella sampling technique. We consider non-electrostatic terms only at the end points, whereas the actually sampled intermediate states did not contain them in our FEP simulations.

In **Chapter 4**, the effect of the non-electrostatic terms has been evaluated by using **Eqn. 62**.

2.3.7 The λ-dependent partial charge model

When the ligands to be simulated have invariable atoms, the hybrid topology model, in which invariable atoms are represented by a single topology, is anticipated to be more efficient than the multiple topology model. The hybrid topology representation avoids uninteresting configurational entropy contributions near end points, which requires additional efforts such as the introduction of "ghost forces"[15] or a restraining potential.[39] The invariable atoms can be assigned constant bonded and van der Waals parameters, but they may have λ-dependent partial charges in order to provide the proper partial charges at end points. For this purpose, we introduced a λ-dependent partial charge model along with the hybrid topology representation. The system is divided into three sets of atoms: (1) the variable part of ligands (x_i); (2) co-located atoms (colo atoms), that are invariable parts of the ligands whose partial charge depends on λ (*r*); (3) the rest of atoms, called “environment atoms” (*X*). The partial charges of the colo atoms depend on {λ} so that the Coulomb potential *in vacuo* and the GB energy related to them are altered according to the movement of {λ}, while any bonded and van der Waals terms are independent of {λ}. The hybrid potential function shown in **Eqn. 14** is re-constructed as follows:

$$V(X,r,\{x\},\{\lambda\}) = V_{env}(X,r) + V_{colo}(X,r,\{x\},\{\lambda\}) + \sum_{i=1}^{L} \lambda_i^2 (V_i(X,x_i) - F_i)\,. \qquad (64)$$

Here, V_{env} is the interaction energy involving the environment atoms only or the non-electrostatic energy excluding any variable part of ligands. $V_i(x)$ is the interaction energy involving ligand *i* except for electrostatic terms involving colo atoms. λ_i is the coupling parameter associated with ligand *i*. When the electrostatic terms are represented by the sum of the Coulomb energy *in vacuo* and the GB energy, the electrostatic terms involving the colo atoms (V_{colo}) are described as follows:

$$\begin{aligned} V_{colo} = {} & 332 \sum_{i \in colo} \left(\sum_{j \in env}^{nonbond} \frac{q_j \sum_{k=1}^{L} \lambda_k^2 C_i^k}{r_{ij}} + \sum_{j>i(j \in colo)}^{nonbond} \frac{\sum_{k=1}^{L} \lambda_k^2 C_i^k C_j^k}{r_{ij}} + \sum_{k=1}^{L} \lambda_k^2 \sum_{j \in k}^{nonbond} \frac{q_j C_i^k}{r_{ij}} \right) \\ & - 166 \left(1 - \frac{1}{\varepsilon} \right) \sum_{i \in colo} \left(2 \sum_{j \in env} \frac{q_j \sum_{k=1}^{L} \lambda_k^2 C_i^k}{\sqrt{\left(r_{ij}^2 + \alpha_i \alpha_j e^{-D_{ij}} \right)}} + \sum_{j \in colo} \frac{\sum_{k=1}^{L} \lambda_k^2 C_i^k C_j^k}{\sqrt{\left(r_{ij}^2 + \alpha_i \alpha_j e^{-D_{ij}} \right)}} + 2 \sum_{k=1}^{L} \lambda_k^2 \sum_{j \in k} \frac{q_j C_i^k}{\sqrt{\left(r_{ij}^2 + \alpha_i \alpha_j e^{-D_{ij}} \right)}} \right), \end{aligned} \qquad (65)$$

where the first term and second term correspond to the Coulomb energy *in vacuo* and the GB energy, respectively, and C_i^k is the partial charge of the *i-th* colo atom in ligand *k* when $\lambda_k^2 = 1$. The first derivative of V_{colo} with respect to {λ} is used for the propagation of {λ} in the λ-dynamics simulation. The λ-dependent partial charge model has been incorporated into the program CHARMM.

The application of the hybrid topology λ-dynamics approach using the λ-dependent partial charge model will be presented in **Chapter 5**.

2.3.8 The Monte Carlo / Langevin dynamics method for efficient sampling

A hybrid Monte Carlo and Langevin dynamics method (MC/LD), which is a modified version of the CMC/MD method, is introduced for efficient sampling of the ligands. In this method, the molecular system is divided into the environment atoms and the focus atoms. The atoms of interest, such as a ligand, are chosen as the focus ones that require enhanced sampling. Thus, the potential energy for the system is

$$V_{real}(X,x)=V_{env}(X)+V_{focus}(X,x)\,. \qquad (66)$$

Here, X and x stand for the coordinates of environment atoms and those of focus atoms, respectively, V_{env} is the potential energy involving the environment atoms only, and V_{focus} is the interaction energy involving the focus atoms. The focus atoms are represented as non-interacting multiple replicas. At any point along the trajectory, one of the replicas (the “selected copy”), and the environment atoms are sampled with the full force field Langevin dynamics (LD) method to generate the canonical ensemble of the system, while the rest of the replicas are propagated via LD with “ghost forces” to explore other local minima. The “ghost forces” due to the interaction between the environment atoms and unselected replicas are applied only on the latter. After every period of LD steps for propagating the atomic coordinates, a Monte Carlo sampling is applied to choose a new “selected copy” from all replicas, which helps to “tunnel” to a different minimum quickly without crossing the barrier. By properly coupling the system with a constant temperature heatbath, the configurational partition function of the real system, comprising the environment atoms and one selected replica, is

$$Z_{real}(X,x)=\int \exp\left[-\beta\left\{V_{\mathrm{env}}(X)+V_{focus}(X,x)\right\}\right]dXdx\,. \qquad (67)$$

The unphysical “ghost force” is derived from the potential $R_i(X, x_i)$. We assume that the probability function of the unselected replica *i* is proportional to *exp(-βR$_i$(X,x$_i$))*, however, the force derived from *R$_i$(X, x$_i$)* on the environment atoms are ignored. Using this approximation, the probability of choosing the trial move to the replica *i* with the coordinates x_i in an MC step can be written down as follows.

$$\alpha_i(X,x_i)=\frac{1}{L-1}\frac{\exp(-\beta R_i(X,x_i))}{\int\exp(-\beta R_i(X,x_i))dXdx_i}, \tag{68}$$

where *L* is the total number of replicas. In **Eqn. 68**, the first term corresponds to the probability of choosing replica *i* among all unselected replicas, while the second term corresponds to the probability of the replica *i* taking the coordinate x_i. Replication of the focus atoms along with an MC step provides an efficient method to explore the configurational space of the focus atoms. Because the forces on the environment atoms derived from {*R*} are masked, the coordinates of the unselected replicas may have a time lag with respect to the coordinates of the environment atoms. As a result, the "ghost force" may deviate from generating the probability distribution of the unselected replica as shown in **Eqn. 68**. Since the trajectory of the real system is generated via LD, small deviations during the MC step will be adjusted after a few LD steps. The detailed balance condition of one MC step between the selected replica *i* with coordinate of x_i and unselected replica *j* with coordinate of x_j can be described as follows.

$$\begin{aligned}&P_0(X,x_i)\alpha_j(X,x_j)A_{i\to j}(X,x_i,x_j)=P_0(X,x_j)\alpha_i(X,x_i)A_{j\to i}(X,x_j,x_i),\\&\exp(-\beta V_i)\frac{1}{L-1}\frac{\exp(-\beta R_j(x_j))}{\int\exp(-\beta R_j(x_j))dx_j}A_{i\to j}=\exp(-\beta V_j)\frac{1}{L-1}\frac{\exp(-\beta R_i(x_i))}{\int\exp(-\beta R_i(x_i))dx_i}A_{j\to i},\end{aligned} \tag{69}$$

where *P$_0$(X,x$_i$)* is the probability of the real system taking the coordinates, (*X*,*x$_i$*), and $A_{i\to j}$ is the acceptance probability of a move from replica *i* to *j*. When the same potential *R* was applied to all unselected replicas, $\int\exp(-\beta R_j(x_j))dx_j$ is equal to $\int\exp(-\beta R_i(x_i))dx_i$. Using this relationship, the acceptance probability is given by

$$\frac{A_{i\to j}}{A_{j\to i}} = \exp\{-\beta(\Delta V_{i\to j} - \Delta R_{i\to j})\} \tag{70}$$

where $\Delta R_{i\to j} = R_j(x_j) - R_i(x_i)$ and $\Delta V_{i\to j} = V_j(x_j) - V_i(x_i)$. Therefore, one acceptance rule [73] that obeys the detailed balance condition and leads to the canonical ensemble is

$$A_{i\to j} = \min(1, \exp\{-\beta(\Delta V_{i\to j} - \Delta R_{i\to j})\}). \tag{71}$$

The MC/LD method can be readily extended to be applied in various systems. For example, to enhance the acceptance ratio, the uniform probability (*1/(L-1)*) used for choosing the trial replica in the MC step (**Eqn. 68**) can be replaced by a weighted probability:

$$\alpha_{j\to i}(X,\{x\}) = \frac{\exp(-\beta V_i(X,x_i))}{W(j)} \frac{\exp(-\beta R_i(X,x_i))}{\int \exp(-\beta R_i(X,x_i))dx_i}, \tag{72}$$

where *W(j)* is the Rosenbluth factor, $W(j) = \sum_{k\neq j}^{L} \exp(-\beta V_k(X,x_k))$.[85] The acceptance criterion [73] that satisfies the detailed balance condition and leads to the canonical ensemble is

$$A_{i\to j} = \min\left(1, \frac{W(i)}{W(j)} \exp(\beta \Delta R_{i\to j})\right). \tag{73}$$

In another example, various types of “ghost forces” such as those derived from multicanonical ensemble,[86-90] or self guided methods[91,92] can be applied to explore a wider configurational space and efficiently detect local minima. In order to achieve both requirements, different “ghost forces” can be assigned where some replicas explore larger configurational space with the smoother “ghost forces”, and the others explore a smaller space with more realistic “ghost forces” to detect the local minima. The type of the “ghost forces” can be exchanged using the replica-exchange method.[93] In this application shown in **Chapter 7**, we adopted a scaled version of the actual potential as the potential between the unselected replicas and the environment atoms (i.e., $R(X,x_i) = \alpha V_{focus}(X,x_i)$).

In a third example of the expandable character of MC/LD, the focus atoms can be chosen manually. In a system including explicit solvent molecules, exhaustive sampling of the solvent configurations yields little information about the solute of interest. Hence, special techniques must be utilized[94] to limit the enhanced sampling only to those regions of the solute. Locally enhanced sampling (LES), which has already been applied to a variety of problems,[95-102] can also choose the focus region manually. However, the focus region should be chosen in order not to yield large errors from the real system since the environment atoms feel the mean force from all replicas in LES.

The combinations of Monte Carlo methods and molecular dynamics have already been applied to improve configurational sampling, such as the hybrid Monte Carlo technique[103] and MC(JBW)/SD method.[104-106] In the hybrid Monte Carlo method, molecular dynamics provides trial move configurations that are then evaluated with Metropolis Monte Carlo criteria to generate a thermodynamic ensemble. The MC(JBW)/SD method uses Monte Carlo steps to jump between conformational minima that are separated by free energy barriers. MC(JBW)/SD requires knowledge of the local minima prior to the simulations. On the other hand, in the MC/LD method, the unselected replicas explore the local minima by themselves. Thus, complicated systems such as protein-ligand complexes, where it is difficult to know the local minima beforehand, are appropriate for the MC/LD method.

An application of the MC/LD method will be presented in **Chapter 7**, where the binding orientations of toluene in β-cyclodextrin are efficiently explored.

Chapter 3

Cytochrome c Peroxidase - Heterocycle Derivatives

Based on

Shinichi Banba and Charles L. Brooks, III,

“Free energy screening of small ligands binding to an artificial protein cavity,”

J. Chem. Phys., **113(8)**, 3423-3433 (2000).

Shinichi Banba, Zhuyan Guo, and Charles L. Brooks, III,

“Efficient sampling of ligand orientations and conformations in free energy calculations using the λ-dynamics method,”

J. Phys. Chem. B, **104(29)**, 6903-6910 (2000).

3.1 Introduction

A multiple topology representation is essential when the topologies or binding modes of the ligands are different. Furthermore, assignment of force field parameters for the common atoms that are defined by the single-topology is complicated because they may depend on $\{\lambda_i\}$. Therefore, extending the λ-dynamics method to a multiple topology representation is very important in applying it to the discovery of new therapeutic ligands. Unfortunately, the multiple topology model is known to yield the slow convergence near the end points of the transformation, λ =0 or 1 even in FEP simulations.[107,108] With a multiple topology model, when λ is close to zero, the ligands become detached from the system and may adopt unphysical high-energy states. The λ-dynamics simulation can trap specific ligands in local minima in λ space because the detached ligands (λ^2 ~0) take mainly high-energy states and seldom return to the important structures (lower-energy states). This phenomenon makes it difficult for the λ-dynamics method employing a multiple topology model to accurately estimate the binding free energy difference within a restricted computational time. Therefore, in this calculation a new restraining potential (**Eqns. 44-45**) is introduced to the λ-dynamics method to enhance the sampling efficiency in λ-space.

By introducing the restraining potential as shown in **Eqn. 44**, at unbound states (λ^2 ~ 0), the barriers between various binding modes are reduced due to the scaling of protein-ligand interactions with the scaling parameter α. This speeds up the exploration of orientational degrees of freedom. For the same reason, the conformational sampling of ligands is also enhanced. In drug design applications, the 3-D structure of a protein-ligand complex is often known for only one or two ligands. The initial structures of chemically related putative ligands are prepared by superimposing them with the known compound’s complex structure or by using computational methods such as docking. Although many docking algorithms and programs have been recently developed,[9,10,109-112] these approaches do not always give the true binding orientation or conformation.[112-114] Furthermore, the fact that the docking scoring function is often not general results in errors in estimating the binding affinity, making it difficult to identify the correct binding structure out of putative binding conformations detected by these docking methods.[115,116] Theoretically rigorous methods, such as FEP[12] or TI,[4] not only are computationally too intensive to become a practical tool in drug design, but they also require that the initial orientation of the ligand is close to its true bound orientation. Furthermore, these methods cannot be applied straightforwardly to ligands that have multiple binding modes. Consequently, there is no viable way to search a preferred binding orientation or conformation on the basis of free energy.

In this application, we are interested in applying the λ-dynamics method using a multiple topology model to identify and rank the tight binding ligands from a large number of compounds within a short simulation time. We also investigate the effectiveness of a newly introduced restraining potential. The results from the λ-dynamics method are compared with experimental binding affinity data and FEP calculations for a set of 10 heterocycle derivatives interacting with cytochrome c peroxidase. Furthermore, we also show that the λ-dynamics method using a multiple topology representation is capable of addressing the inefficient sampling problems and provides a means to explore binding orientations or conformations on a free energy basis.

3.2 Computational details

As shown in **Table 1**, the system we have studied is a set of small cationic molecules that bind to an artificial cavity created inside cytochrome c peroxidase (CCP).[117,118] CCP is an enzyme containing heme. To perform its function, the enzyme is first oxidized by hydrogen peroxide, then it oxidizes its substrate, cytochrome c. As an intermediate in this process, the tryptophan near the active site is oxidized to a free radical. In order to design an enzyme that oxidizes a specific molecule, mutagenesis was used to create an artificial cavity by replacing the tryptophan by a glycine residue. [117,118] X-ray crystallographic structures of the ligands complexed with the protein indicate that certain molecules bind specifically to the cavity. [117,118] The binding affinity of a series of structurally characterized protein-ligand complexes has been measured using an optical spectroscopy technique. The initial coordinates of the complexes were taken from the X-ray crystallographic structures (provided by Professor D. Goodin at The Scripps Research Institute). Although the 10 ligands are chemically very related and the protein structure is preserved, the binding orientations of the ligands observed in X-ray structures are different. **Figure 2** shows the initial structure for the λ-dynamics simulations prepared by superimposing all crystal structures. Chemically related compounds mostly take similar orientations, however one substitution sometimes changes the binding orientation (e.g. nmei → dime). Because of such changes, application of the single topology model is not appropriate even though assigning a single topology to the common atoms is possible.

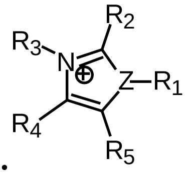


Table 1. The structures of five-member ring ligands.

Name	Z	R_1	R_2	R_3	R_4	R_5
3met	S	-	H	CH_3	H	H
2am4	S	-	NH_2	H	CH_3	H
2am5	S	-	NH_2	H	H	CH_3
34di	S	-	H	CH_3	CH_3	H
234t	S	-	CH_3	CH_3	CH_3	H
345t	S	-	H	CH_3	CH_3	CH_3
nmei	N	CH_3	H	H	H	H
nvi	N	$CH=CH_2$	H	H	H	H
2eti	N	H	CH_2CH_3	H	H	H
dime	N	H	CH_3	CH_3	H	H

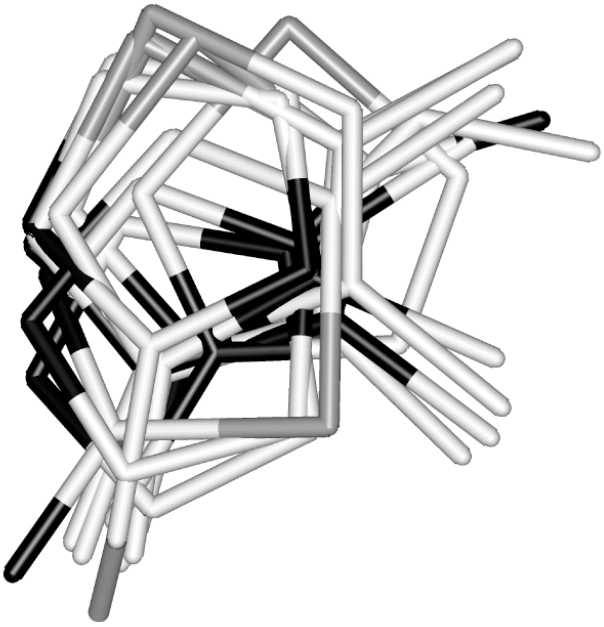


Figure 2. The initial orientations of the ligands in the protein-ligands model prepared by superimposing the C_{α}'s of the protein structures obtained by X-ray crystallography. Key: Nitrogen-black, Sulfur-gray, Carbon-white. Hydrogen atoms are not shown for clarity.

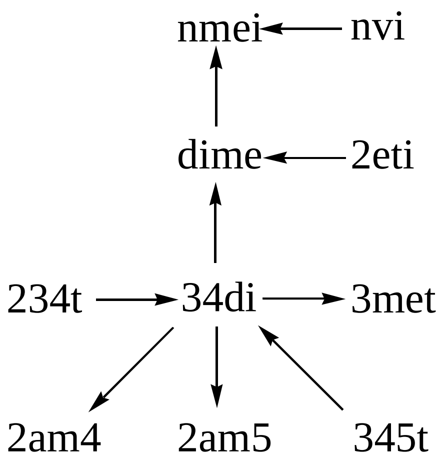


Figure 3. Pairing of ligands for the relative binding free energy (*ΔΔG*) calculations using the FEP method. The arrow shows the direction of chemical modification in the FEP calculation.

All computations were performed using the CHARMM molecular dynamics package and force field.[119] The missing parameters for the ligands, including bond, angle, dihedral and improper energies, were obtained based on QUANTA parameters for similar atom types and scaled up or down to be consistent with CHARMM parameter set. The charges of the ligands were obtained from quantum chemical calculations (GAUSSIAN 94, calculations was carried out at the US Army Research Laboratories by Drs. S.W. Bunte and G.M. Jensen). Water molecules were represented by the TIP3P water model of Jorgensen.[120] All bonds containing hydrogen atoms were constrained to their parameter values using the SHAKE algorithm.[121] Nonbonded interactions were treated using a cutoff of 12.4Å along with van der Waals switching between 8.0Å and 10.0 Å and electrostatic shifting functions. The temperature of the system was maintained near 300K by coupling the non-hydrogen atoms to a Langevin heatbath. The λ-dynamics and FEP simulations used a time step of 1fs and 1.5fs, respectively. Molecular dynamics simulation of the solvated protein-ligand complex, which was initially equilibrated in a 20Å sphere of water and then partitioned using a 12Å reaction zone and a 3Å buffer region, was carried out using the stochastic boundary molecular dynamics method.[122] The final solvated protein-ligands system contained 1677 protein atoms, one heme group, 21 crystal water molecules, and 88 solvent water molecules. The λ-dynamics calculations were carried out with the solvated protein-ligands system including 4,6, or 10 ligands for 300ps with the initial 30ps used as an equilibration phase. The λ trajectories were saved every 15fs and were used for later analysis. The 4-ligand system included four imidazolium derivatives (i.e. 2eti, dime, nmei, and nvi), while the 6-ligand system included the six thiazolium derivatives (i.e. 234t, 2am4, 2am5, 345t, 34di, and 3met). In this study, we are interested in not only identifying the better binders from a number of putative candidates but also the assessment of the λ-dynamics method with a multiple topology model. Therefore, it is important to have a large overlap of energy distributions of the ligands in order to obtain sampling of multiple ligands in the λ-dynamics simulation. From earlier results with the trypsin-benzamidine derivatives system, it has been observed that molecules with a binding free energy 5 kcal/mol higher than the lowest one would not compete.[14] Short λ-dynamics simulations showed that one ligand, 2am5, was dominant throughout the simulation and the other ligands showed no sign of competing. Therefore, a 5 kcal/mol bias was added to the F_i of 2am5 in order to render the other ligands competitive in all λ-dynamics simulations. The effect of the constant bias was removed by subtracting 5 kcal/mol from $\Delta\Delta G$ of 2am5.

In order to explore the efficiency of sampling by the λ-dynamics method, we also prepared the initial structures with putative orientations that mimic the actual situation wherein the 3-D structures of protein-ligand complex were known only for some ligands. The X-ray crystallographic structure of 234t was chosen as a reference and the structures of the other ligands were generated by superimposition of their five membered rings. Thus, the generated binding orientations of 2am5, nvi, and nmei were very different from their X-ray crystallographic structures, while those of the other ligands were close to the conformation found in their crystallographic structures.

To assess the results of the λ-dynamics simulation method, we also performed conventional FEP calculations with the solvated protein-ligand complex system. To get *ΔΔG* between all 10 ligands, FEP calculations should be carried out with nine ligand pairs as shown in **Figure 3**. The nine pairs were decided by considering chemically related ligands and similarity of their binding orientations. The pair (34di and dime) between the thiazolium and imidazolium derivatives was selected due to the large overlap of their van der Waals volumes. The larger volume ligand was defined as reactant in each pair. We performed simulations with λ=0.03125, 0.125, 0.325, 0.5, 0.675, 0.875, and 0.96875, respectively. The free energy change for each simulation was calculated using the double-wide sampling technique.[4] For each λ, a 30 ps equilibration period was followed by 120 ps of data collection.

As shown in **Eqn. 14**, in order to perform the λ-dynamics calculation of the relative binding affinity of the ligands, the free energy of the ligands in the unbound state (F_i in **Eqn. 14**) has to be predetermined. This is also done using the FEP simulation method. In the solvation FEP simulations, the systems consisted of the selected two ligands and 498-503 water molecules in 24.8 Å cubic box with periodic boundary conditions. The same nine pairs shown in **Figure 3** were also chosen to determine the set of F_i values. Water molecules whose oxygen overlapped within 2.8 Å of any non-hydrogen atoms in the solute molecule were removed. We performed simulations with λ = 0.125, 0.5, and 0.875, respectively. For each λ, a 30 ps equilibration period followed by 60 ps of data collection was performed.

3.3 Results and discussion

3.3.1 Use of a restraining potential to enhance λ-space sampling

For the 6-ligand system that includes 234t, 2am4, 2am5, 345t, 34di and 3met, λ-dynamics simulations with and without the restraining potential (**Eqn. 44**) were carried out. While the energy of the dominant ligand was low, the potential energy of the other ligands increased rapidly in the early stages of the λ-dynamics simulations without the restraining potential as shown in **Figure 4**. The unbound ligands (λ^2~0) finally reached a high-energy state causing instability in the integration algorithm. As the energy difference (V_i –F_i) governs the movement of λ_i, the ligand selected initially remains dominant throughout the simulation. The distributions of (V_i-F_i) also indicate that overlap between the potential energy distribution of the dominant ligand and the others is very poor. Such a poor overlap prevents the λ-dynamics simulation from changing the dominant ligands. Addition of the restraining potentials resulted in considerable overlap in the distribution of (V_i-F_i) of the ligands and enhanced the λ-space sampling as shown in **Figure 4**. Furthermore, stability of the λ-dynamics increases dramatically and no simulations yielded instability in the integration algorithm. These results clearly show that restraining potentials are required to avoid trapping in local minima in the λ space when we adopt the multiple topology representation.

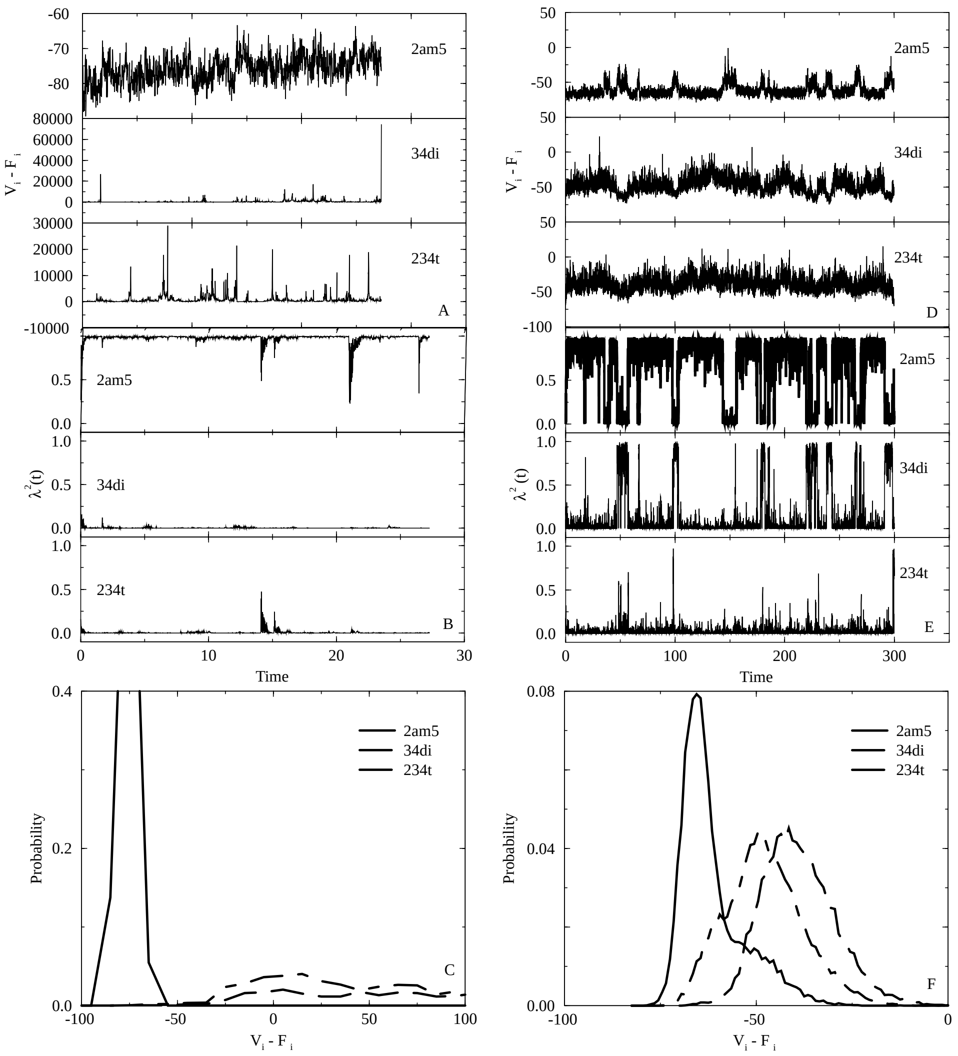


Figure 4. Comparison of sampling in λ-dynamics when different restraints are used. (A-C) Data from a λ-dynamics run with a 6-ligand system and no restraining potential. Only results for the three ligands (234t, 2am5, and 34di) are shown for clarity. The trajectories of the energy differences in kcal/mol (V_i -F_i) (Eqn. 14) and λ_i^2 are shown in (A) and (B), respectively. (C) shows the distribution of (V_i - F_i). The simulation was terminated after 27ps due to instability of the numerical integrator. A bias of 5 kcal/mol was subtracted from F_i of 2am5 to make the other ligands competitive. (D-F) The simulation conditions, except the addition of the restraining potentials, are identical to the run shown in (A-C).

3.3.2 Relative binding free energy estimation by λ-dynamics

In the binding affinity calculations, as shown in **Eqn. 46**, the relative binding free energy of a ligand is related to the amount of time that the ligand has λ^2=1 during a λ-dynamics simulation. Since λ is a continuous variable, we need to select a cutoff value to approximate the λ^2=1 state. Here we chose the cutoff as λ^2=0.8. From each λ trajectory, the number of times that a ligand reaches λ^2>0.8 can be computed. The relative binding free energy is calculated according to **Eqn. 46**. Correction terms as shown in **Eqn. 43** also have to be calculated. According to **Eqn. 43**, the average of the restraining potential of ligand *i* when λ_i =0 corresponds to its correction term. We assume that λ=0 if λ^2 is smaller than 0.05. In order to verify the effect of the cutoff values, the relative binding free energy differences estimated with two different cutoffs (0.8 and 0.9) for the dominant state were compared. They were in very good agreement (correlation coefficient = 0.997, slope = 1.02 for α=0.3). Moreover, a different cutoff (0.05 and 0.1) for correction terms also gave good correlation to $\Delta\Delta G$ (correlation coefficient = 1.00, slope = 0.998 for α=0.3). These results showed that the free energy landscape in the λ-dimensions is smooth at least for the CCP system.

As shown in **Figure 5**, a short time λ-dynamics simulation with the 10-ligand system successfully estimated the binding free energy differences as compared with those from FEP simulations and experiment. The scaling parameter α (**Eqn. 44**) also should be selected to obtain optimal sampling in the λ-space. To assess the effect of α, five different values (0,1, 0.2, 0.3, 0.5, and 0.75) were used as α. As shown in **Table 2**, all α values except for 0.1 provided reasonable estimates of the relative binding free energy as compared with FEP results. A small α value only weakly restrains the unbound ligands in low-energy regions. Therefore, by using a very small α value (e.g. α=0.1), the λ-dynamics simulation was trapped in local minima in λ-space and failed to yield converged $\Delta\Delta G$ values (data are show) within a restricted computational time. On the other hand, a large α value strongly biases the unbound ligands to low-energy regions. The total probability that any ligand stays within λ^2> cutoff decreases rapidly as α increases, since large α values stabilize the intermediate states. Even though the probability of effective sampling, in which any ligand reaches the dominant state (λ^2>cutoff), is very small when α=0.75, good agreement with the FEP result was obtained. The correction term mainly contributes to $\Delta\Delta G$ when α is large, and the statistical error in the probability term shown in **Eqn. 46** can be mostly ignored when α=0.75.

Table 2. Summary of relative binding free calculations.[a]

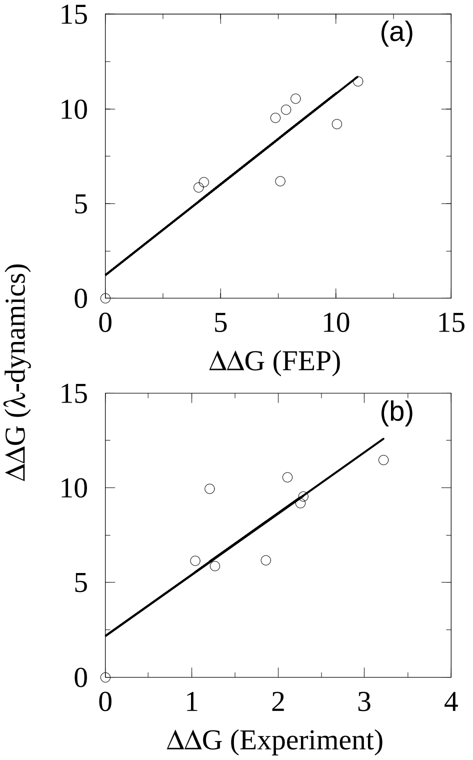
ligand	ΔΔG(bind) (Exp.)	ΔG(free)[b,c] (FEP)	ΔG(bound)[b] (FEP)	ΔΔG(bind) (FEP)	ΔΔG(bind), α=0.2 (λ-dynamics)	ΔΔG(bind), α=0.3 (λ-dynamics)	ΔΔG(bind), α=0.5 (λ-dynamics)
234t	3.22	18.00	28.96	10.96	10.57	11.46	11.90
2am4	2.11	-23.90	-15.65	8.25	9.12	10.55	8.46
2am5	0	0	0	0	0	0	0
345t	2.29	10.63	18.02	7.39	8.42	9.53	9.80
34di	1.21	-4.14	3.71	7.85	6.85	9.96	10.76
3met	2.26	-7.67	2.40	10.06	8.04	9.20	8.50
2eti	2.79	15.55	25.47	9.92	N.D.[d]	N.D.[d]	11.29
dime	1.04	12.89	17.18	4.28	5.04	6.15	6.43
nmei	1.27	8.66	12.71	4.05	6.03	5.87	6.47
nvi	1.86	23.39	30.99	7.60	6.63	6.18	7.58

[a] Free energy changes are in kcal/mol and relative to 2am5.

[b] Statistical uncertainities are ~±0.5 kcal/mol for all FEP calculation results.

[c] Free energy half-cycle with ligand free in solution.

[d] Not determined because they does not reach λ^2>0.8 during the entire simulation time.



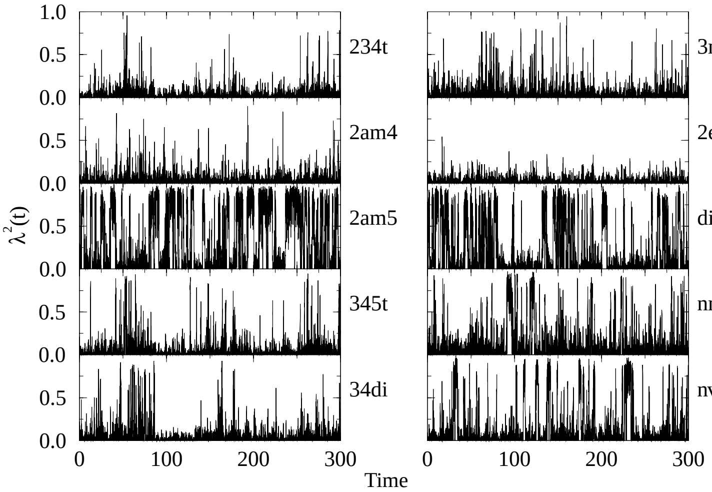


Figure 5. Comparison of relative binding free energy of the ligands between the λ-dynamics and FEP method (a) or experimental results (b).

Figure 6. Trajectory of λ^2 values for 10 ligands with α=0.3.

To demonstrate how the λs evolve during the simulation, the λ-trajectories are shown in **Figure 6** (α =0.3). The tight binding ligands like 2am5 and dime have sampled the dominant states often enough to give converged *ΔΔG*s, while weak binding ligands such as 234t or 345t have not been sampled enough. 2eti does not reach λ^2>0.8 during the entire simulation time (noncompetitive) when α = 0.3. 2eti is considered to have a higher binding free energy than the others, according to our λ-dynamics calculations, which is consistent with the FEP results

3.3.3 Convergence of free energies with the restraining potential

As shown in **Eqn. 46**, relative binding free energies are estimated by two terms (i.e. probability term and the correction term). The cumulative running average of each term is shown in **Figure 7**. The *ΔΔG*s of poorly binding ligands show large jumps occasionally due to their poor convergence of the probability terms, but the probability terms for better ligands reach constant values within 270ps of simulation time. The convergence of the correction terms (U_i*'*) seems to be faster than that of probability term.

A 1350 ps λ-dynamics simulation with 10-ligand system was carried out with α =0.3 to verify the convergence. The result was divided into five segments (0-270ps, 270-540ps, 540-810ps, 810-1080ps, 1080-1350ps) and *ΔΔG*s were estimated independently in each segment. The five *ΔΔG*s estimated in each segment were used to get average *ΔΔG* and statistical deviation as shown in **Figure 8**. In general, strongly binding ligands tend to have small statistical errors, while *ΔΔG* of weakly binding ligands include large statistical errors as predicted by the running average of *ΔΔG*. Nevertheless, a 270ps simulation is enough for the λ-dynamics method to screen out the tight binding ligands from the putative candidates. If we want to estimate accurate binding free energy differences efficiently even with poor ligands, constant biasing potentials are required.

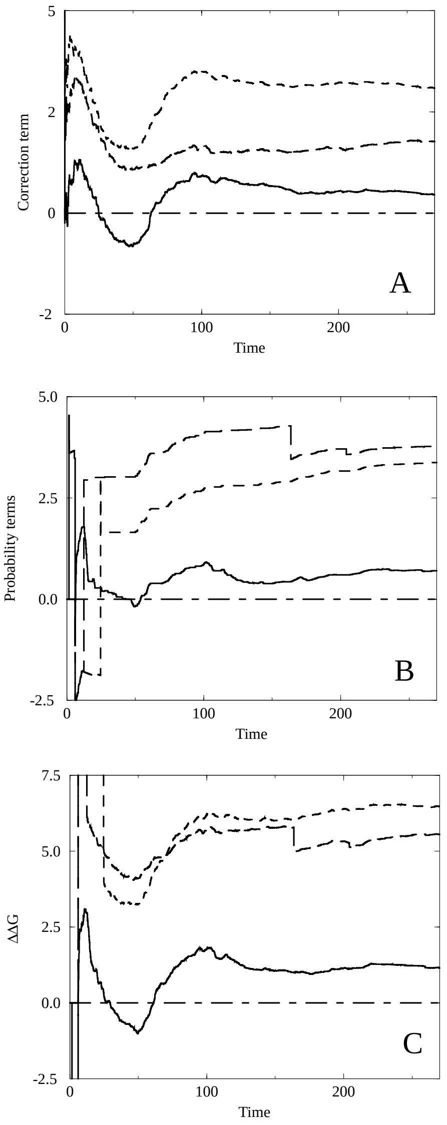


Figure 7. The cumulative averages of the probability term, the correction term (see Eqn. 46) and the total *ΔΔG* (in kcal/mol) from a λ-dynamics simulation of a 10-ligand system. The biasing potential of 2am5, which is selected as reference, is not removed. A value of 0.3 was used for *α*. Only the results of four ligands (234t, 2am4, 2am5 and dime) are shown for clarity. Key: 234t-dotted line, 2am4-dashed line, 2am5-dotted and dashed line, dime-solid line.

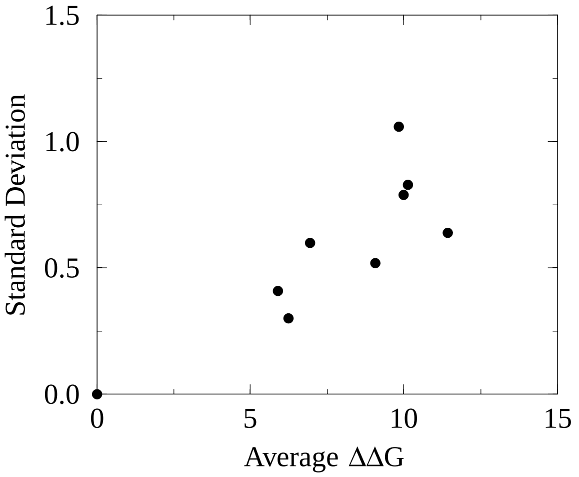


Figure 8. The relationship between the standard deviation and average *ΔΔG* (both in kcal/mol). The data were estimated from five 270ps λ-dynamics trajectories. 2eti was not included in the analysis. The standard deviation of *j*-th ligand was calculated as

$$s.d._j = \sqrt{\frac{\sum_{i=1}^{N} \left| \Delta\Delta G_j^i - \overline{\Delta\Delta G_j^i} \right|^2}{N-1}}$$

***N* indicates the number of trajectories; $\Delta\Delta G_j^i$ indicates the relative binding free energy of *j*-th ligand from the *i*-th trajectory.**

To further explore the convergence of $\Delta\Delta G$ with the λ-dynamics method, we also carried out a λ-dynamics simulation using a system that contained ten copies of the ligand 34di. The free energy differences among these ligands should be zero if the λ-dynamics simulation had fully converged. As shown in **Figure 9**, when α=0.5, convergence is slow. The total probability of effective sampling, where any ligand remains at λ^2 > cutoff is very small with α=0.5 because a large α value keeps all ligands in very low-energy states and intermediate states are relatively stable. Especially when all ligands are identical, the intermediate states have similar potential energies. With a small α value like 0.2 or 0.3, $\Delta\Delta G$ between the most favored and least favored reaches 0.6 kcal/mol after only 270ps sampling. These small statistical errors are comparable with those in FEP calculations, whereas the simulation time required for the λ-dynamics simulation is much smaller than for FEP method. The results of $\Delta\Delta G$ estimated by the λ-dynamics method indicate that the choice of α influences the convergence of $\Delta\Delta G$, but does not affect the $\Delta\Delta G$ itself. The proper choice of the α value is important for faster convergence because it influences the ratio of the effective probability that ligands are in the dominant state and the change of the dominant ligand. The optimal value of α can be selected by a short time test simulation with the system that contains the identical ligands.

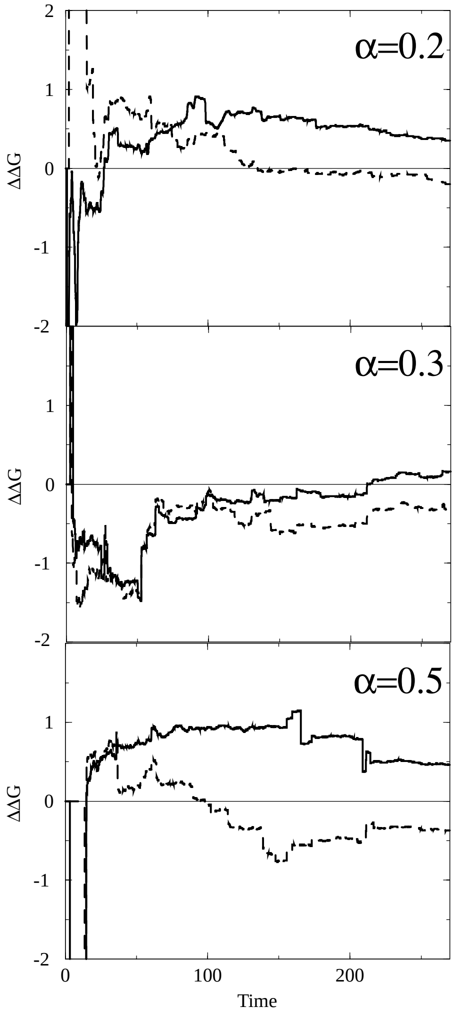


Figure 9. The cumulative running averages of the total *ΔΔG* (in kcal/mol) from a λ-dynamics simulation with ten copies of the ligand (34di). One of the ten 34di ligands was chosen as reference. The results for two ligands, which had the highest and the lowest *ΔΔG*, are shown.

3.3.4 Orientational motion of ligands inside the pocket

In the course of the λ-dynamics trajectory, we noted significant motion of the ligands within the binding pocket. We thus examined the orientational sampling of the ligands in the binding pocket. The efficient sampling of ligand orientations during a simulation is important for a good estimate of ligand binding free energy. **Figure 10** gives the distribution of the dipole moment of the ligands projected along a fixed (lab frame) direction. Since the dipole moment is somewhat ill-defined and coordinate system dependent when the total charge of each ligand is not zero, in this analysis, the center of geometry of the selected ligand atom was used as origin for the coordinate system in which the dipole moment is calculated. It is clear that from **Figure 10** that there are two dominant orientations for the ligand 34di in the bound state (λ^2>0.8), the peak to the left is in alignment with the crystallographic binding orientation, while the one to the right is not. Judging from the relative bound state population of the two orientations, which could be used to determine the relative binding free energy of the two modes, the crystallographic binding orientation is more stable by about 0.4kcal/mol than the alternative binding orientation detected in the λ-dynamics simulation. However, the alternative binding orientation is also stable enough to contribute to the binding free energy since 34di reached the λ^2 threshold in this alternative orientation as well. Although the other ligands adopt dominant states (λ^2>0.8) only when their binding orientations are close to X-ray orientations, the λ-dynamics method also explores a larger ligand orientational space than conventional MD as shown for nvi in **Figure 10-b**. **Figure 11-a,b** illustrate snapshots of X-ray and the alternative binding orientation sampled for 34di, respectively. The coordinates were extracted from the portions of the λ-dynamics trajectory having λ^2>0.8 for this ligand, which corresponds to the ligand in the bound state. The relative orientations of the ligand are obtained by superimposing the protein backbone structures. The root mean square deviation of protein backbone is ~0.6Å. The presence of the alternative orientation in 34di proposed here awaits experimental verification.

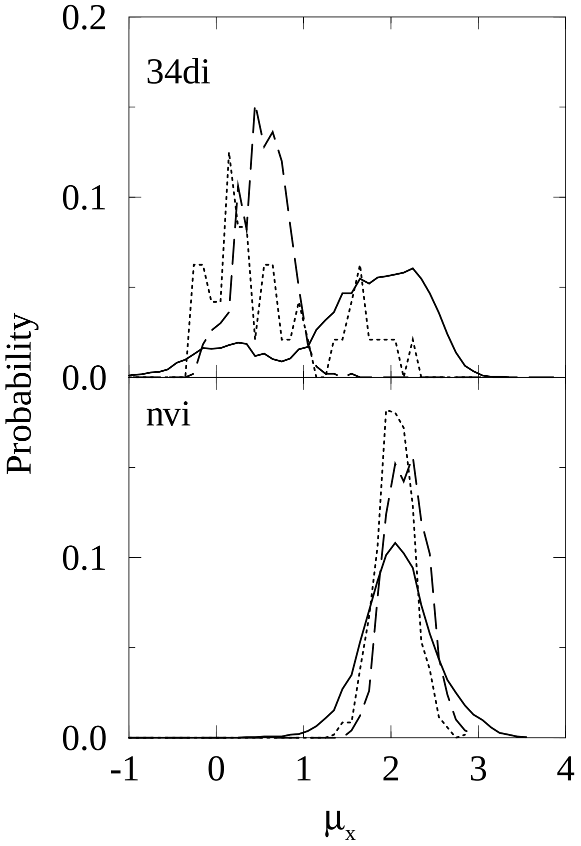


Figure 10. Distribution of the dipole orientation projected along a certain direction (X-axis) of 34di and nvi from a 270ps λ-dynamics trajectory for a 10-ligand system. The dipole moment of the initial X-ray crystallographic structures with 34di and nvi are 0.1 and 2.2, respectively. The dashed line shows the results of conventional MD including only a single ligand. The solid line shows data from the entire λ-dynamics trajectory. The dotted line shows the distributions that include only the bound states (λ^2>0.8) from the λ-dynamics trajectory. There are two dominant orientations with 34di; the X-ray crystallographic orientation and an alternative orientation. Sampling of the ligand orientation by conventional MD was restricted around the initial X-ray crystallographic structure.

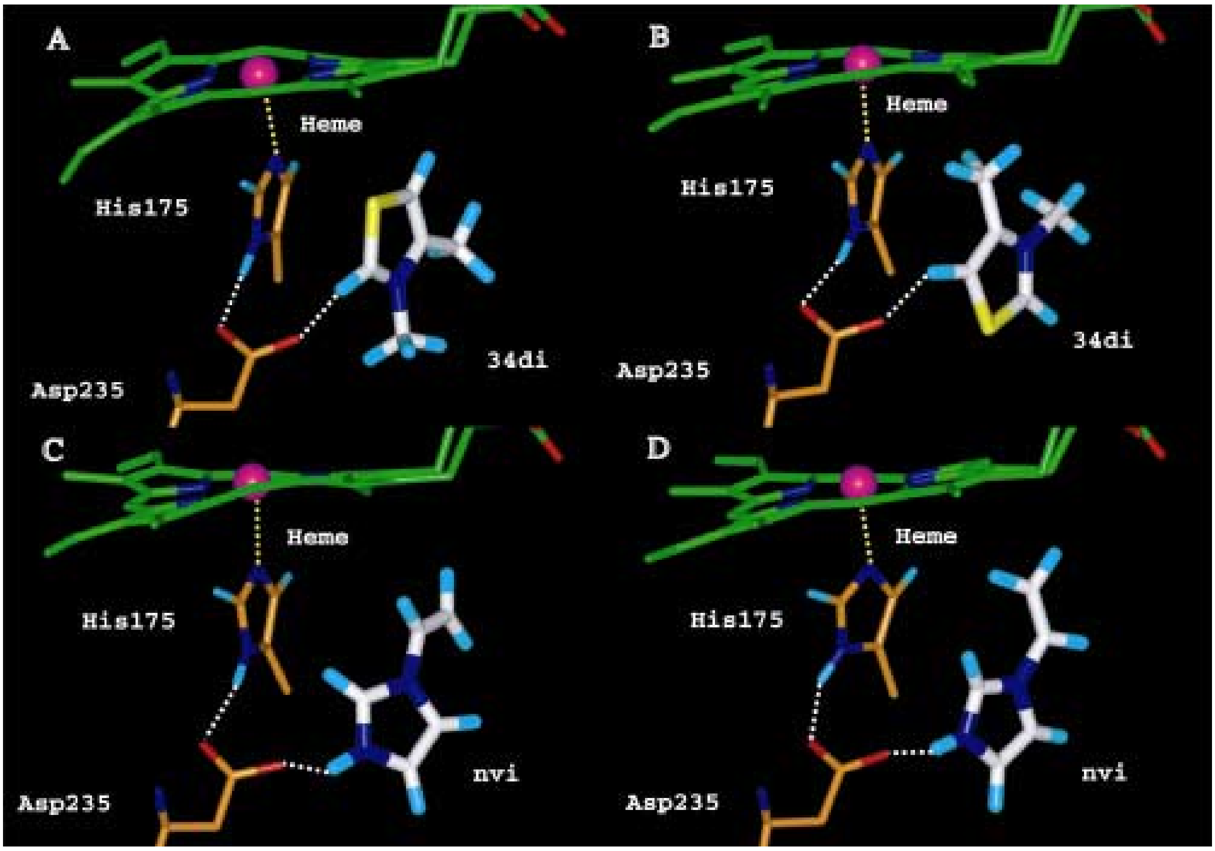
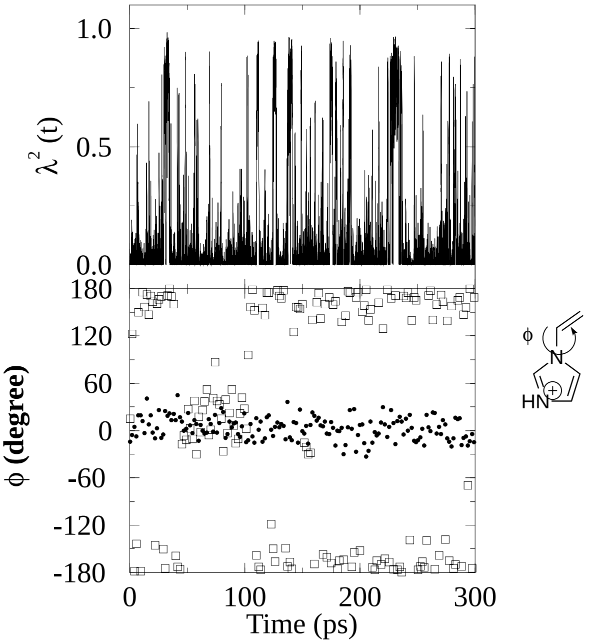


Figure 11. Snapshots of the binding modes for 34di and nvi inside the binding pocket taken from a λ-dynamics trajectory with λ^2 >0.8, which corresponds to the bound states. The X-ray binding orientation (a) and alternative binding orientation (b) are shown for 34di. (c) and (d) show the X-ray conformation and alternative conformation for nvi. Only the ligands, heme, Asp-235, His-175 are shown for clarity. The carbon atoms for heme, protein, and the ligands are shown in green, orange, and white, respectively. The oxygen, nitrogen, hydrogen are shown by red, blue, cyan, respectively. Iron atom is shown by the sphere.

3.3.5 Conformational sampling of the ligands inside the binding pocket

The sampling of ligands inside the protein cavity was also investigated. It should be noted here that we are not addressing the sampling of protein conformations. Although the protein is free to move, we are not focusing on the issue of protein conformational changes upon binding. The protein conformation remains close to its initial conformation during the λ-dynamics simulation (total backbone rmsd ~0.6Å). Following the same argument as in the previous section, the sampling of ligand conformations is enhanced due to the scaling of forces. Therefore, if a ligand has multiple binding conformations, the method will be more likely to identify them than regular dynamics simulations. Since the conformation of a molecule is determined by its torsional angles, we chose the torsional sampling of nvi as an example (see **Figure 12** for a definition of the torsion). This torsion seems to be appropriate for investigating the efficiency of conformational sampling because it has two local minima and the barrier between them is too high for conventional MD to sample both states within the restricted computational time. **Figure 12** shows that this torsional angle sometimes changes between two local minima ($\phi=0^o$ and 180^o) within a 300ps simulation time, while in conventional MD it stays in one local minimum ($\phi=0^o$) within the same simulation time. These results indicate that the sampling of torsional degrees of freedom is enhanced by the λ-dynamics method. Thus, the λ-dynamics calculation predicts an alternative binding conformation ($\phi=180^o$) for nvi because this ligand reached $\lambda^2>0.8$ (see **Figure 12**) at both the X-ray crystallographic conformation and $\phi=180^o$. The alternative binding conformation is shown in **Figure 11d**. It is difficult to get the free energy difference between these two conformations due to the restricted samplings. But, according to the force field that we used in this simulation, the alternative conformation seems to be more stable than the X-ray crystallographic conformation in our force field. The consideration of alternative conformations might lower the $\Delta\Delta G$ of nvi estimated by the λ-dynamics method over that estimated by the FEP method. Enhanced torsional sampling is an interesting feature of the method.



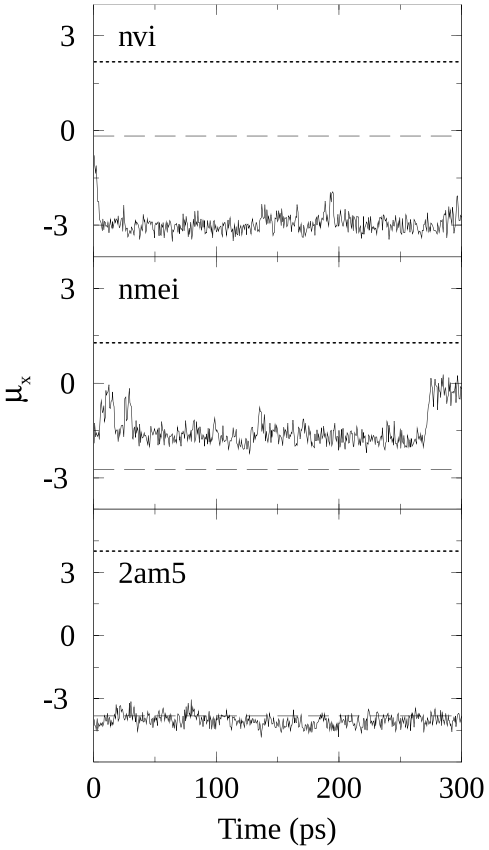


Figure 12. Change of torsion angle (ϕ) for nvi during simulations. The dot and squares show the results of a conventional MD simulation and λ-dynamics simulation, respectively. The torsion angle sometimes makes transitions between two minima (0° and 180°) only in λ-dynamics simulation. The trajectory of λ^2 for nvi is also shown.

Figure 13. Change of dipole orientation projected along the x-direction (μ_x) of 34di, nmei and nvi from a conventional MD simulation. Each simulation started from a putative and incorrect initial orientation. The dashed line and dotted line show the dipole orientation of the initial orientation and X-ray orientation, respectively.

3.3.6 Exploring stable ligand binding orientations with λ-dynamics

λ-dynamics simulations from putative (and incorrect) initial conformations were carried out to illustrate its sampling efficiency. To contrast the results from the λ-dynamics method, conventional MD simulations of single ligand-protein complexes were also carried out with three ligands (2am5, nmei, and nvi). The same initial structures were used for both the λ-dynamics simulations and the conventional MD simulations. As shown in **Figure 13**, conventional MD from our initial structures were all trapped in local minima near the initial structures and failed to move to the X-ray crystallographic binding structures. On the other hand, in the λ-dynamics simulation with the 6-ligand system, 2am5 was trapped in two local minima but reached the X-ray crystallographic orientation within 300ps (**Figure 14**). In a run with a 4-ligand system, which includes the 4 imidazolium derivatives, both nmei and nvi also reached the X-ray crystallographic structures within 300ps simulation (**Figure 14**). The 10-ligand system, including both thiazolium and imidazolium derivatives, was also tested to confirm sampling efficiency. The three ligands (2am5, nmei, and nvi) successfully reached their crystallographic orientations within the 300ps of the λ-dynamics simulation (data not shown). Moreover, these ligands reached the bound states (λ^2>0.8) only after they reached their X-ray crystallographic binding modes with all λ-dynamics simulations started from the incorrect orientations (**Figure 14**). These results clearly show that when a ligand adopts λ^2 values near zero rapid exploration of low energy orientations and conformations occur due to the scaling of the potential. Then, at a later time, fluctuations in the protein configuration, the ligand conformation, or both occur to induce a "binding mode" configuration of λ. Furthermore, these results also show that the λ-dynamics method can be applied to explore the docking of the ligands on a free energy basis.

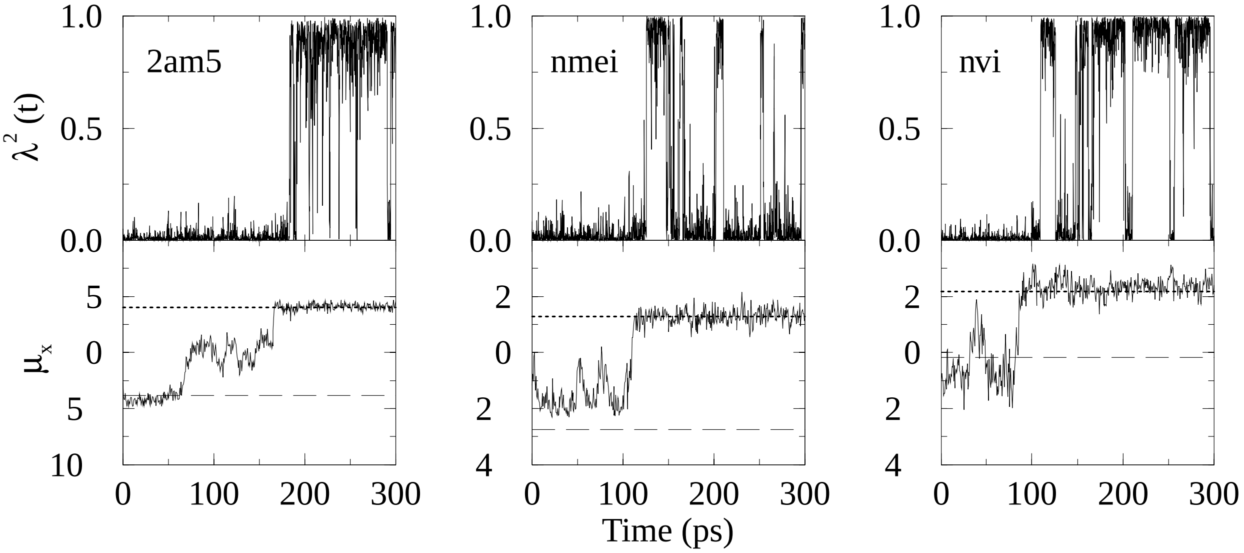


Figure 14. Change of dipole orientation (μ_x) of 34di when (a) 6 ligands are considered and (b) those of nmei and nvi with a 4-ligand system. The λ-dynamics simulations started with incorrect initial orientations for the ligands. The trajectory of the λ^2 values for 34di, nmei and nvi are shown at the same time. The dashed line and dotted line show the dipole orientation of the initial and X-ray orientation, respectively.

3.4 Conclusions

We have presented a set of the promising observations for ligand binding ranking and exploring ligand binding orientations and conformations using the λ-dynamics method. These include the consistency of the λ-dynamics calculations with FEP calculations and experimental results, the enhanced sampling of orientational and conformational degrees of freedom, and rapid search of binding orientations during a λ-dynamics simulation. The restraining potential is very effective and important for the application of λ-dynamics with a multiple topology representation. The λ-dynamics results were obtained at far less computational cost than FEP calculations due to addition of the restraining potentials.

The approximations embodied in **Eqn. 45** may be problematic if the ligand bound conformation of the protein is different for each ligand and large structural changes of the environment atoms take pace during binding. Such large motions in protein structure seldom happen within the limited sampling time of λ-dynamics simulations, while the movement of ligand atoms is enhanced by scaling the potential.[40,123] Therefore, multiple ligand free energy methods like λ-dynamics or CMC/MD[15,16] are basically limited to the groups of ligands in which the optimal binding conformation of the protein is similar for all ligands. Therefore, the approximation shown in **Eqn. 45** is valid in principle with the λ-dynamics method. To ameliorate this problem, all candidate ligands can be divided into a few groups which only have structurally similar ligands by using recently developing molecular similarity methods[124,125] and the λ-dynamics simulation can be carried out on the ligands in each of those groups. By partitioning the ligands into groups such that they have common members, $\Delta\Delta G$ for structurally dissimilar ligands belonging to the different groups also can be estimated. This strategy may overcome the limitation that binding modes should be similar in all ligands. Moreover, the alternative binding orientations detected by the λ-dynamics method were restricted to those that did not include a large conformational change in the protein structure because the interaction energy within environment atoms were not scaled by λ^2. To restrict the enhanced sampling region within the ligands is one of the merits in the λ-dynamics method because the expansion of the enhanced sampling region to environment atoms may result in the collapse of protein structure owing to slow convergence of $\Delta\Delta G$. Although large sampling space delay the convergence, enhanced sampling in environment atoms may be possible by redefining some of the environment atoms as multiple conformations that are scaled by using second coupling parameters.

For orientational and conformational sampling, this method is much more efficient than the FEP method. This is because when a ligand is not bound (small λ_i^2) to the protein, it rapidly explores its low energy binding orientations or conformations in order to be able to compete with the dominant ligand at a later time. These results show that the efficient sampling of ligand binding orientations in the λ-dynamics method removes the restriction that the initial orientation of the ligand inside the binding pocket must be close to its true bound orientation in order to get a reasonable estimate of binding free energy - a prerequisite for other free energy calculation methods. This feature is particularly important in drug lead discovery and optimization when the binding mode is unknown, or the modification of ligands causes the change of binding mode. It is also important in the case where a single ligand could have multiple binding modes. It may be true that enlarging sampling space delays the convergence. But many ligands compete at same time in the λ-dynamics method, and most of the time one of them reaches a low energy state and competes with the dominant ligand. It may be argued that the enlarged sampling space obtained by the scaling of the potential energies in the λ-dynamics method may contain the ligands in unphysical conformations or orientations. This phenomenon was observed in λ-dynamics without the restraining potential and resulted in the slow convergence.[39] By using this potential and the scaling parameter α (**Eqn. 44**), one can decide for oneself the extent to which the sampling space of ligand orientations and conformations is enlarged. These merits make it possible to achieve not only the calculations of relatively correct binding free energy differences but also the identification of alternative binding orientations and conformations in some ligands from a single simulation.

Chapter 4

Trypsin - Benzamidine Derivatives

Based on

Shinichi Banba, Komath V. Damodaran, and Charles L. Brooks, III,

“Free energy simulations with generalized Born implicit solvation I: formulation and exploratory application,”

J. Phys. Chem. B, submitted.

4.1 Introduction

As shown in the previous chapter, free energy based screening methods such as λ-dynamics[6,13-16,39-47,56] are very effective. These methods can be much faster than FEP or TI, and they give the same relative free energies within statistical errors. For example, the λ-dynamics method[13] was successfully applied to protein-ligands systems.[6,14,39,40] Despite its success and relative speed compared with FEP, the λ-dynamics method with explicit solvent still requires considerable computational time to obtain the correct ranking of ligands when the solvation environment for the ligand or the size of the ligands are different. This is because the re-orientation of solvent molecules is infrequent, and thereby inhibits the transition of the chemical states between the ligands. Since explicit solvation models require averaging of over a large conformational space to yield converged thermodynamic properties, slow convergence has been observed for many biological systems.[107,126] Thus, semi-quantitative models employing continuum solvation have been introduced as an intermediate approach. Finite difference solutions of the PB equation have been successfully applied for many systems.[127-129] For example, a continuum solvent approach using the PB equation was successfully applied to the ranking of ligands in trypsin[130], arabinose binding protein and sulfate binding protein systems.[63] However, the explicit numerical solution of the PB equation is also too costly to permit useful long time dynamics of biological molecules. Since the GB model is fully analytical,[65,76] derivatives of the energy with respect to individual atoms are available and allow the effects of solvation to be efficiently included in molecular dynamics. Furthermore, the GB models can be essentially as accurate as more elaborate finite difference PB calculations.[65,78,83] Therefore, MD simulations using a GB approach have recently been applied to many systems (e.g. estimation of pKa shift[77,79,80], binding affinities or binding structure of ligands[64,81], loop structure[82], and so on). These factors motivate us to apply a combination of λ-dynamics and the GB approach (λ-dynamics/GB) for elucidating better binding protein inhibitors while using a restricted amount of computational time. This is, to our knowledge, the first application of free energy calculation methods, such as FEP and the λ-dynamics methods ,with a GB model.

When one focuses on inhomogeneous systems, standard periodic boundary conditions are not always efficient. Many methods have been presented to address this problem.[122,131-137] For example, the stochastic boundary approach was successfully applied for the calculation of relative binding free energies with explicit water molecules.[6] Even though a continuum solvation model ignores explicit water molecules, inclusion of the whole protein is still computationally too expensive for semi-quantitative computational screening purposes. Therefore, an efficient boundary model with the continuum solvent representation is still useful for structure-based drug design. For this reason, we also assess whether a simple spherically truncated model with the GB solvent representation gives the correct ranking of the ligands in the trypsin-benzamidine derivatives system.

4.2 Computational details

4.2.1 System under study

The system studied was benzamidine and three of its para-derivatives, namely, p-amino benzamidine (p-NH_2), p-methyl benzamidine (p-CH_3), and p-chloro benzamidine (p-Cl) complexed with trypsin. The detailed force field for the ligands was described in a previous paper.[6] This system has been examined fully by the FEP method[7,46] and the λ-dynamics method[6] with explicit water. Thus, it is a good test case for the λ-dynamics/GB approach.

4.2.2 Dynamics simulations

All simulations used the CHARMM version 22 all hydrogen parameter and topology set.[138] Explicit water molecules were represented by the TIP3P model.[120] All bonds containing hydrogen atoms were constrained to standard values using SHAKE.[121] The temperature of the system was maintained near 300K by coupling the non-hydrogen atoms to a Langevin heatbath using a frictional coefficient 50ps^{-1}. The λ-variables were also coupled to a Langevin heatbath using a frictional coefficient 5ps^{-1} to keep the system near 300K in all λ-dynamics simulations. Nonbonded interactions were truncated using a switching function between 15Å and 8 Å for the explicit solvent calculations and between 10 Å and 8 Å for the GB model. The time step used in all simulations was 1.0fs. The masses of the fictitious λ degrees of freedom were chosen to be 5 amu•Å^2. The GB parameters fitting to single amino acids and proteins were used for the unbound ligand state and bound state, respectively.[83] We set all hydrogen radii in the CHARMM param22 parameter set to 1.5Å for the calculation of the effective Born radii as discussed in **Appendix A**. In all simulations including the GB energy, the Born radii and corresponding forces were updated at every MD time step, thereby ensuring the correct relationship between the energy function and its derivatives. All calculations were done using the CHARMM molecular dynamics package.[119]

The complete trypsin-benzamidine complex was used for assessment of the free energy simulation methods with the GB energy. After removing all crystal water molecules from the initial X-ray structure, the variant groups of the ligands and the hydrogen atoms were added according to the appropriate parameter/topology set. Following, the complex structure was minimized under successively reduced harmonic restraints for the heavy atoms. We adopted the final structure as an initial structure for all later simulations of the whole protein GB model. The final system contains 3250 protein atoms, 29 ligand atoms, and one calcium atom. After a 50ps equilibrium MD simulation of benzamidine bound to trypsin in the GB model, the ligand had shifted about 1.5Å from the X-ray structure, which was also observed in the explicit water simulation using either a 20 or 25Å spherical stochastic boundary. Therefore, this structural change is due not from the GB model but from possible inaccuracies in the assignment of the force field for the ligands or the nonbonded cutoffs. Our primary interest in this study is in assessing the free energy calculation methods with the GB model. Thus, to maintain closer structural correspondence with X-ray structure, all heavy backbone atoms were restrained to remain near this reference structure using harmonic restraints with a force constant of 4 kcal/mol/Å^2. The harmonic restraining potential applied to the backbone atoms is similar to those employed by others in MD or MC simulations in the same trypsin-benzamidine derivative system.[7,46] In the λ-dynamics and CMC/MD simulations, a 30ps equilibration period was followed by 200ps production runs.

In this study, the solvation free energy of the ligands in explicit water $\{F\}$ was taken from previous FEP calculations.[14] The solvation free energy of the ligands in GB solvent was calculated with one minimized structure. This is sufficient due to the inflexibility of the ligands.

In the FEP calculations of the bound state, three transformations - *p*-Cl to *p*-H, *p*-NH_2 to *p*-Cl, and *p*-CH_3 to *p*-NH_2 - were considered. A 30ps equilibration period was followed by a 60ps production run for every λ (0.125, 0.5, 0.875) using double-wide sampling to span the entire λ space.

In this study, a hybrid topology, in which invariable ligand atoms are represented with a single topology and the variant groups with a multiple topology, [14] was used so that the invariable ligand atoms are independent with scaling factors $\{\lambda^2\}$. The bonded terms (i.e. bond, angle and improper dihedral terms) related to the variant groups were not scaled by $\{\lambda^2\}$. When the variant groups are small, the hybrid topology and unscaled bonded terms are enough to keep the unselected ligands in the low energy states. The unscaled bonded terms are expected to cancel in the full thermodynamic cycle (i.e. the bonded terms were excluded in the solvation half cycle as well).In this application, variant groups were small so that any restraining potentials (**Eqn. 44**) were not added for the unselected ligands in the λ-dynamics and CMC/MD simulations The *ΔΔG*s are estimated by using **Eqn. 16**, instead of **Eqn. 46**.

4.2.3 The conditions for Chemical Monte Carlo / molecular dynamics

The CMC/MD method used in this study is basically the same as CMC/MD developed by Kollman group[15,16]. In the previous studies of Kollman, a “ghost force” and harmonic restraining potentials between the ligands’ centers of mass were applied for the unselected ligands to keep them in lower energy states and to avoid low acceptance ratios in MC steps. In this study, any “ghost force” and harmonic restraining potential are not included to the unselected ligands. As mentioned above, restraining potentials for the unselected ligands were also not included. Therefore, the variant groups of unselected ligands are evolved according to only non-scaled bonded interaction (i.e., bond, angle and improper dihedral terms) in the CMC/MD simulations.

One advantage of CMC/MD is that sampling the space of {λ} can be as restricted as we like, whereas an umbrella potential is required to control the chemical sampling in the λ-dynamics method. In this study, sampling along λ coordinates are restricted to only chemically important end points (see **Eqn. 25**) or $\{\lambda_i^2=0.5, \lambda_j^2=0.5, \lambda_{k\neq i,j}^2=0\}$. The λ-variables were sampled every 10fs. One Monte Carlo step for the chemical variables {λ} was attempted at every 10 MD steps of the atomic variables in the CMC/MD simulations. The CMC/MD method has been incorporated into the program CHARMM.

4.3 Results and discussion

4.3.1 Explicit water versus the generalized Born solvent

The relative binding free energy differences were calculated with the GB model using the whole protein. The relative binding free energy differences calculated by FEP, λ-dynamics and the CMC/MD method are tabulated in **Table 3**. The previous results from an explicit solvent simulation using a 20Å stochastic boundary condition are also listed in **Table 3** for comparison.[6]

Table 3. Results of relative binding free energy calculations in the whole protein system.[a]

R	FEP EW[b]	λ–dynamics EW[b]	FEP[c] GB *type1*	λ–dynamics GB,*type1*	λ-dynamics GB,*type2*	MC/MD[d] GB,*type2*	MC/MD[e] GB,*type2*
H	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NH_2	0.4	0.7	0.37	0.40	0.19	0.45	0.32
CH_3	2.3	1.9	0.52	0.41	0.25	0.49	0.37
Cl	2.2	2.6	1.55	1.03	1.51	1.36	1.44

[a] Free energy changes are in *kcal/mol* and relative to Benzamidine (R=H).

[b] The results were taken from the previous paper, [6] in which explicit water model with 20Å spherical stochastic boundary condition was used.

[c] Statistical uncertainities are ~±0.3 *kcal/mol* for all FEP calculation results.

[d] The sampling of the chemical space in MC steps are restricted only at end points.

[e] The sampling of the chemical space is allowed at both {λ_i^2=1 ,$\lambda_{k\neq i}^2$=0}and {λ_i^2=0.5, λ_j^2=0.5, $\lambda_{k\neq i,j}^2$=0}.

In spite of different simulation conditions, qualitatively good agreement is obtained between the GB model and the stochastic boundary explicit water model. With the respect to the GB model, the FEP, λ-dynamics, and CMC/MD methods gave very similar results. Furthermore, the λ-dynamics simulations with both coupling schemes, “Type 1” and “Type 2” (see **Eqns. 53** and **56**), were in good agreement. Although both definitions take different paths between the end points, they give exactly the same GB energy at all end points. This result means that the exact expression of the end points is key to getting the correct free energy differences. “Type 2“ required more CPU time (about 50 - 60%) as compared with “Type 1“ in λ-dynamics simulations, so that “Type 1“ may be more appropriate. To check the convergence of each simulation, the cumulative running average of the relative binding free energy differences are shown in **Figure 15**. The three stronger binders rapidly yield converged *ΔΔG* values in all λ-dynamics and CMC/MD simulations. Since more than 80 percent of the λ-dynamics trajectories were spent in unphysical intermediate states and the weaker binder (R=*Cl*) was sampled infrequently, its binding free energy converged more slowly than in the CMC/MD simulations, in which only end points are sampled. We found that the CMC/MD simulations with explicit solvent tended to get trapped in one state when different sized ligands were exposed to solvent and the intermediate states were not sampled (data not shown). In contrast, in this study, CMC/MD simulations, which sample only end points, yield converged *ΔΔG* values without getting trapped in a local minimum. Therefore, additional intermediate states in the MC procedure do not improve the final results. We speculate that the faster convergence of the CMC/MD simulations observed in this study may come from not only the similarity of the ligands but also introduction of the GB model. Furthermore, the faster convergence partially comes from the merit in the λ-dynamics and CMC/MD methods that some ligands can work as the intermediate states to connect dissimilar ligands.

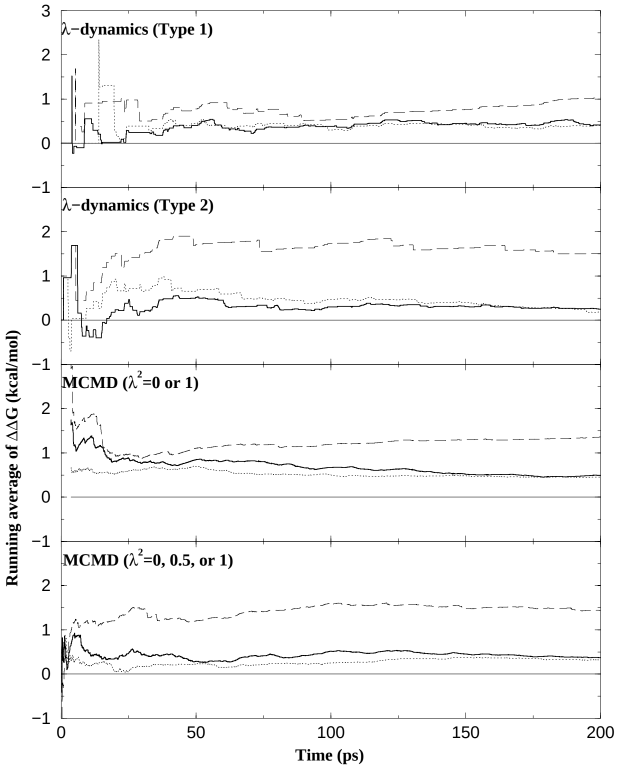


Figure 15. The cumulative running average of *ΔΔG* relative to *p*-H. Key: dotted line: *p*-NH2, solid line: *p*-CH3, dashed line: *p*-Cl.

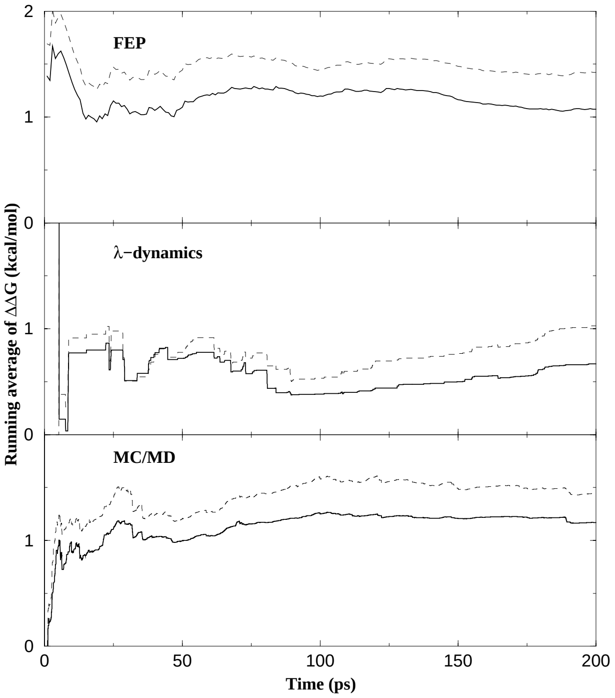


Figure 16. The cumulative running average of *ΔΔG* relative to *p*-H including the non-electrostatic solvation energy (V_{SA}). Only the results of R=Cl, which converge slowest, are shown for clarity. In FEP results, the X-axis represents the sampling time in each windows (λ=0.125, 0.5, and 0.875). The *ΔΔG* without V_{SA} and with V_{SA} are shown by dashed line and thick solid line, respectively.

4.3.2 The contribution from the non-electrostatic terms

We assume that the non-electrostatic terms can be linearly related to solvent-accessible surface area as shown in **Eqn. 62**. As a preliminary value for the empirical atomic solvation parameter (σ_i) shown in **Eqn. 59**, *7* cal/(molÅ^2) was chosen for all heavy atoms. We used the trajectories of the whole protein system with GB solvation to estimate the non-electrostatic terms. At each snapshot, the non-electrostatic energy (V_{SA}) was calculated using CHARMM package with 1.4Å radius probe atom. The effect of the non-electrostatic term is tabulated in **Table 4**.

Table 4. Results of relative free energy calculations including the non-electrostatic terms.[a]

R	ΔG(free)	ΔG(bound) (FEP)[b]	ΔΔG(bind) (FEP)	ΔΔG(bind) (λ–dynamics, *Type1*)	ΔΔG(bind) (MC/MD, *Type2*)[c]
H	0.0	0.0	0.0	0.0	0.0
NH_2	-17.97 (0.14)	-17.79(-0.05)	0.18(-0.19)	0.31(-0.12)	0.10(-0.22)
CH_3	-6.70 (0.21)	-6.30(0.09)	0.40(-0.12)	0.18(-0.23)	0.14(-0.23)
Cl	-1.95 (0.25)	-0.75(-0.10)	1.20(-0.35)	0.68(-0.35)	1.17(-0.27)

[a] **Free energy changes are in *kcal/mol* and relative to Benzamidine (R=H). The free energy difference ΔG_{SA} ($\Delta G_{SA} = \Delta G_{GB/SA} - \Delta G_{GB}$) are listed in parentheses. $\Delta G_{GB/SA}$ and ΔG_{GB} represent free energy differences with non-electrostatic terms and without them, respectively.**

[b] **Statistical uncertainities are within 0.3 kcal/mol for all FEP calculation results.**

[c] **The sampling of chemical space is allowed at both $\{\lambda_i^2=1, \lambda_{k\neq i}^2=0\}$and $\{\lambda_i^2=0.5, \lambda_j^2=0.5, \lambda_{k\neq i,j}^2=0\}$.**

The contributions from the non-electrostatic energy are relatively small in both half cycles due to the small difference in the size of the ligands in this system. Although the total non-electrostatic energy was considerably larger ($V_{SA} \approx 64$ kcal/mol), the umbrella sampling successfully estimated the effect of non-electrostatic terms and gave converged *ΔΔG* values for all ligands in the λ-dynamics and CMC/MD simulations, as shown in **Figure 16**. This occurred, in part, because $V_{SA,i}$ was nearly constant during the simulations. The FEP simulations also yielded converged results. These findings suggest that the effect of the non-electrostatic terms can be estimated from λ-dynamics and CMC/MD trajectories by using the umbrella sampling techniques as long as the protein remains in the native states. However, since the total non-electrostatic energy is large, it may bias the configurational sampling of the protein-ligand complex structure. Studies on other systems need to be done to clarify this issue. The application of the recently developed efficient method for the estimation of SASA would be another choice.[139-142]

4.3.3 15Å spherical boundary in the generalized Born solvent

When considering large inhomogeneous systems like the protein-ligand complexes here, one is sometimes required to limit the physical system to an interesting region around the ligand. To fully exploit the merits of the implicit solvent model (e.g. faster convergence with a smaller computational cost), development of efficient boundary methods is very important. For this purpose, a simple spherically truncated GB model, in which atoms outside a spherical region of interest were removed, was prepared as follows. The protein-ligands complex structure was centered at the carbon atom of the ligand connected to the amidine group. Any residue with all atom distances greater than 15Å from this central point was removed, as shown in **Figure 17**. The heavy atoms outside of a 10Å sphere were restrained by harmonic restraints with a force constant of 4 kcal/mol/Å^2, in order to keep them near the X-ray structure. The 15Å spherical protein system contained 1478 protein atoms, and 29 ligand atoms. To compare with the 15Å spherical GB model, an explicit water model using a 15Å stochastic boundary potential was also prepared according to the previous protocols.[39] The system was partitioned using a 10Å reaction zone, with a 5Å buffer region. Non-hydrogen buffer region atoms were restrained by harmonic restraints with a force constant of 4 kcal/mol/Å^2. All simulations (i.e. FEP, λ-dynamics and CMC/MD) used the same equilibrium and production protocols described above.

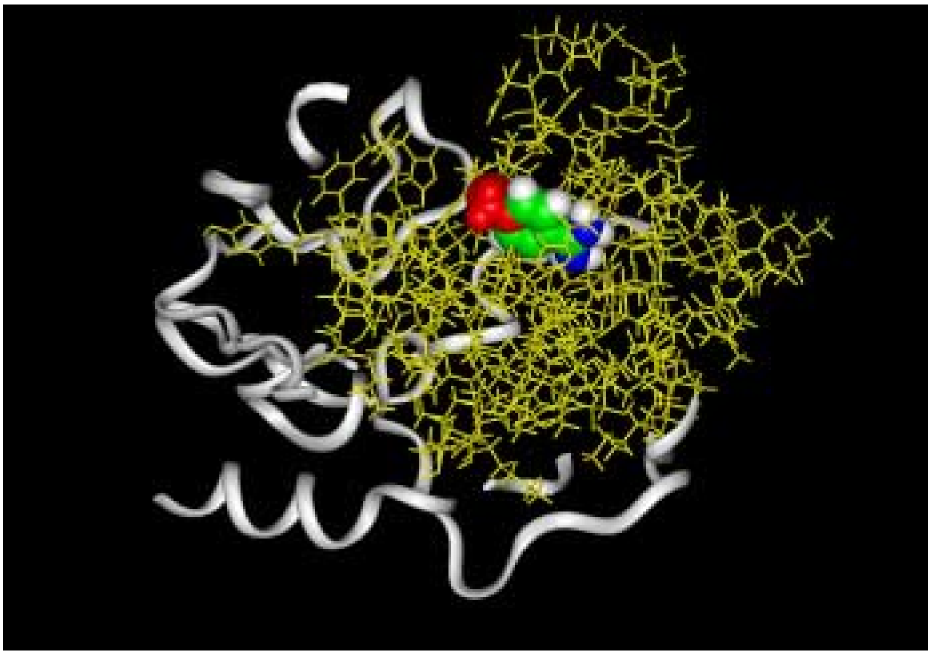


Figure 17. Four benzamidine derivatives in trypsin. The residues included in 15 Å spherical boundary model are shown by red. The removed residues in this boundary model are shown by cyan. The variant groups of the ligands are shown by yellow. Carbon atoms, Nitrogen atoms, and hydrogen atoms represented by single topology are shown by green, blue, and white, respectively.

Table 5. Results of relative binding free energy calculations with 15Å spherically truncated models.[a]

R	FEP[c] GB *type1* whole protein	FEP[c] GB *type1*	λ–dynamics GB *type1*	MC/MD[d] GB *type2*	FEP[c, e] GB *type1*	λ–dynamics[e] GB *type1*	MC/MD[d, e] GB *type2*	FEP[c] EW[b]	λ–dynamics EW [b]
H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NH_2	0.37	0.43	0.40	0.48	0.42	0.22	0.32	1.59	1.55
CH_3	0.52	0.44	0.26	0.28	0.39	0.41	0.32	3.46	3.56
Cl	1.55	1.40	1.30	1.13	1.60	1.57	1.26	4.70	- [f]

[a] **Free energy changes are in kcal/mol and relative to Benzamidine (R=H).**

[b] **The explicit water molecules with 15Å spherical stochastic boundary condition are used.**

[c] **Statistical uncertainities are ~±0.2 kcal/mol for all FEP calculation results.**

[d] **Only end points {λ_i^2=1,$\lambda_{j\neq i}^2$=0} are included as the sampling of the λ space in MC steps**

[e] **Born radius is estimated roughly including the deleted atoms outside the 15Å sphere.**

[f] **R=Cl does not take the dominant states during the production run.**

The values of *ΔΔG* of the 15Å spherical GB model calculated by FEP, λ-dynamics and CMC/MD are tabulated in **Table 5**. The results of the 15Å spherical GB model are in good agreement with those for the whole protein GB model. The computational time required for the 15Å spherical GB model was about three times smaller than the time for the whole protein GB model. The most crucial differences between the whole protein and the spherically truncated model are the elimination of interactions between the ligand and the environment atoms which are removed and the underestimation of the effective Born radii for atoms near the 15Å sphere boundary. In this study, shorter cutoff distances minimize direct interactions between the ligand and the removed atoms. To assess the effect of the underestimation of the effective Born radii near boundary, we also carried out FEP, λ-dynamics and CMC/MD simulations in which Born radii were calculated including the contributions from the atoms outside of the 15 Å spherical region. These contributions from the removed atoms are assumed to be constant and evaluated only once using the initial structure of whole protein model. These constant offsets are always included in the estimations of the effective Born radii in later simulations. The results of FEP, λ-dynamics and CMC/MD simulations that include contributions from the removed atoms do not appreciably differ from those that neglect these contributions, therefore, the underestimation of the effective Born radii was little influence on the free energy differences. As shown in **Figure 18**, the GB energy is almost cancelled with the Coulomb interaction energy *in vacuo* and the effect of the Born radius quickly decreases when the atoms are separated by more than 15 Å. Furthermore, most of these differences are cancelled among the ligands if they are similar. A larger sphere may be necessary for highly polarized systems or those in which the electrostatic potential is very different between the ligands.[143] Therefore, when cut-off distance and restraints on the native structure are introduced, this spherically truncated model may save time without significant loss of accuracy. Further improvement may be achieved by the addition of grid based inclusion of the removed atoms for both the estimation of the Born radius and electrostatic potential. A similar idea has already been used in docking studies.[81] However, an interpolation technique will be required to yield continuous energy and derivatives used in MD simulations. As a matter of course, if the low frequency motions expanding to the whole protein are important to express the binding free energy differences, a spherically truncated model with GB may not yield binding free energy differences correctly, as would be true with other non-periodic boundary models.[16,134]

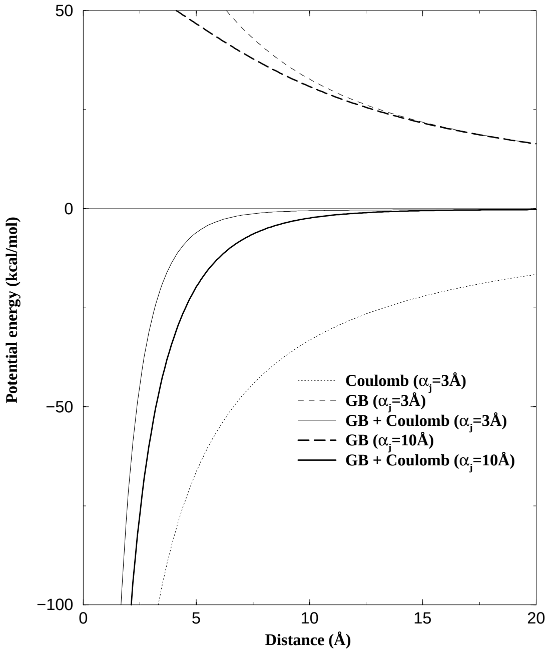


Figure 18. The electrostatic solvation energy (GB energy + Coulomb potential in vacuo) between two atoms who have the opposite unit charge. The X-axis represents the distance between them. The effective Born radius (α_j) of one atom is changed, whereas the other is fixed at 3Å.

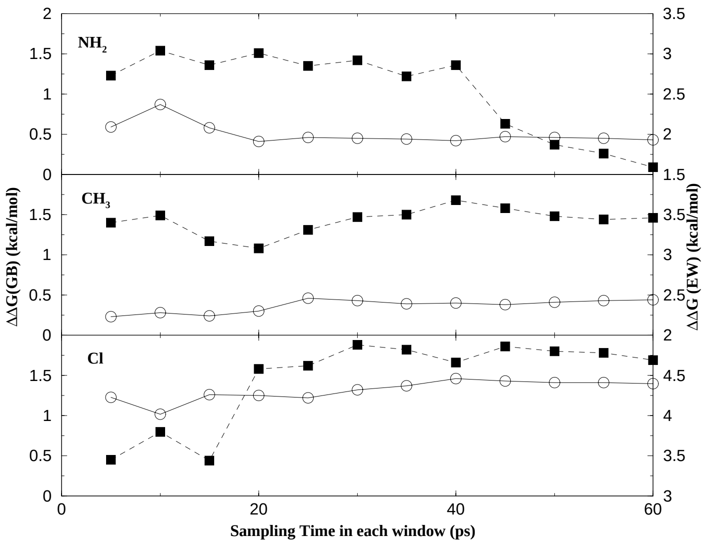


Figure 19. The trajectories of *ΔΔG* calculated by FEP methods using 15Å spherical model. *p*-H was chosen as reference. The results using GB model and explicit water model are shown by solid line and dashed line, respectively.

FEP and λ-dynamics simulations were also carried out with the explicit water model using a 15Å spherical stochastic boundary. Qualitative agreement was obtained between the GB model and the explicit water model as shown in **Table 5**. The quantitative agreement between the FEP and λ-dynamics methods using explicit water again demonstrates the validity of the λ-dynamics method. The relative binding free energy differences calculated by the FEP method, as a function of sampling time in each window, are shown in **Figure 19**. The GB model converges faster than the explicit water model. The faster convergence with implicit solvent is an important advantage over the explicit water model. This advantage may come from the fact that the averaged solvent-solute interaction energy is calculated from a single solute structure in the GB model. In contrast, the explicit water model requires many configurations of the solvent for one solute configuration. Furthermore, van der Waals clashes between the unselected ligands and the mobile solvent, which can lead to the instabilities in integrating the equations of motion or trapping of one ligand in multiple ligand free energy methods, can be avoided by using the GB model.

4.4 Conclusions

In this chapter, the GB model has been extended to describe the hybrid intermediate states associated with free energy simulation methods and incorporated into these methods. Promising observations for the ranking of ligand binding were obtained from the combination of continuum solvent models using GB with free energy simulation methods (FEP, λ-dynamics and CMC/MD). The GB model gave good agreement with explicit water models. Furthermore, the CMC/MD method was assessed and demonstrated to yield good agreement with FEP and λ-dynamics methods. In our study, the non-electrostatic solvation energy (V_{SA}) varied only small amount so that the umbrella sampling techniques could be used for the incorporation of V_{SA}, instead of calculating V_{SA} and its first derivative at every MD step. The simple spherically truncated model also showed good agreement with the whole protein model. Such a boundary model will help to accelerate the qualitative computational screening of promising compounds.

The use of a GB model, with advantages such as faster convergence and fewer degrees of freedom, save a significant amount of computational time. The λ-dynamics and CMC/MD methods are also known to screen for better ligands from putative candidates with much smaller time than the conventional free energy simulation methods. The combination of the GB model with λ-dynamics or CMC/MD has a significant potential in the application for drug lead optimization. These combinations may fill the gap between the empirical methods using a single minimized complex structure and the theoretically rigorous methods like FEP or thermodynamic integration.

Chapter 5

β-Cyclodextrin - Benzene Derivatives

Based on

Komath V. Damodaran, Shinichi Banba, and Charles L. Brooks, III,

"Application of multiple topology λ-dynamics to a host-guest system: β-cyclodextrin with substituted benzenes,"

J. Phys. Chem. B, **105(38)**, 9316-9322 (2001).

Shinichi Banba, Komath V. Damodaran, and Charles L. Brooks, III,

"Free energy simulations with generalized Born solvent II: applications to solvent exposed macrocycles,"

J. Phys. Chem. B, submitted.

5. 1 Introduction

We have been investigating new ways to render free energy based screening methods still more efficient and as part of this effort we have attempted to carry out the λ-dynamics simulations in an implicit solvent environment using the GB model.[84] In the previous chapter, I demonstrated that the incorporation of the GB model into the λ-dynamics method could be successfully applied to rank the four benzene derivatives bound to trypsin.[84] For small molecules, the GB model has been shown to reproduce solvation energies while utilizing less CPU time when compared to the PB model. However, the current analytical GB models are known to underestimate the effective Born radii of deeply buried atoms and this has limited its application. In one attempt to address these issues, Onufriev et al. introduced a single parameter to account for the nonzero size of solvent molecules buried in the interior region as the system increases in size.[65,144] This problem in the analytic GB approximation indicates that the GB parameters should be chosen carefully when large molecules are studied. Moreover, the GB energy involves only the electrostatic polarization even though the non-electrostatic terms (e.g., hydrophobic contribution) can play a vital role. Thus, before applying the λ-dynamics/GB method, we investigated how the non-electrostatic terms and the set of the GB parameters influenced binding free energy results using β-cyclodextrin (β-CD)-toluene complex system.

To increase the overlap at the end points of chemical free energy perturbation calculations, FEP and λ-dynamics require the use of the intermediate states. The introduction of these intermediates can be represented using different strategies to mimic the molecular topologies, such as single or dual topology models. In this work, “multiple topology” means that each guest is represented by the independent molecular topology or structure and “hybrid topology” means that the invariable guest atoms (non-varying framework) are represented by a single molecular fragment and the variant atoms of the guests are represented by the separate topologies connected to this single fragment, like a Hydra with multiple chemical “heads” pendant to a single molecular framework. The multiple topology representation allows the comparison among molecules with quite different molecular structures. However, in investigations of relative binding affinity among a family of similar compounds, the hybrid topology representation seems to be the more promising one. The use of this framework requires the introduction of λ-dependent partial charges, as described in what follows.

In this chapter, firstly, we further validate the techniques of λ-dynamics using the multiple topology representation for the ligands in the presence of the restraining potential by applying the method to a β-CD-benzene derivatives system with explicit solvent in a periodic environment. We have also examined the ligand dynamics under restraining potentials of different strengths and biasing conditions. Secondary, the GB solvent model was investigated in β-CD-toluene complex system. Finally, we have applied hybrid topology λ-dynamics with the GB solvation model to investigate the relative binding free energies of seven mono substituted benzene derivatives in β-CD. To validate the hybrid topology λ-dynamics/GB method using λ-dependent partial charge model, its results are compared with multiple topology FEP simulations using explicit solvent, highlighting the merits and weaknesses of both topology representations.

5.2 Computational details

5.2.1 System under study

The system chosen for this investigation is a series of monosubstituted benzenes binding to a host molecule, namely β-cyclodextrin. Cyclodextrins (CDs) are cyclic oligosaccharides consisting of glucopyranose units linked using α-1-4 glycosidic bonds. The most widely investigated variants have six (α), seven (β) and eight (γ) sugar units.[145-147] The shape of β-CD is described as truncated cone with the primary hydroxyl (CH_2OH) groups occupying the narrower rim[148]. The secondary hydroxyl (OH) groups at the 2' and 3' positions form hydrogen bonds with adjacent sugar units, adding to the rigidity of CDs.[148] Particularly, they show only limited rotation around the bonds linking the sugar units.[149,150] CDs have been of great interest because of their propensity to form inclusion compounds with small, hydrophobic molecules, a property that has been widely made use of as a facilitator of chemical reactions[151] and as a drug delivery agent.[152] They are also simple models for studying host-guest chemistry. The interior of these molecules are hydrophobic and the host-guest interactions are primarily van der Waals in nature. Early molecular modeling studies have suggested CDs to be flexible.[153] However, these calculations did not include any solvent environment, which may have contributed to the observed flexibility. Dynamic Monte Carlo simulations of complexes have shown that even inclusion of solvent effects using implicit models results in a reduction of conformational space.[154] MD simulations of α- and β- CDs have been carried out in both the crystalline and explicit solvent environments by Koehler et. al.[155-157] These studies not only reproduced the structural properties from neutron diffraction studies, but the dynamics of the hydrogen bonds between adjacent sugar units crystalline model agreed excellently with the experimental results as well. Kohler et al. also showed that the solution structure of deviates from the crystalline structure by ≈1Å and is somewhat more mobile.[155]

5.2.2 Dynamics simulations with the explicit water

The benzene derivatives used as guests with the explicit water model have the following substituents: R=*H, -CH_3, -CH_2Cl, -Br, -NO_2, -OCH_3, -Cl,* and *-F*. Association constants for these and other related ligands with β-CD have been determined and analyzed for quantitative structure-property relationships by using multiple regression methods by Liu *et al*.,[158] We used the structure of the β-CD-benzyl alcohol complex taken from the 'DEBGOG' entry of the Cambridge Crystallographic Database as a template to build the initial structures, using the 3-D molecular builder facility in the program QUANTA. We adopted the structural model as suggested in Liu et al., [159] as our starting structure, namely the monosubstituted benzenes reside in the cavity such that the long axis of the guest is perpendicular to the plane of the β-CD host and the hydrophobic substituent of the guest resides at the narrow rim of the truncated cone (**Figure 20**). The sugar forcefield (par_all22.sugar) developed by Guyan Liang and John Brady (Cornell University 09/08/1995) was used to model β-CD. Param-22 force field parameters and partial charges obtained using electrostatic potential (ESP) fitting[160] on ab initio optimized structures at the Hartree-Fock/6-31G* level were used for the guest molecules. All bonds containing hydrogen atoms were constrained to standard values using SHAKE.[121] The temperature of the system was maintained near 300K by coupling the non-hydrogen atoms to a Langevin heatbath using frictional coefficient of 50 ps^{-1}. The λ-variables were also coupled to a Langevin heatbath using frictional coefficient of 5 ps^{-1} to keep the system near 300K in all λ-dynamics simulations. The time step used in all simulations was 1.0 fs. The masses of the fictitious λ degrees of freedom were chosen to 5 amu•$Å^2$. All calculations were done using the CHARMM molecular dynamics package.[119]

The host-guest complex was solvated in a pre-equilibrated water box containing 1000 TIP3P water molecules[120] and the overlapping water molecules within 2.8Å were removed. The final configuration consisted of the host, the guest molecules and 923 water molecules. MD trajectories were generated for the solvated empty β-CD as well as for its complexes with toluene (R=CH_3), nitrobenzene (R=NO_2), methoxybenzene (R=OCH_3) and fluorobenzene (R=F). The solvation free energies of the ligands were calculated by solving the PB equations on energy minimized structures using a grid size of 0.1Å. The relative free energies (ΔGs) of the guests in the complexed form were calculated by FEP simulations. Mutations were carried out over 9 windows using toluene (R=CH_3) as the reference. Both the reactant and the product were explicitly represented (dual topology model). In the case of the substituents CH_2Cl, *Br*, NO_2 and OCH_3, mutations were carried out from the substituent to CH_3, while for *H*, *F*, and *CHO*, the mutations were carried out from CH_3 to the substituent. An equilibration phase of 15ps and sampling phase of 45ps was used in each window. λ–dynamics simulations were carried out using 8 guest molecules in the β-CD cavity at three different scaling parameters (α) for the restraining potential – 0.10, 0.30, 0.50. In the unbiased run, all the F_i values were set to the respective solvation free energies. In the run with α=0.1, some ligands did not sample the dominant state (λ^2 > 0.8) under this condition. The relative binding free energies from the unbiased simulations were used as additional biases to enhance the sampling. Typically, the λ-dynamics simulation time was 450ps of which the initial 50ps part of the trajectory was discarded from the estimation of free energies.

5.2.3 Dynamics simulations with the generalized Born model

The system studied with the GB model is benzene and six of its mono-substituted derivatives, namely, toluene (R=CH_3), aniline (R=NH_2), phenol (R=*OH*), nitrobenzene (R=NO_2), bromobenzene (R=*Br*), and fluorobenzene (R=*F*) with β-CD. The carbon atoms of the phenyl ring and the five hydrogen atoms directly attached to those carbon atoms are colo atoms and expressed by the single topology. The angle θ, which represents the binding orientation of the guest in the host, is defined between the long axis of the guest and the molecular axis of the host as shown in **Figure 20**.

In this study, we used the GB parameters specifically optimized for the CHARMM force field, as described by Dominy and Brooks.[83] The van der Waals scaling parameter for single amino acids was used for the unbound state,[83] while that for bound state was re-fitted to β-cyclodextrin. The details of the GB parameter fitting for β-CD system is described in **Appendix**. We set all hydrogen radii to 1.5Å for the estimation of the effective Born radii.[84] In all simulations including the GB energy, the Born radii and corresponding forces were updated at every MD time step, thereby ensuring the correct relationship between the energy function and its derivative. The free energy of the guests in the unbound state was calculated from one minimized structure taking advantage of the rigid nature of the guests. We use the same simulation protocols in the GB model as those in the explicit water model, except for no cutoff used in the GB model.

In FEP calculations of the bound state using the GB model, six transformations – *Br* to *H*, NH_2 CH_3, NO_2, *OH*, and *F* – were considered. A 50ps equilibration period was followed by a 200ps production run for three λ values (0.125, 0.5, 0.875). Double-wide sampling was used to span the entire λ space. In the λ-dynamics/GB simulations, a 30ps equilibration period was followed by 270ps production runs. The λ-variables were sampled at every 10fs. The total CPU time for 300ps λ-dynamics/GB simulation requires only less than 9 hours using SGI 250 MHz R10000 processor within an SGI Octane workstation, while the 300ps FEP simulation with the explicit water model requires more than 120 hours as a CPU time.

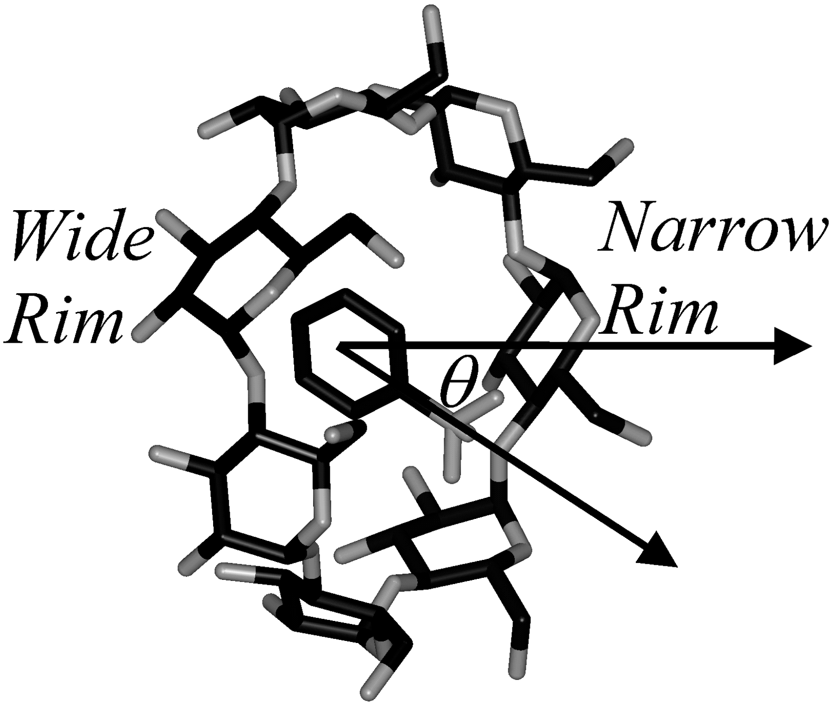


Figure 20. Schematic of β-cyclodextrin with monosubstituted benzene guests illustrating the orientation of the guests in the initial structure.

5. 3 Results and discussion

5.3.1 Structure and dynamics of the host

The β-CD host showed some degree of flexibility in all of the MD/FEP/λ-dynamics simulations. The maximum root-mean-square deviation (RMSD) for the host heavy atoms was ≈1.0Å, as evident from a typical RMSD plot shown in **Figure 21**. This compares well with the RMSD deviations observed between the solution and crystal structures of the six-membered varient (α-CD) by Koehler *et al.*,[155]. We also show, in **Figure 22**, the distribution of distances between bridging oxygen atoms (O_4) separated by 3 sugar units in the β-CD host from 9 different trajectories which includes MD/FEP/λ-dynamics runs. The distributions from all trajectories showed 9.8Å as the most probable value. The average of all distributions shows a full-width at half-maximum of 0.75Å. There was no constant deviation of any one distance from others, which would have indicated a distortion of the host.

The truncated cone shape of the host shows fluctuations both in the empty and complexed form. This is illustrated by the time dependence of the average distance between C_5 carbons separated by 3 sugar units, in comparison with the average distance between similarly situated C_3 carbons. The C_5 carbons are located on the narrow rim of the macrocycle while C_3 carbons are on the wider rims. Hence a smaller distance between the C_5 carbons is indicative of the truncated cone shape. The empty host appears to undergo oscillatory motions while retaining the truncated cone shape (**Figure 23-a**), whereas in the λ-dynamics trajectory with ligands present the truncated cone has relaxed somewhat **Figure 23-b**.

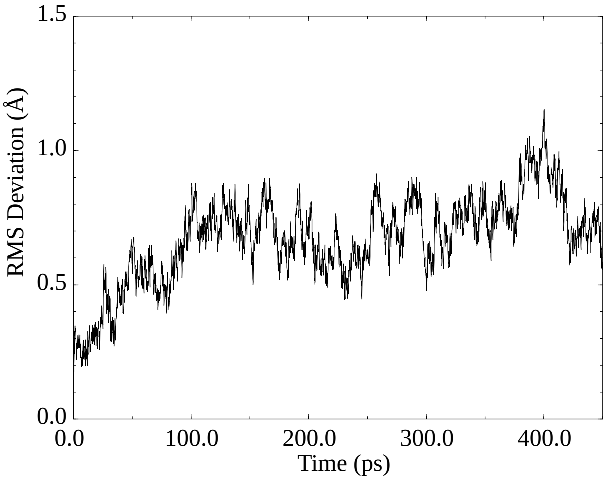


Figure 21. Time dependence of the root-mean-square deviation (RMSD) of the β-CD heavy atoms along a typical dynamics trajectory.

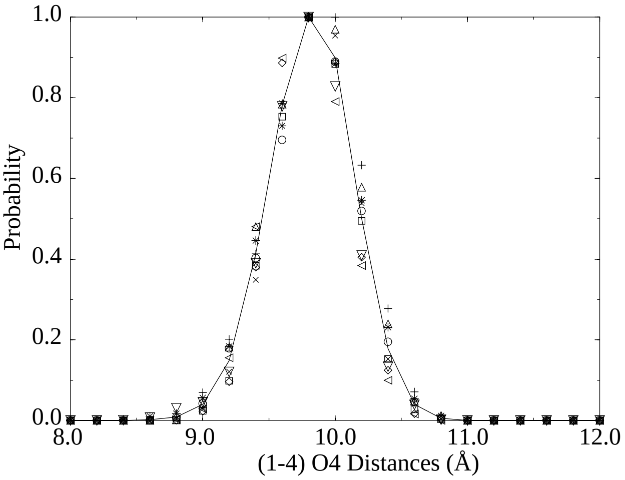


Figure 22. Distribution of the distances between bridging O_4 oxygens of the α-1-4 glycosidic bonds separated by three sugar units. The data points are from MD/FEP/λ-dynamics trajectories. The distribution from each trajectory has been normalized by the number of frames used. The solid line represents the average values.

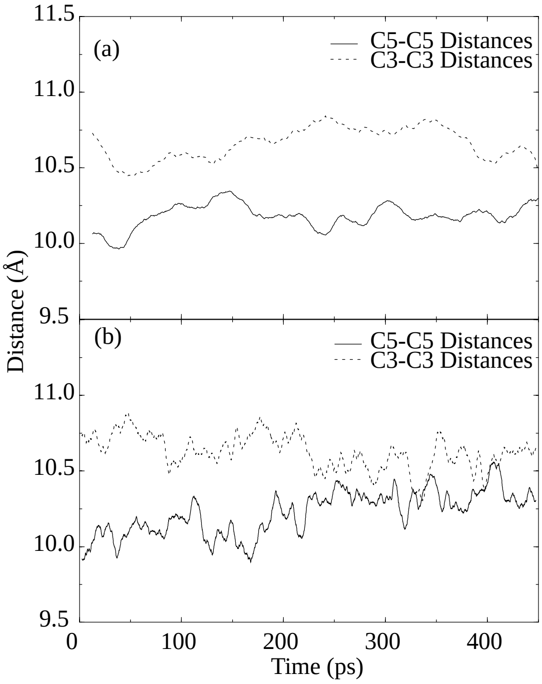


Figure 23. Distances between C_5 carbons and C_3 carbons separated by three sugar units from (a) MD trajectory of the solvated β-CD and (b) a typical λ-dynamics trajectory. Running averages of the instantaneous values are shown.

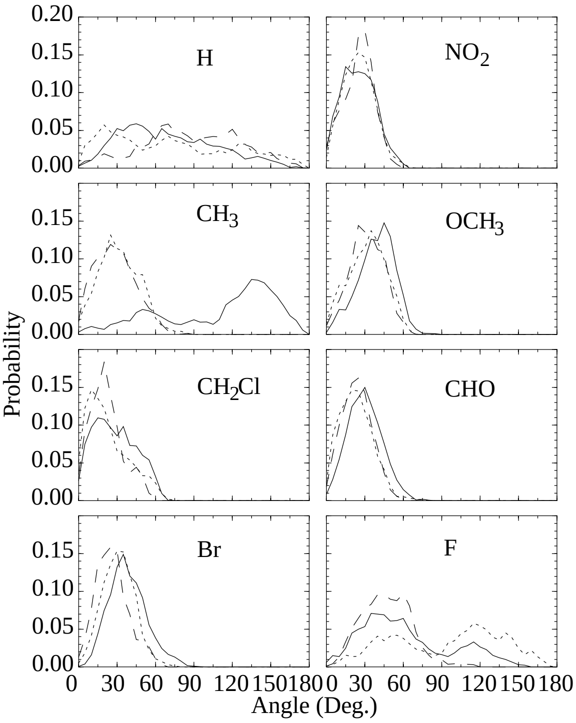


Figure 24. Distributions of the angle between the long molecular axis of the ligands and the molecular axis of the β-CD host from λ-dynamics trajectories for *α*=0.1 (solid lines), *α*=0.3 (dotted lines) and *α*=0.5 (dashed lines). Because Benzene is completely symmetric, it shows almost no orientational preference.

5.3.2 Ligand conformation and dynamics in MD simulations

The guest molecules with larger substituents remained in the same average conformation as the initial structure with the substituent near the narrow rim of the host. The mean angle of the long axis of the guest from the molecular axis of the host (see **Figure 20**) was ≈30°. Only fluorobenzene (R=*F*) underwent rotational motion during the simulations.

Overall, the RMS deviations in **Figure 21** and the distance data in **Figure 22** and **Figure 23** suggest that the environment provided by the host does not vary significantly during the course of the λ-dynamics simulations. Further, the solvent interacts with the ligands only from the regions beyond the two rims of the cavity. Thus using the instantaneous coordinates *X(t)* as an approximation to time invariant coordinates (see **Eqn. 45**) is a reasonable one.

5.3.3 Relative binding free energies with the explicit water

The relative binding free energies ($\Delta\Delta G$s) obtained from FEP simulations have been tabulated in **Table 6** with those calculated from experimental association constants (K_a) using the relation $\Delta\Delta G_{ij} = -(1/\beta)\ln(K_a^i/K_a^j)$. The experimental values were taken from Liu *et al.*.[159] The calculated values span a range much larger than the observed values. Further, the correlation between the calculated and observed data is not significantly high. Particularly, we have not been able to reproduce the relative trend for nitrobenzene (R=NO_2). The lack of significant correlation may be due to several factors including the force field, as in the case of all computer simulations. For example, we have used the solvation free energies derived from the PB equation, which includes only the electrostatic contributions. One may also obtain better correlation by refining the ligand force field parameters. We have not made any attempt in this direction since the objective of the present work is to investigate the efficiency of the λ-dynamics approach as a faster alternative for evaluating the relative binding free. We rely on the correlation between relative binding free energies obtained from FEP and λ-dynamics to demonstrate this.

The detailed values of $\Delta\Delta G$s obtained from λ-dynamics simulations are also listed in **Table 6**. Data from multiple trajectories with identical scaling parameter α, but under different biasing conditions, were combined using the WHAM equation (**Eqn. 48**). There is very good correlation between FEP and λ-dynamics data with correlation coefficients equal or better than 0.9. In the case of α=0.5, free energies from a 450ps long λ-dynamics run with no additional biases (i.e., F_i have been set to the solvation free energies of the ligands) shows 95% correlation with the FEP data.

Table 6. Relative binding free energies calculated from FEP and λ-dynamics simulations.[a]

substituent (R)	Exp.[b]	FEP[c]	λ-dynamics				
			α=0.3		α=0.5		α=0.1[c]
			no bias	biased[b]	no bias	biased[b]	biased[b]
-*H*	0.14	2.75	3.79	3.07	3.72	3.43	3.21
-CH_3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-CH_2Cl	-0.16	-0.52	0.46	0.10	0.08	0.40	0.75
-*Br*	-0.23	0.43	0.43	0.31	0.54	0.63	1.26
-NO_2	-0.16	4.80	3.60	3.78	4.70	4.41	3.60
-OCH_3	0.02	0.50	-0.03	0.08	-0.46	-0.03	0.46
-*CHO*	0.21	3.02	4.14	3.71	3.94	3.80	3.84
-*F*	0.51	1.80	2.00	1.67	1.50	1.63	2.0

[a] All values are in kcal/mol and relative to R=CH_3.

[b] The values “Biased” were calculated from No Bias and Biased trajectories using WHAM.

[c] For α=0.1, “No Bias” trajectory was not considered, since some ligands did not sample the dominant state.

Additional simulations were carried out using the calculated $\Delta\Delta G$s from the unbiased run as additional biases to the F_i. However, combining these data improves the correlation only marginally. In the case of α=0.3, $\Delta\Delta G$s from the unbiased λ-dynamics have a lower correlation coefficient (0.89) with the FEP data than in α=0.5 case, because the stronger restraining potentials confines the ligands more to their binding orientations. This is noticeable in the distribution of angular orientation of the ligands shown in **Figure 24**. Ligands such as (R=*F*), (R=CH_2Cl) and (R=NO_2) have slightly broader orientational distribution in the α=0.3 trajectory than in α=0.5 trajectory. However, using the $\Delta\Delta G$s from the unbiased run as additional biases in the subsequent λ-dynamics runs improves the correlation coefficient to 0.95. In the case of α=0.1, weaker guests such as (R=NO_2) and (R=*CHO*) did not sample the dominant states during the entire simulation. So *ΔΔG* from the unbiased simulation with α=0.3 was used as biasing potentials, which helped these ligands also dominate. The *ΔΔG*s calculated from the biased simulations still has only 91% correlation, lower that with α=0.3 and α=0.5. We note that the particular value of α used in the biasing potential serves to influence sampling efficiency and not (in principle) the final free energy estimates.

5.3.4 Ligand dynamics

An attractive feature of λ-dynamics is its ability to explore alternate binding conformations for the ligands. This is due to the fact that when λ≈0 for a particular ligand, the only interaction that it is subjected to is the restraining potential. By controlling the scaling parameter for the restraining potential we can control the extent of conformational space to be sampled. The validity of alternative binding conformations can be assessed from the probabilities that such conformations sample the dominant λ-states. In λ-dynamics simulations, increasing the scaling parameter α for the restraining potential restricts conformational space sampled by the ligands. The effect of the scaling parameter α is illustrated in **Figure 25** where the variation of the orientation of the ligand (R=*F*) and the coupling coefficient (λ^2) along the MD and λ-dynamics trajectories have been shown. In our MD simulations of β-CD/ligand complexes, the ligands except (R=*F*) showed a distribution of orientations centered around about 30° from the molecular axis of the host. This corresponds to the substituents confined at the narrow rim of the host. However, (R=*F*) starting from this initial orientation quickly adopted an orientation centered about 130° (i.e., fluorine near the wider rim), as shown in **Figure 25-a.** In the λ-dynamics trajectories, when the restraining potential was weak (α=0.1, 0.3), starting from the same initial orientation (R=F), explored both orientations in the dominant state (**Figure 25-b** and **Figure 25-c**).

However, in the trajectories with α=0.50, the restraining forces were stronger and ligand remained in the vicinity of the initial orientation, as shown in **Figure 25-d**.

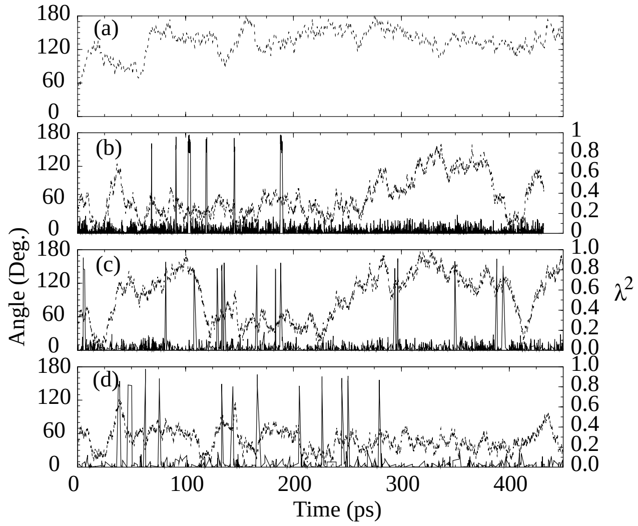


Figure 25. Variation of the orientation in the host cavity (dotted lines) and the coupling parameter λ^2 (solid lines) of the ligand (R=*F*) along the (a) MD and unbiased λ-dynamics trajectories with (b) α=0.1, (c) α=0.3 and (d) α=0.5.

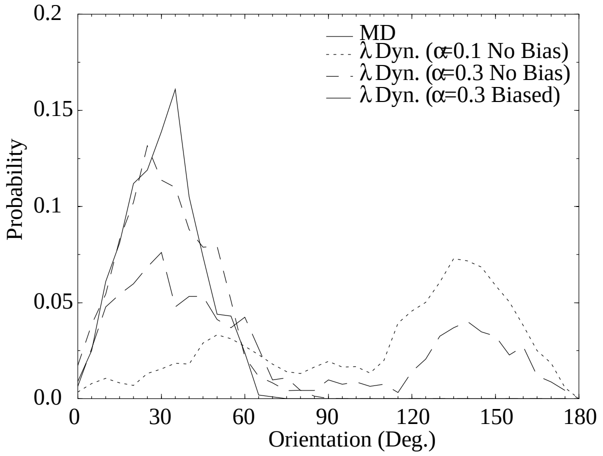


Figure 26. Distributions of the orientation in the host cavity of the ligand (R=*CH_3*) from MD and λ-dynamics trajectories.

Another interesting case of enhanced sampling is provided by the ligand toluene (R=*CH_3*), whose orientational distributions are shown in **Figure 26**. In the unbiased run with α=0.1, this ligand explored the inverted orientation (with the methyl group near the wider rim), which also sampled the dominant state. Increasing α to 0.30 restricted this ligand from sampling this orientation. However, when the calculated *ΔΔG*s from the unbiased run were added as bias potentials, the poor binding ligands became more competitive. As a result, the fraction of the time the strong binding ligands such as (R=*CH_3*) and (R=*OCH_3*) spend in the dominant state was reduced. This enabled (R=*CH_3*) to sample the alternative orientation with α=0.3 also. The alternative orientation was not sampled when α was increased to 0.5. To evaluate the relative free energies of the two orientations, a 200ps λ-dynamics trajectory was generated with two toluenes in the β-CD host, one in the "regular" and the other in the alternate orientation. The initial conformation was built such that the benzene rings of the two ligands overlapped. Further, harmonic restraints were applied between the overlapping ring carbon atoms of the two ligands. This restricted the sampling space of the ligands to the vicinity of their respective initial orientations. The calculated free energy difference between the two orientations from the λ-dynamics trajectory was 0.20 kcal/mol, in favor of the orientation with the methyl group near the narrow rim. A FEP run on the same model using 5 windows (mutation time = 300ps) yielded a relative free energy of 0.18, but in favor of the inverted orientation (i.e. methyl group near the wider rim). Judging from the small value of the relative free energies, we conclude that there may not be any preference for one orientation over the other, although the ligand may not undergo free rotations inside the cavity due to the energy barrier associated with the "flat" orientation (i.e., ligand in the plane of the host).

5.3.5 Validation of the generalized Born solvent model

The total solvation free energy is composed of the GB electrostatic part and a non-electrostatic part. Traditionally, the latter is assumed to be linearly related to the solvent-accessible surface area as shown in **Eqn. 59**. [65,67] In order to assess the relative importance of the surface area term in the implicit solvent model, 500ps conventional MD simulations of the β-CD-toluene system (100ps equilibration and 400ps sampling) were carried out using the GB energy with the surface area term (labeled GB/SA) and without (labeled GB). The van der Waals scaling parameter λ_α (see **Enq. 51**) was modified to fit for the various sized molecules as shown in **Appendix**. In β-CD system, λ_α=0.74 is chosen. Moreover, the same simulations were also carried out with λ_α=0.705 which was optimized for a protein database. The system was coupled to a 300

K heatbath. In this study, 7 cal/(molÅ^2) was chosen for the empirical atomic solvation parameter (σ_i) of all heavy atoms.[76] (see **Eqn. 59**) The averaged root mean square deviations for the host heavy atoms were about 0.9 Å in both the GB and GB/SA models, which is slightly larger than the value of 0.8 Å obtained with explicit water.[161] We show, in **Figure 27**, the distribution of distances between bridging oxygen atoms separated by three sugar units in the β-CD host to illustrate the sampling of host configurations. Although the implicit solvent models (GB and GB/SA) produce broader distributions, both implicit and explicit models yielded 9.8Å as the most probable value. As the GB and GB/SA models gave similar distributions, we conclude that the surface area term had little effect on the sampling of host configurations. The GB model with λ_α=0.705 yielded a distribution of distances that was centered at larger values. The distributions of the angle θ, indicating the binding orientation of toluene, are also shown in **Figure 28**.

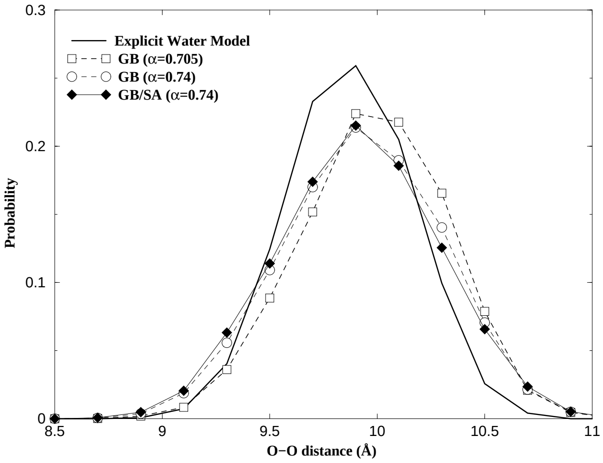


Figure 27. Distribution of the distances between bridging oxygen atoms of the α-1-4 glycosidic bonds separated by three sugar units obtained from the conventional MD trajectories of β-CD-toluene system for explicit water model (solid line), GB model (dotted line with circle plots), and GB/SA model (long dashed line with square plots).

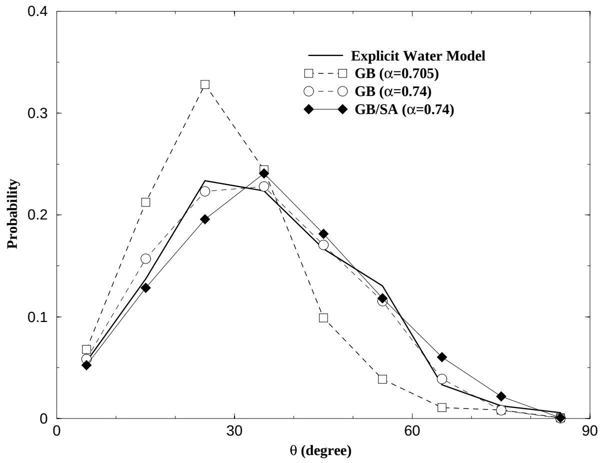


Figure 28. Distributions of the angle θ from the conventional MD trajectories of β-CD-toluene system for explicit water model (solid line), GB model (dotted line with circle plots) and GB/SA model (long dashed line with square plots).

Both explicit and implicit water models adopted θ=20~30° as the most probable value and produced similar distributions of distance when λ_α=0.74. However favored binding orientations using the GB model are more parallel to the molecular axis of the host when λ_α=0.705 is used. The GB and GB/SA models show good agreement in the sampling of guest orientations as well as host configurations. We conclude that the non-electrostatic contribution is negligible in its influence on the configurational ensemble, since it is mostly constant throughout the simulation (i.e. the largest differences of the surface area part and the GB part from 400 ps trajectories are 0.6 and 14.5 kcal/mol, respectively). Therefore, in this application, the non-electrostatic contribution was ignored in the following simulations with the GB model. The value of λ_α (0.74), chosen by considering the number of atoms, reproduced the sampling of both guest orientations and host configurations, however, the previously optimized value (λ_α=0.705) for a protein database generated larger deviations from those of the explicit water model.

5.3.6 Relative free energy differences using the generalized Born model

The relative binding free energies ($\Delta\Delta G$s) obtained from the hybrid topology λ-dynamics simulations using the GB model are displayed in **Figure 29** and compared to those from the hybrid topology FEP simulations using the GB model as well as the multiple topology FEP simulations using an explicit TIP3P water model. The detailed values of these free energy changes are also tabulated in **Table 7**. Since the restraining potentials for the unselected ligands are not included in the hybrid topology λ-dynamics/GB simulations, $\Delta\Delta G$s are calculated from **Eqn. 16**.

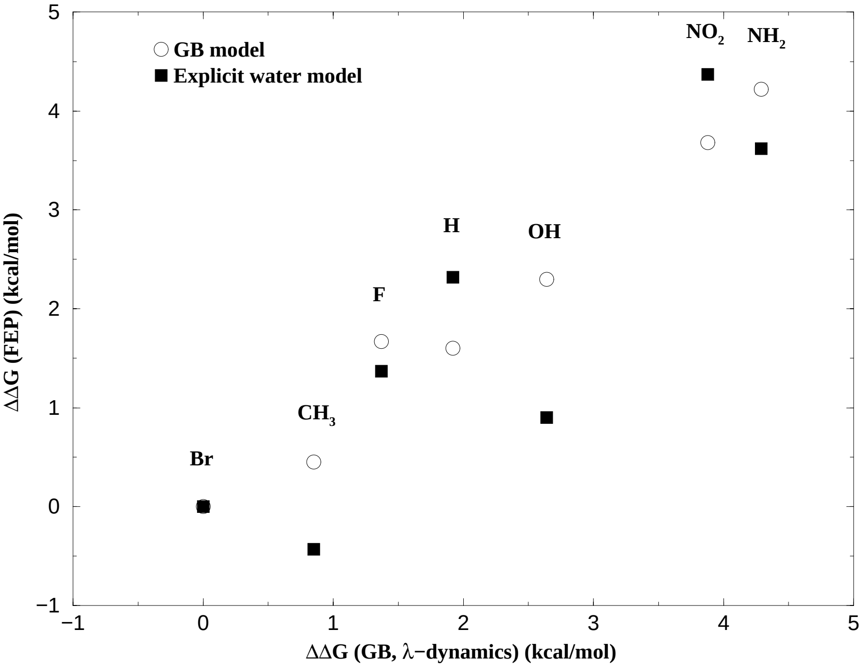


Figure 29. Correlation plot of the calculated relative free energies of binding from λ-dynamics and FEP simulations. The correlation coefficients between *ΔΔG* (λ-dynamics/GB) and *ΔΔG* (FEP/GB) and between *ΔΔG* (λ-dynamics/GB) and *ΔΔG* (FEP/Explicit water) are 0.99, 0.89, respectively.

Table 7. Results of relative binding free energy calculations.[a]

R	ΔΔG (Exp.)[b]	ΔG (EW) (FEP)[b]	ΔG(GB) (FEP)[c]	ΔΔG(GB) (λ–dynamics)
Br	0.0	0.0	0.0	0.0
H	0.38	2.32	1.60	1.92
CH_3	0.23	-0.43	0.45	0.85
NH_2	1.10	3.62	4.22	4.29
NO_2	0.08	4.37	3.68	3.88
F	0.75	1.37	1.67	1.37
OH	0.72	0.90	2.30	2.64

[a] Free energy changes are in kcal/mol and relative to Bromobenzene (R=Br).

[b] Statistical uncertaininities are ~±0.2 kcal/mol for all FEP calculation results.

These plots show very good correlation. In spite of different simulation conditions (cutoff distance and topology model), good agreement is obtained between the GB model and the explicit water model. We can observe in **Figure 29** that the GB solvent model tends to overestimate the $\Delta\Delta G$ values of the guests that include the hydrogen atoms in the variable parts (i.e. R=*OH*, NH_2, CH_3) when compared with those from the explicit water calculations. This may partially come from an inconsistent estimation of the effective Born radii between the heavy atoms and the hydrogen atoms (see **Appendix**). For example, the Born radii of the hydrogen atoms belonging to the variable parts remained unchanged as the guests bind to the host (i.e. 1.9 - 2.1Å for R=CH_3, ~1.5Å for R=NH_2, ~1.6Å for R=*OH*). While those of the heavy atoms of the variable parts increase when the guests bind to the host such as 2.4 - 3.2Å for R=CH_3, 2.1 - 2.8Å for R=NH_2, 1.9 - 2.6Å for R=*OH*. The smaller Born radii of the positively charged hydrogen atoms in the interior portion of the host may decrease the favorable electrostatic interaction between the guest and the ligands. These results indicate that a new methodology to estimate the consistent effective Born radii for both hydrogen and heavy atoms is very important for both small and large molecules in order to apply the GB model to ligand screening problems.[162] Results from FEP and λ-dynamics runs using the GB solvent model were in very good agreement, which validated the λ-dependent partial charge model. The partial charge of the guest carbon atom attached to the variable groups varies most among all *colo* atoms, changing from -0.1417 (R=*Br*) to 0.5523 (R=*OH*). The trajectory of its partial charge is shown in **Figure 30** together with the λ^2 value of bromobenzene. Its partial charge fluctuates during the λ-dynamics simulation with the λ-dependent partial charge model. When bromobenzene occupied the dominant states, its partial charge reaches -0.14.

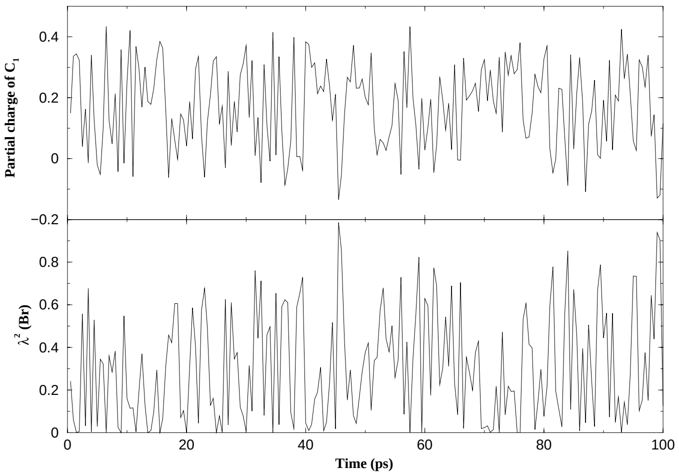


Figure 30. The trajectory of the partial charge of the guest carbon atom connected to the variable groups in atomic unit. The trajectory of λ^2 value of bromobenzene is also shown.

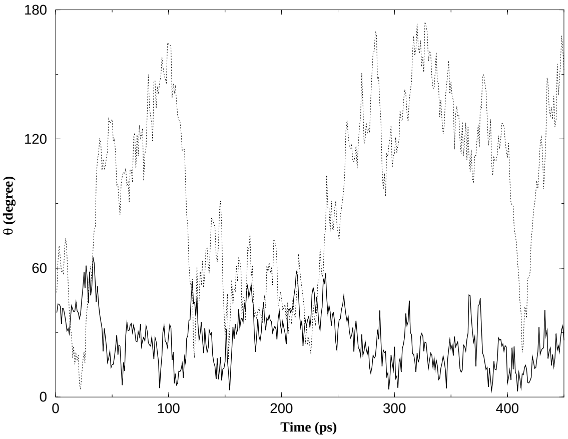


Figure 31. Time dependent angle *θ* of the guest (fluorobenzene) from the λ-dynamics simulations. The dotted line shows the result taken from the previous multiple topology λ-dynamics simulation with the explicit water model. [161] The solid line shows that from the hybrid topology λ-dynamics/GB simulation.

5.3.7 Multiple and hybrid topologies

Free energy results from simulations using hybrid topology and multiple topology representations show good agreement. Here we discuss the differences between them. FEP simulations using hybrid topology models converge much faster than those using multiple topology models since in the latter the guests having λ≈0 adopt unimportant high energy states, adding uninteresting configurational entropy contributions (data not shown). For the same reason, the hybrid topology λ-dynamics method yields converged results without the restraining potential that is essential for rapid convergence in the multiple topology λ-dynamics method. However, the use of a hybrid topology results in more limited sampling of conformational space. For example, an unselected guest is restricted to an orientation dictated by the dominant guest, since they share the *colo* atoms. In the multiple topology model the non-selected guest is subjected to only the restraining potential, the strength of which can be selected at will, and adopts a wider range of orientations within the binding site. This is demonstrated by the time dependence of the angle θ of the guest (fluorobenzene) from the λ-dynamics simulations (shown in **Figure 31**). By using a multiple topology representation, fluorobenzene adopts multiple bound conformations, i.e., near the initial structure ($\theta < 90°$) and also an alternate orientation ($\theta > 90°$). On the other hand, the molecule is restricted to regions near the initial orientation throughout the simulation using a hybrid topology representation. To obtain reliable free energies, one must ensure both satisfactory convergence and sampling of all the relevant minima. Thus, one must consider the configurational sampling of all ligands and use the hybrid topology representation with caution. The iterative procedure with a biasing potential may partially solve any sampling problems in the hybrid topology model.

5. 4 Conclusions

β-CD provides a simple prototype for ligand-enzyme systems, involving mainly van der Waals and hydrophobic interactions. We have chosen this system as a test case for the λ-dynamics methodology due to its simplicity. The excellent correlations we have obtained between the free energies of binding from λ-dynamics and FEP simulations reaffirms the capability of this approach as a faster alternative to FEP, particularly when a large number of ligands are involved. A significant difference in free energy of binding between the “best” and “worst” members results in the latter not sampling the dominant state. However, when such a situation arises, additional biases can be applied and the free energies estimated using the enhanced sampling. These simulations have also demonstrated how λ-dynamics may be used to sample the conformational space for alternative binding orientations by tuning the strength of the restraining potential and by using biased sampling.

In this application of the GB solvent model, a hybrid topology λ-dynamics/GB method was used to study the binding of seven mono-substituted benzene derivatives to β-CD. The effect of non-electrostatic solvation terms approximated by solvent accessible surface area was very small at least in this system. The binding free energy differences obtained from the λ-dynamics/GB method agreed well with those of FEP simulations in both explicit and implicit solvent models. Carefully chosen GB parameters, which depend on the size of the system, are very important to yield results that are comparable with those obtained using explicit water. We demonstrated that a λ dependent partial charge model introduced for the hybrid topology representation worked well. The λ-dynamics/GB simulations converged without any restraining potential, although the configurational space sampled was restricted as compared to sampling using the multiple topology model. The hybrid topology λ-dynamics method may be a useful tool for screening out the slightly varying ligands.

Chapter 6

Protein Stability Analysis:

c-Myb DNA-Binding Domain

Based on

Shinichi Banba and Charles L. Brooks, III,

“Application of multiple topology λ-dynamics to protein stability analysis: c-Myb DNA-binding domain,”

in preparation.

6.1 Introduction

Enzymes are excellent biocatalysts so that they are expected to have applications in industry, however, they are very sensitive to chemical and physical factors.[163] Therefore, it is very important to enhance the protein thermal stability and the tolerance for chemical compounds in their applications as catalysts. For example, enhanced thermal stability is a key factor in polymerase chain reactions (PCR) applications or in the detergent industry.[164] To create the artificially thermostable proteins, mutational analyses have been carried out systematically in many proteins such as T4 Lysozyme[165] and staphylococcal nuclease.[166,167] Moreover, the 3D structures of many thermophilic proteins and thermostable mutants have been solved and published. Hence, by increasing our understanding of the molecular-level origins of protein stability, we will be able to move toward protein design. Such rational design is still relatively elaborate and slow because it is an iterative process and each design must be characterized biophysically before making the next change. Therefore, an efficient and accurate methodology, which predicts the thermal stability of extensive mutants, will play an important role for the protein design process.

For this purpose, many researchers have recently developed computational methods for predicting the effect of mutations on protein stability.[168-171] Until now such calculations have been performed mostly with expensive free energy simulation methods. The expensive free energy simulation methods such as free energy perturbation (FEP) have successfully been applied to rationalize changes in protein stability caused by mutation.[172-175] However due to the computationally intensive nature of these methods, a more efficient computational protocol is required as a practical tool. In addition to the fact that long simulations are often necessary to obtain satisfactory convergence, FEP is not the method of choice when a large number of mutations have to be screened.

In the previous chapters, the λ-dynamics and CMC/MD methods have been demonstrated to be effective in the rapid evaluations of binding free energies of a large number of ligands from a single simulation.[13] The λ-dynamics method has already been successfully applied for protein-ligands or host-guest systems. [6,14,39,40,84,161,176] A variant of the λ-dynamics approach, CMC/MD, has been applied to protein stability analysis by Pitera and Kollman. The CMC/MD method successfully identified and ranked the thermal stability of 8 different mutants at a specific site of T4 lysozyme.[177]

Our objective of this chapter is to further expand the λ-dynamics and CMC/MD approaches for protein stability analysis. In the previous applications, multiple ligands are prepared against a specific protein or host. Then, the λ-dynamics and CMC/MD methods are used to screen out the ligands that bind to the protein or host most tightly. In this study, multiple amino acids at a given site of the protein are prepared and then the λ-dynamics and CMC/MD methods screen out the most stabilizing mutations from putative candidates.

6.2 Computational details

The system we study is the DNA-binding domain of c-Myb (R2 unit) with different amino acids at position 103 (*Val*); a-butyric amino acid (*Abu*), alloisoleucine (*Ail*), Alanine (*Ala*), isoleucine (Ile), Leucine (*Leu*), and, norvaline (*Nva*) (see **Figure 33**). c-Myb is a transcriptional regulatory factor, playing an vital role in the regulation of the proliferation of hematopoietic cells, mature T-cells and muscle cells.[178] The DNA-binding domain of c-Myb has three imperfect structural repeat units (R1, R2, and R3). NMR analyses revealed that the thermodynamically less stable R2 unit has a cavity in the hydrophobic core as shown in **Figure 32**.[179] Free energy calculations and cavity-filling mutants proved that the decreased thermal stability of R2 domain is derived from this cavity,[174,175,180] however, it plays an important role in DNA binding. In fact, a cavity-filling mutant, V103L, had a decreased binding affinity for DNA.[179] In this system, the free energy changes associated with unfolding were measured using urea denaturation experiments.[174,180] The tertiary structure of the wild type was determined by NMR analysis (PDB code: 1MBG).[179] Saito, et al. have already investigated this system to clarify the stabilization mechanism by using AMBER force field.[174,175] Therefore, this system is an appropriate choice for investigating the λ-dynamics approach for protein stability analysis.

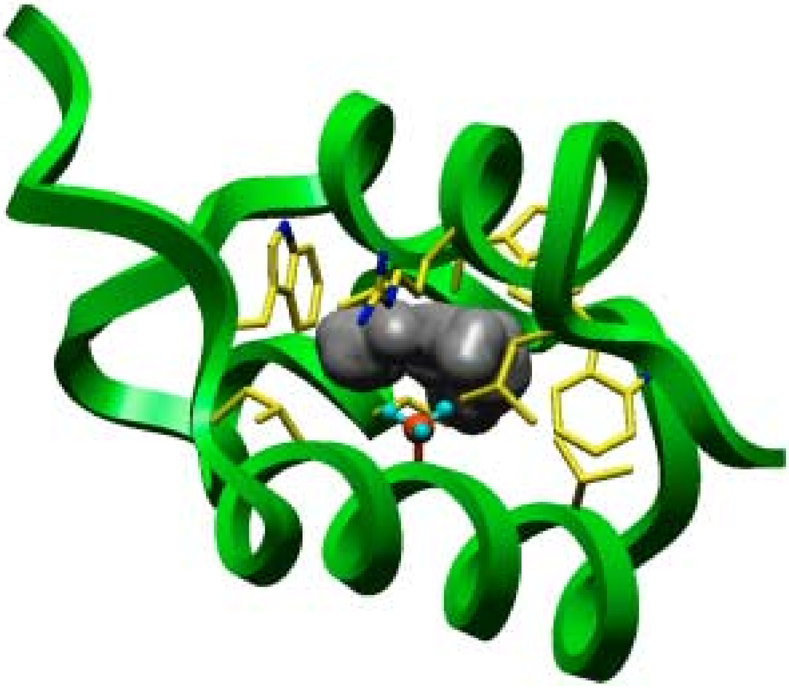


Figure 32. c-Myb R2 structure. The cavity surface detected with a mutant V103A is shown by gray .The alanine mutated here is shown by red. The residues forming the cavity wall are shown by sticks. The figure is drawn by InsightII.

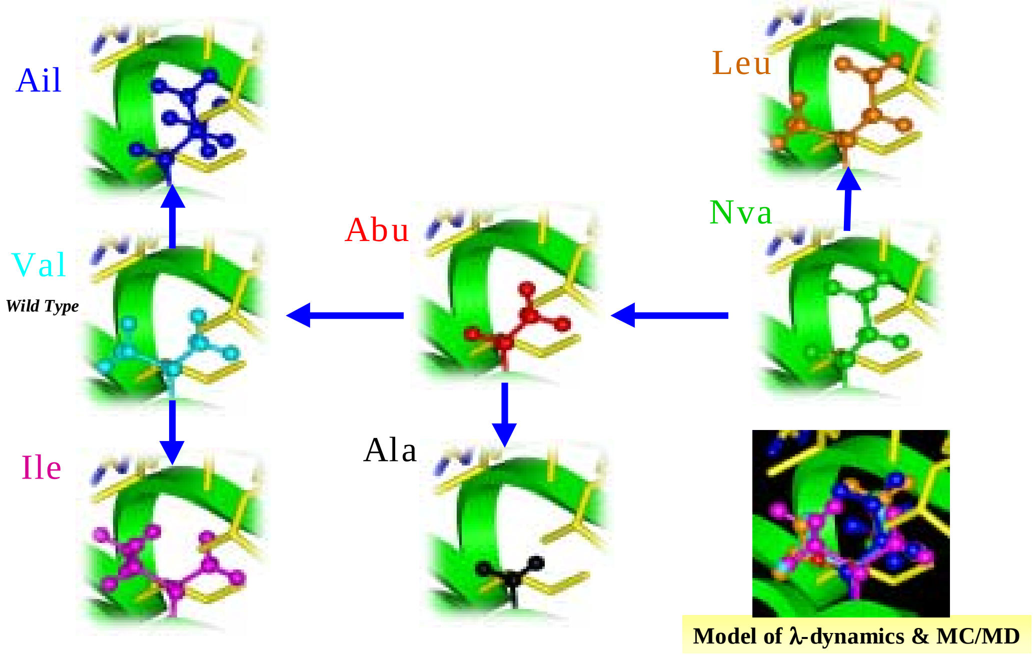


Figure 33. Amino acids and the mutation paths calculated in this study. The conformations shown here roughly represent the initial conformations used in this study.

In this chapter, we calculate the folding free energy of the protein using the thermodynamic cycle as shown in **Figure 34**. In computational approaches, the relative folding free energy difference($\Delta\Delta G_{fold}$) is calculated from ΔG_{fold} - ΔG_{unfold}.The values of ΔG_{unfold} for the half-cycles with unfolded states were pre-calculated using the FEP method, and then, λ-dynamics or CMC/MD are used to evaluate the ΔG_{fold} to obtain the folding free energy differences.

$$\begin{tikzcd} P^{wild}_{unfold} \arrow[r, "\Delta G_{wild}"] \arrow[d, swap, "\Delta G_{unfold}"] & P^{wild}_{fold} \arrow[d, "\Delta G_{fold}"] \\ P^{mutant}_{unfold} \arrow[r, swap, "\Delta G_{mutant}"] & P^{mutant}_{fold} \end{tikzcd}$$

Figure 34. Thermodynamic cycle used for relative folding free energy calculations

All computations were performed using the CHARMM molecular dynamics package.[119]

Mutated amino acids used in this study was shown in **Figure 33**. The backbone atoms and C_β of the mutated amino acids are represented by single topology, while variable side chain atoms of the mutated amino acids are represented by independent multiple topologies. We use the CHARMM version 22 parameters and topologies except for the partial charges of some variable side chains, which were slightly modified to confer total partial charge of the mutants zero.[138]

The initial coordinates of the native state were taken from the NMR structure (PDB code: 1MBG). Each mutated residue was modeled in by hand using QUANTA. Its initial conformation was chosen following the work of Saito et al.[174] The system was first solvated in a 20 Å sphere of water using the stochastic boundary molecular dynamics method,[122] and those water molecules whose oxygen overlapped within 2.8 Å of any non-hydrogen protein atoms were removed. Water molecules were represented by the TIP3P water model of Jorgensen.[120] The values of λ^2 for each mutant are kept constant (e.g. λ^2 = 1/6, for all mutants) while preparing the initial solvated structures. The water molecules were only allowed to equilibrate for 30ps of MD simulation, using a 12Å reaction region and a 8Å buffer region. The geometric center of Val103 was used as the center in partitioning the system. Additional 20Å spheres of water were then overlaid to fill any ‘holes’ because water may have moved into protein cavities during thermalization. This was followed by another 30ps of water thermalization and energy minimization without any constraints. We adopted the final structure as an initial structure for all later simulation. For example, to get $\Delta\Delta G_{fold}$ between any two mutants using conventional FEP methods, the other five mutants were deleted from the final structure and used as initial structure.

To assess the results of the λ-dynamics and CMC/MD methods, we also performed conventional FEP calculations with the native state. FEP calculations was carried out with six mutation pairs as shown in **Figure 33**. Only one methyl group is different in each mutation pair. The larger volume mutant was defined as a reactant in each pair. Mutations were carried out over 9 windows using a dual topology model. After an initial 100ps MD simulation for equilibration, an equilibration phase of 20ps and sampling phase of 50ps was carried out in each window.

To obtain the folding free energy difference, FEP calculations were performed with the denatured states using the same mutation pairs as those with the native states (**Figure 33**). At first, the denatured state was modeled with penta-amino acids taken from the actual sequence. The denatured states were also represented by mono peptide model (Acetyl-Xaa-Methylamino, where Xaa denoted the mutated amino acids). Side chain conformations were chosen to be the same as those in the native states. Since the mutants used here are non-polar aliphatic mutants, the mono-amino acid model and penta-amino acids model yielded the almost identical results. Thus, the denatured state was modeled as a mono-amino acid. This peptide model was immersed in a periodic box of 1000 TIP3P water molecules, and again those water molecules whose oxygen overlapped within 2.8 Å of any non-hydrogen atoms were removed. For each pair, a 100ps MD simulation for equilibration was followed by an equilibration phase of 20ps and a sampling phase of 50ps over 9 windows. This free energy is used as the reference free energy $\{F\}$.

All bonds containing hydrogen atoms were constrained to their parameter values using the SHAKE algorithm.[121] Nonbonded interactions were treated using a cutoff of 12.4Å along with van der Waals switching between 8.0Å and 10.0 Å and an electrostatic shifting function. The temperature of the system was maintained near 300K by coupling the non-hydrogen atoms to a Langevin heatbath using frictional coefficients of 50 ps^{-1} and 5 ps^{-1} for atoms and λ variables, respectively. All simulations used a time step of 1.0 fs. The masses of the fictitious λ degrees of freedom were chosen to be 5 amu•Å^2 in all λ-dynamics simulations.

To keep the side chains of the unselected mutants in the lower energy states, the restraining potentials (see **Eqn. 44**) are added in the λ-dynamics and CMC/MD simulations in this application. Therefore, the $\Delta\Delta G_{fold}$ of the mutants is estimated by using **Eqn. 46** for a single trajectory and **Eqn. 48** when combining multi-trajectories with WHAM. MC steps occurred every 10 MD steps in all CMC/MD simulations. The λ trajectories obtained from λ-dynamics and

CMC/MD simulations were saved every 10 fs and were used for later analysis. By applying biasing offsets for {F} and an iterative technique, we expect to achieve better sampling of the phase space of chemical coordinates and therefore faster convergence of the calculations. Thus, we performed the λ-dynamics and CMC/MD simulations using an iterative procedure with WHAM up to 8 iteration. In each iteration, a 50 ps equilibration period was followed by 250 ps of data collection. As described before (**Eqn. 47**), biasing offsets {F} for *n*-th iteration are estimated from previous (*n-1*) trajectories.

6.3 Results and discussion

6.3.1 Relative folding free energy differences

The relative folding free energy differences calculated by FEP, λ-dynamics and CMC/MD are tabulated in **Table 8**. The FEP results using Amber force field are also listed in **Table 8** for comparison.[174] From a single 250ps simulation, both the λ-dynamics and CMC/MD methods successfully identified the best stabilized mutant (*Leu*). However, a single run was insufficient to obtain the $\Delta\Delta G_{fold}$ for noncompetitive mutants such as *Abu* and *Ala*. Therefore, the iterative procedure with biasing offsets {F} was carried out to get the $\Delta\Delta G_{fold}$s for all mutants. As a result, all methods gave good agreement with experiments as shown in **Figure 35-a**, however, our simulation results overestimated the relative folding free energy differences. Both the λ-dynamics and CMC/MD methods were in good agreements with FEP results (**Figure 35-b**) and consequently demonstrated their methodological validity. As shown in **Figure 35-c**, the different scaling parameters (α=0.1 and 0.3) for the restraining potential gave really good agreement. Thus, the scaling parameters may influence very little on the final results as long as the convergence is enough among similar mutants. Our simulations using CHARMM force field were also in good agreement with the FEP simulations using AMBER force field (**Figure 35-d**).

Table 8. Summary of relative folding free energy calculations [a]

amino acid	Exp.[b]	FEP[c]	FEP	λ-dynamics				MC/MD			
				α=0.1		α=0.3		α=0.1		α=0.3	
				1st	5th[d]	1st	5th[d]	1st	5th[d]	1st	5th[d]
Leu	-3.96	-4.45	-3.32	-2.6	-3.32	-4.1	-3.18	-4.32	-3.58	-3.05	-2.94
Nva	-2.4	-2.94	-1.14	-0.65	-1.43	-1.4	-0.96	-1.39	-2.20	-0.91	-0.84
Ail	-1.27	-1.77	0.12	-0.51	0.31	-0.15	-0.34	0.4	-0.98	0.61	-1.00
Ile	-1.07	-0.13	-0.2	0.02	-2.19	-1.81	-0.78	-2.5	-2.15	-1.78	-1.49
Abu	-0.2	0.15	1.34	-[e]	1.20	-[e]	1.06	-[e]	0.10	2.27	1.65
Val	0	0	0	0	0	0	0	0	0	0	0
Ala	1.47	2.79	3.9	-[e]	4.90	-[e]	5.19	-[e]	3.87	-[e]	5.65

[a] All values are in kcal/mol, and relative to Valine (wild type).

[b] $\Delta\Delta G_{fold}$s are taken from Saito et al.[174]

[c] Calculated values using Amber force[174]

[d] The values were calculated with 5 trajectories using WHAM

[e] Not Determined because they did not reach dominant states ($\lambda^2 > 0.9$).

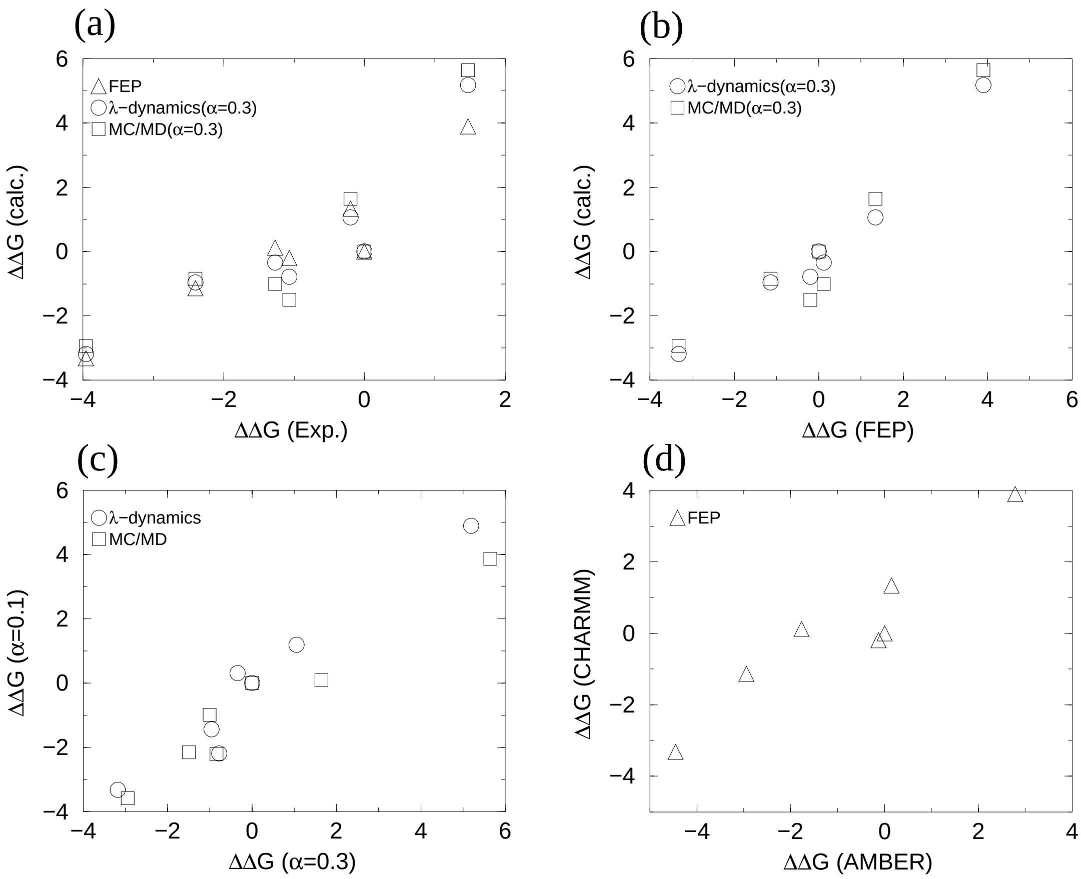


Figure 35. Correlation plots of the relative stability free energy of the mutants in kcal/mol. (a) Comparison with observed values; (b) λ-dynamics and CMC/MD versus FEP; (c) Comparing the different scaling parameters (α=0.3 versus α=0.1); (d) $\Delta\Delta G_{fold}$ using CHARMM force field versus $\Delta\Delta G_{fold}$ using AMBER force field.

The λ^2 trajectories are shown in **Figure 36**. Since *Leu* provides the greatest stability by 2 kcal/mol in the FEP simulations, *Leu* mostly occupied the dominant states in the first λ-dynamics run without any biasing offsets (i.e., relative free energies in the denatured states $\{F^0\}$ were used). In contrast, all mutants compete reasonably well in the fifth run with biasing offsets, $\{F^{(4)}\}$, which were estimated from the four λ-dynamics trajectories.

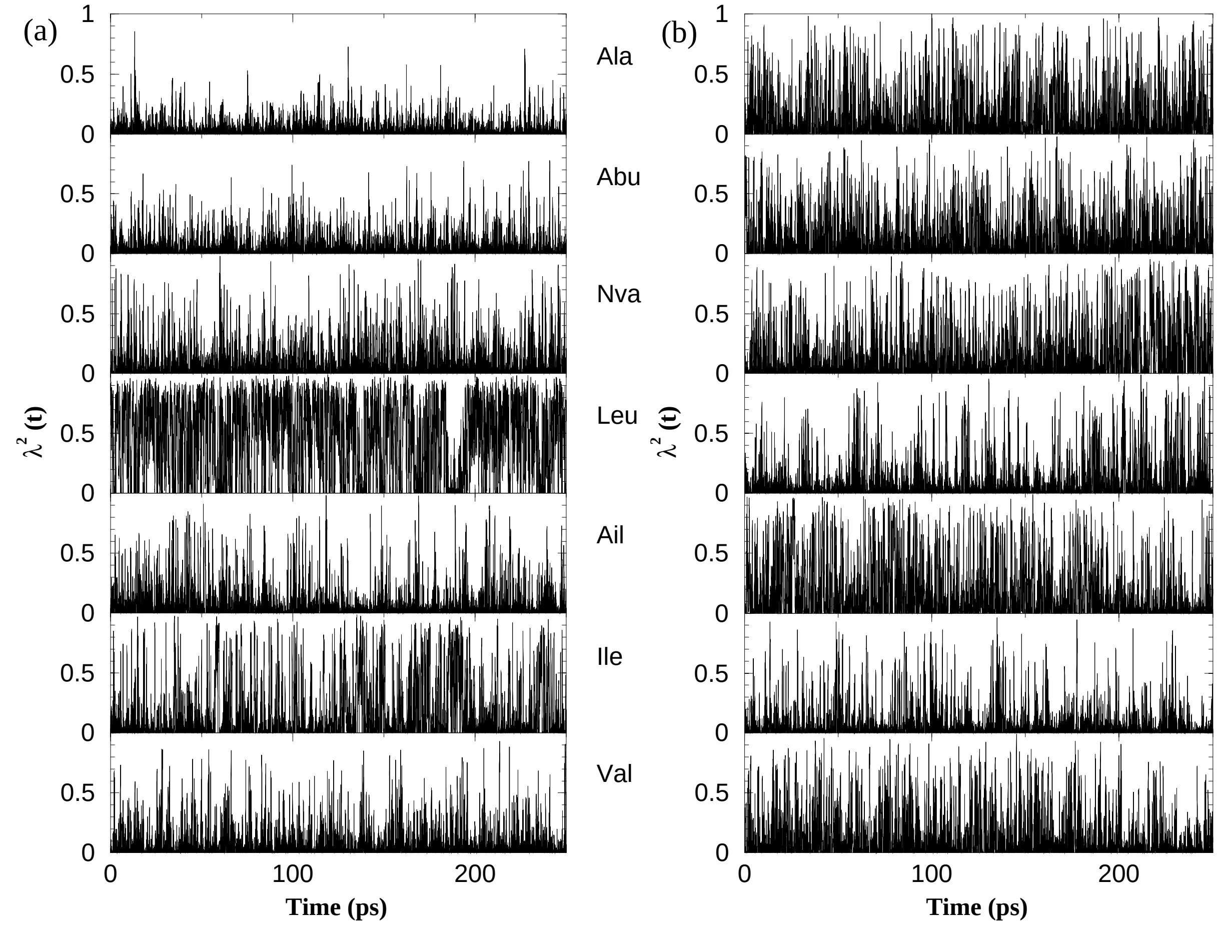


Figure 36. Trajectory of λ^2 values obtained using the λ-dynamics method with α=0.3. (a) Data from first λ-dynamics run with {F^0} corresponding to the respective free energies at the denatured states. (b) Data from 5th λ-dynamics run with {$F^{(4)}$}, which were estimated by using four λ-dynamics trajectories based on WHAM.

The convergence profiles of the $\Delta\Delta G_{fold}$s along iterative cycles (**Figure 37**) indicate that two or three iterations are enough to obtain the converged folding free energies for all mutants in this system. Although this behavior may depend on the system and quick convergence may partially come from the simplicity of this system, such a fast convergence demonstrates the effectiveness of the iterative techniques with WHAM.

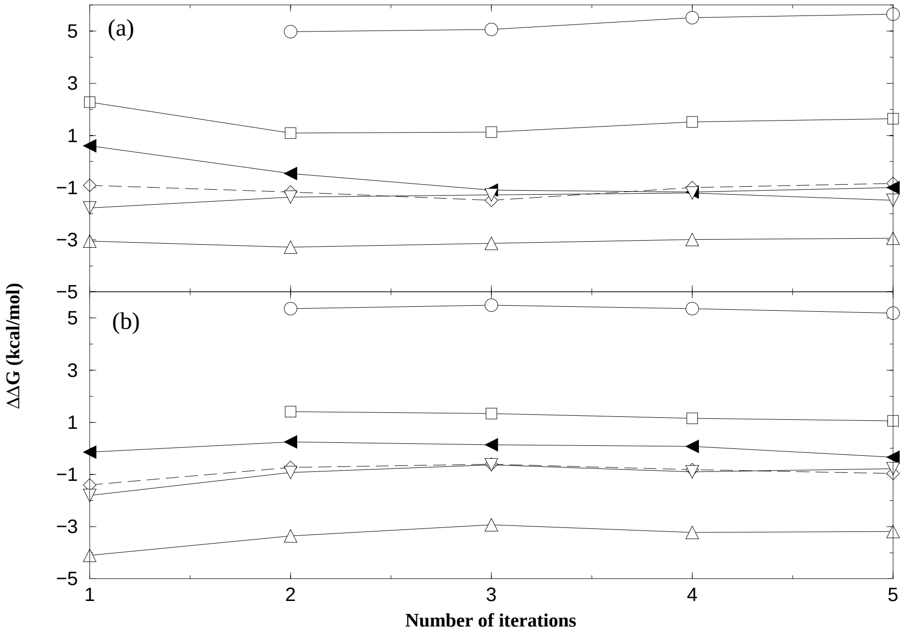


Figure 37. Estimations of $\Delta\Delta G_{fold}$ as a function of the number of iterations for (a) CMC/MD, (b) λ-dynamics. A value of 0.3 was used for α. *Val* was chosen as reference. Key: *Ala*-circle, *Abu*-square, *Ail*-filled triangle, *Nva*-diamond, *Ile*-triangle down, *Leu*-triangle up.

6.3.2 Biasing potential along the λ coordinates

In the CMC/MD method, the stochastic sampling by MC step permits one to restrict the sampling of the chemical space {λ} such as described in **Eqn. 25**. This condition allows us to sample exclusively the end states of interest. However, the smaller overlap of the end points has a risk of inefficient sampling along λ coordinates and getting trapped in a chemical state.

On the other hand, in the λ-dynamics method, the λ-variables are treated as continuous variables and explore both end points and the intermediate states. Using biasing offsets {F} in the iterative procedure, the ratio of sampling the end states decreased in the λ-dynamics simulations as shown in Table 9. This is because the number of the intermediate state is much larger than those of the end states and biasing offsets {F} make the intermediate states competitive as well as the end states. To increase the ratio of the end states, an additional biasing potential V_{bias} (see **Eqn. 27**) along the λ coordinates is added as following.

$$V_{bias} = \sum_{i=1}^{n} \begin{cases} -10(\lambda_i^2 - 0.5)^2 & (\lambda_i^2 > 0.5) \\ 0 & (\lambda_i^2 \le 0.5) \end{cases} \tag{74}$$

This biasing potential stabilized the end states about 2.5 kcal/mol. In this application, the same V_{bias} was applied to all mutants. In this case, $\Delta\Delta G_{fold}$ can be calculated without including the effect from V_{bias} since its effect is completely canceled (see **Eqn. 29**). As shown in **Table 9**, the additional biasing potentials successfully increased the ratio of sampling the end states about 30 times as compared to that without them.

Table 9. Number of sampling the end states in λ-dynamics simulation (α=0.3).[a]

	1st	2nd	3rd	4th	5th	5th(biased potential)[b]
Number of End states(λ^2>0.9)	780	626	174	257	264	7595
Number of Transitions	68	44	57	83	74	162

a There are 25,000 samples in each iteration.

b 5th iteration with the additional biased potential shown in **Eqn. 74**.

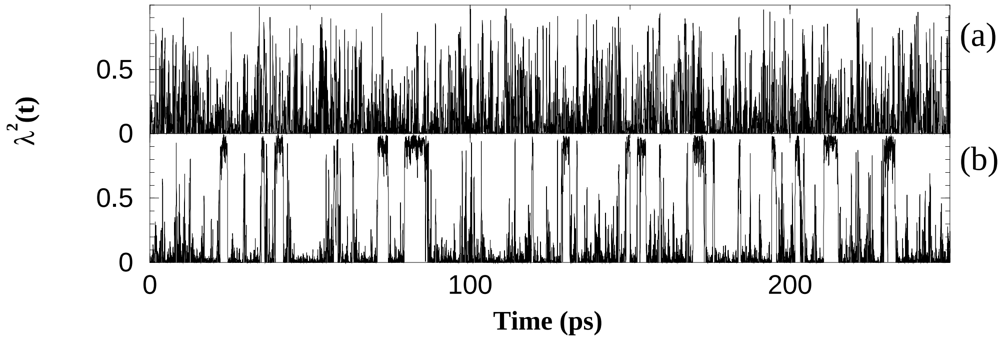


Figure 38. Trajectory of λ^2 values of R=*Ala* obtained from the 5th λ-dynamics trajectories with α=0.3. (a) No biasing potential; (b) with the biasing potential shown in Eqn. 74.

To further demonstrate the effect of V_{bias}, the λ^2 trajectory of the *Ala* mutant from the 5th run is presented in **Figure 38**. Without V_{bias}, the *Ala* mutant occupied the dominant state for a very short period and return to the unselected states, while it occupied the dominant states for a long time with this biasing potential. Furthermore, the transitions between the end points happen more frequently without a trap in an end point (see Table 9). Thus, the properly chosen biasing potentials successfully permit both the smooth transitions and efficient sampling of the end points.

6.3.3 Conformational sampling of the mutant side chains

An attractive feature of the λ-dynamics method is its ability to explore a larger conformational space. This is due to the fact that when λ is nearly zero, the only interaction is the restraining potential. By controlling the scaling parameter α for the restraining potential, we can control the extent of conformational space to be sampled. The distribution of χ_1, and χ_2 observed during the λ-dynamics simulation is illustrated in **Figure 39**. With the weaker restraining potential (α=0.1), the mutants explored different local configurational minima. Furthermore, *Ail* and *Leu* occupied the dominant states with a different conformation with respect to χ_2. In the CMC/MD simulations, Pitera and Kollman introduced rotameric states for each mutant in order to compensate for the inefficient conformational sampling of the side chain. In this application, the smaller scaling potential (α=0.1) yields the broader sampling of the conformational space and consequently is substituted for the set of rotamers used in their CMC/MD study.

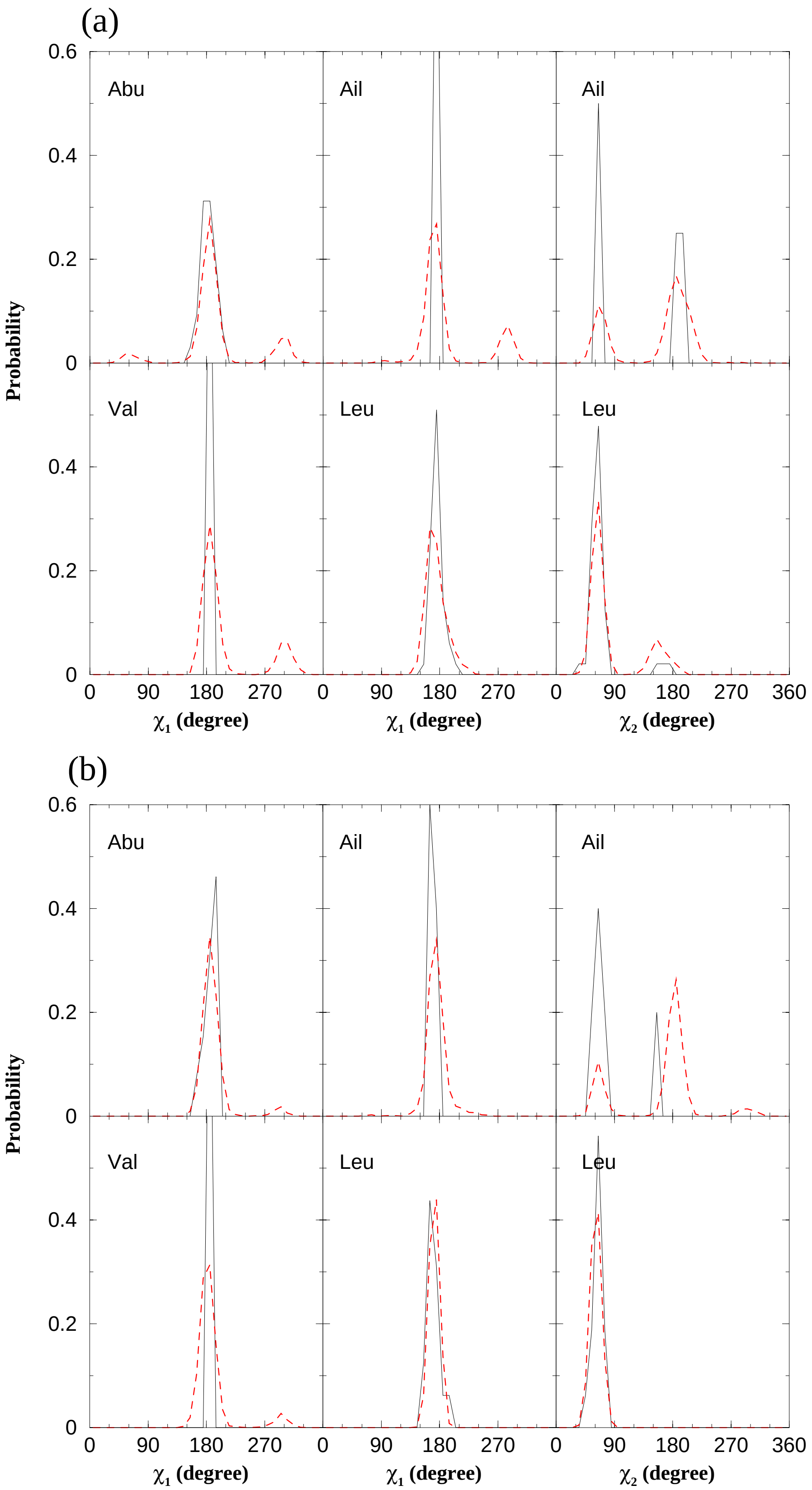


Figure 39. Distributions of the dihedral angle of the side chains from the 5th iterate of λ-dynamics trajectories for α=0.1 (a), α=0.3 (b). The dotted lines represent those using the whole trajectories, while the solid lines show those only occupying the dominant states.

6.3.4 The effect of mass of the fictitious λ degrees of freedom

To provide a standard protocol for future applications, we examined the effect of mass of the fictitious λ variables (m_λ), which provides additional control over the dynamics of the λ variables. Since the derivatives of the λ variables are related to the potential energy $V_i(X,x_i)$ in **Eqn. 14**, the distribution of $V_i(X,x_i)$ has an influence on the proper value of m_λ. For example, when a large number of atoms are assigned as the variable atoms (x_i in **Eqn. 14**), the distribution of the potential energy $V_i(X,x_i)$ becomes wide and then large m_λ may be preferred. Here, we limit our discussion to the point mutations cases.

The λ-dynamics simulations were carried out using the conditions of the 5th iteration (α=0.3) except for the values of m_λ. The smaller mass (m_λ =1) made the numerical integration of the λ variables unstable and insufficient coupling of the λ variables with a heatbath resulted in a higher temperature for the λ variables as shown in **Table 10**. With even smaller masses (m_λ = 0.01), the unselected mutants yielded the larger fluctuations near zero point and the ratio of sampling the end states became small. With m_λ =0.01, the λ-dynamics method yielded the incorrect $\Delta\Delta A_{fold}$s even with 2 ns simulation. Although smaller time step is inefficient for evolving the atomic coordinates, smaller time step may be necessary to evolve the λ coordinates when such small masses are chosen.

On the other hand, the larger m_λ (m_λ= 500) showed slow evolution in the λ-space, but $\Delta\Delta A_{fold}$s calculated with m_λ =500 with 2 ns trajectory gave good agreement with those using m_λ=5 (data not shown). Too large a mass (m_λ= 5000) tended to trap the system in a state where one mutant occupied the dominant states and then much longer simulation time is required to explore the whole λ space. In this case, each mutant spent about a few tens of picoseconds to reach the dominant state from the unbound state. Much longer time step is efficient for evolving λ variables, however, such a large time step yields unstable numerical integration of the atomic coordinates. From this study, we conclude that the proper mass for the λ variables is about 5 ~ 50 amu•Å^2. This is small enough for efficient sampling of the λ space and large enough for stable numerical integration at a given time step (1 fs) for the atomic coordinates. The properly chosen m_λ is important for efficient λ-space sampling.

Table 10. The effect of mass for the λ-dynamics simulations.[a]

m_λ **(amu•Å^2)**	**0.01**	1	**5**	**10**	**50**	500	**5000**
Number of end states [b]	16(12)	259(5)	264(73)	256(73)	350(30)	113(3)	456(2)
Average temperature of {λ}	1023	1010	310	301	301	301	301

[a] All λ-dynamics simulations were carried out using the conditions for the 5th iteration except for the values of m_λ.

[b] There are 25,000 samples in each simulation. The number of transitions between the end states is also shown in the parentheses.

6.3.5 The limitations in the free energy based screening methods

The mutations investigated here were restricted to similar aliphatic residues. To further evaluate free energy based screening methods, we added five structurally and/or electrostatically dissimilar mutants: asparagine (*Asp*), methionine (*Met*), phenylalanine (*Phe*), serine (*Ser*), and threonine (*Thr*). The additional mutation paths for the FEP simulations are shown in **Figure 40**. In this case, a total of 12 mutants were examined simultaneously using λ-dynamics and the CMC/MD methods. All simulations (i.e., FEP, λ-dynamics and CMC/MD) used the same equilibrium and production protocols described above.

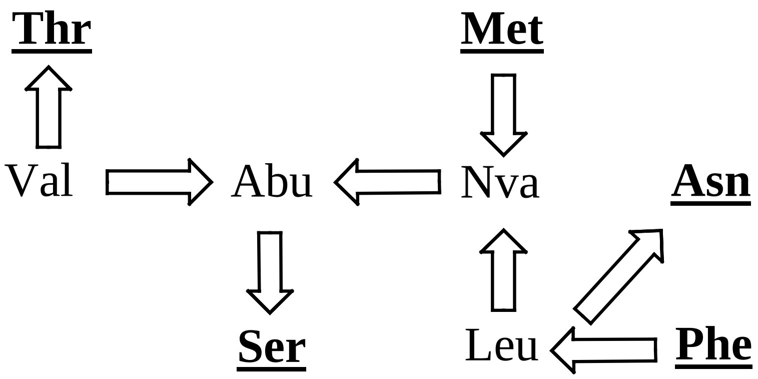


Figure 40. Amino acids and the mutation paths calculated in 12-mutants system. ***Asn, Phe, Met, Ser*** **and** ***Thr*** **are the additional mutants.**

In the CMC/MD method, the stochastic sampling by an MC step restricts sampling exclusively to the end states of interests (**Eqn. 25**). However, inefficient sampling of the chemical states, such as trapping in an end state, may occur. This is prevalent when there is a large free energy gap between the side chains. Trapping may, however, be partially avoided by the addition of a few chosen intermediate states to bridge the end points. In the 12-mutant system, the CMC/MD simulations, which sample only the end states, did not reproduce the FEP results even after the eighth iterative cycle. This is because the MC steps tended to get trapped in one of the mutant states (see **Table 11**). Additional intermediate states seem to be important to avoid trapping in local minima, however, allowing one intermediate state (λ_i^2=0.5 and λ_j^2=0.5) was insufficient to enhance the efficiency of sampling the chemical space. The CMC/MD simulations with nine intermediate states increased the transitions of the end states and yielded relatively smaller absolute free energy errors after the eighth iteration.

In the λ-dynamics method, the λ-variables are treated as continuous variables, so smooth transitions between the end points are expected. After the eighth iterative cycle, the λ-dynamics method successfully predicted the folding free energy as compared to FEP results, except for the *Phe* mutant (see **Figure 41**). Both the λ-dynamics and the CMC/MD methods predicted the $\Delta\Delta A_{fold}$ of the *Phe* mutant to be much higher than that from the FEP result. Since the *Phe* residue is large compared to the cavity space, the protein has to change its structure. In **Table 11**, we did not observe transitions to the end states in any of the CMC/MD trajectories except for those that included nine intermediate states when the *Phe* mutant was in the dominant state. Thus, once the *Phe* mutant occupies the dominant state, it is difficult for the other mutants to compete, however, *Phe* rarely occupies this state.

Table 11. Number of sampling the end states [a] and the average absolute free energy errors along the iteration procedures.[b]

Iteration No.	**1st**	**2st**	**3nd**	**4th**	**5th**	**6th**	**7th**	**8th**
λ-dynamics (α=0.1)	4096(168) -	1375(132) -	1041(147) 1.39	887(158) 1.18	1314(174) 1.03	909(144) 0.95	757(141) 0.85	626(117) 0.81
λ-dynamics (α=0.3)	272(78) -	270(84) -	205(47) 2.26	513(17) 2.07	139(39) 1.89	221(58) 1.71	118(49) 1.61	114(38) 1.57
MC/MD [c] (α=0.1)	25000(2) -	25000(34) -	25000(0) -	25000(29) 2.32	25000(7) 2.82	25000(39) 2.88	25000(0) 2.90	25000(0) 2.90
MC/MD [c] (α=0.3)	25000(10) -	25000(45) -	25000(59) 3.96	25000(54) 3.87	25000(0) 3.74	25000(0) 3.75	25000(0) 3.85	25000(0) 3.87
MC/MD [d] (α=0.1)	23803(13) -	25000(0) -	24077(8) 2.0	22388(57) 2.40	21293(94) 2.33	21632(85) 2.37	24881(0) 2.30	25000(0) 2.28
MC/MD [e] (α=0.1)	16339(12) -	18167(13) -	12186(22) 2.06	24887(0) 2.09	24909(0) 2.21	24146(0) 2.13	4380(47) 1.99	3629(42) 1.88

[a] The number of the end states are shown at upper line in each cell. The number of the transitions among the end states is also shown in the parentheses. There are 25,000 samples in each iteration.

[b] Average absolute free energy errors (*kcal/mol*) against the FEP results are shown at lower lines in each cell. *Phe* is excluded due to its large errors.

[c] The sampling of the chemical space in MC steps are restricted only at end points.

[d] The sampling of the chemical space is allowed at both $\{\lambda_i^2=1, \lambda_{k\neq i}^2=0\}$and $\{\lambda_i^2=0.5, \lambda_j^2=0.5, \lambda_{k\neq i,j}^2=0\}$

[e] The chemical space sampling is allowed at $\{\lambda_i^2=0, 0.1, ..., 0.9, 1\}$ with the conditions $\sum_{i=1}^{L} \lambda_i^2 = 1$.

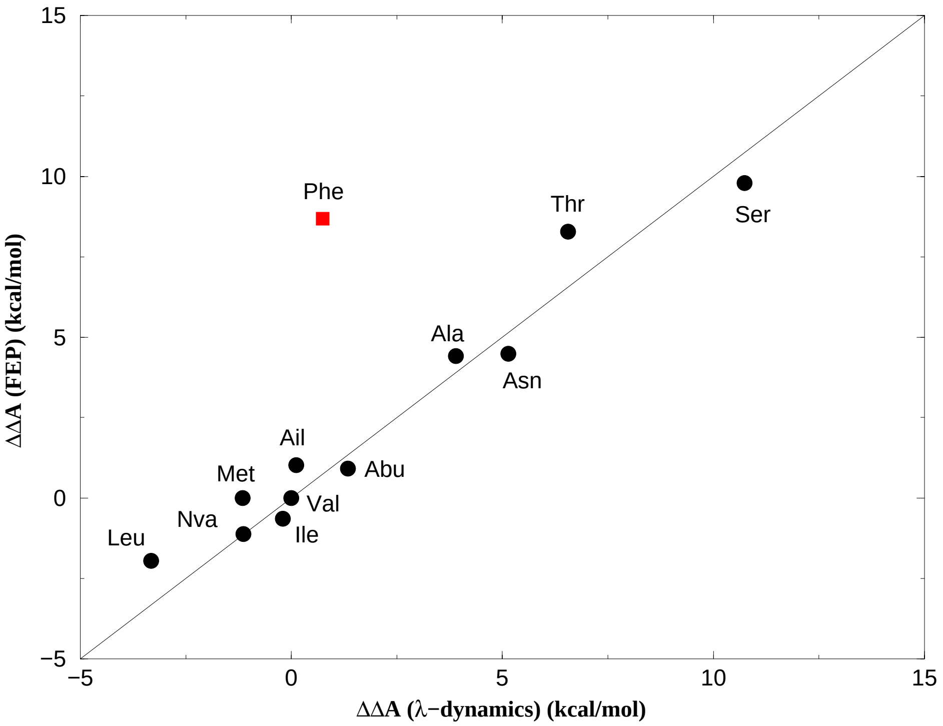


Figure 41. Correlation plot of the calculated relative folding free energies from λ-dynamics and FEP simulations in the 12 mutants system. There is good correlation, except for *Phe* mutant.

To clarify why the λ-dynamics ended in the incorrect estimation of $\Delta\Delta A_{fold}$ for *Phe*, constant temperature MD simulations were carried out with the wild type (i.e., *Val*) and *Phe* mutant. Judging from the conventional MD simulations, there are two distinct differences in the protein configuration between the wild type and *Phe* mutant (see **Figure 42**). One is the side chains of two *Trp* residues, which move to accommodate *Phe*. The other is the shifted third helix to provide the larger cavity space. When the *Phe* mutant occupies the end-point state during the λ-dynamics simulations, the two *Trp* side chains moved in a similar way, but the third helix remained in conformations near the wild type structure. The *Trp* side chains relax rapidly to accommodate the *Phe* mutation, whereas, the shift of the third helix requires a longer time and is hardly observed in the λ-dynamics simulations. Thus, we speculate that the λ-dynamics method predicted the incorrect folding free energy for the *Phe* mutant because the third helix remained in positions where the *Phe* mutant disfavored. On the other hand, both the *Trp* side chains and third helix shifted in the eighth CMC/MD trajectory (no intermediate condition) in which *Phe* was dominant nearly 1.2ns and the other mutants hardly occupy the dominant states. Inefficient λ space sampling in the CMC/MD simulations permit the *Phe* mutant to remain in the dominant state for a longer period of time, which makes it possible for the third helix to shift. Once the third helix moves, the other mutants do not have favorable van der Waals interactions and do not return to the dominant states without the additional biasing offsets. In this study, the additional biasing offsets used in the iterative procedure only permit the third helix to transit between the minima for the *Phe* mutant and other mutants.

With both λ-dynamics and CMC/MD, the smaller scaling parameter (α=0.1) yielded smaller absolute free energy errors than α=0.3, as shown in **Table 11**. We assumed that the entropy in the correction (second and third terms in **Eqn. 46**) cancels. In the 12-mutant system, dissimilar mutants are included and entropic contributions from the correction terms may not cancel among them. Since the contribution from the correction terms increase as the value of the scaling parameter increases, the correction terms yielded larger errors with α=0.3. Furthermore, with the larger scaling parameter (α=0.3), the unbound mutants also occupied the lower energy states and then the intermediates states are competitive. The ratio of sampling the end states is so small that the convergence of the probability term (first term in **Eqn. 46**) is also slower (see **Table 11**). We speculate that both the inaccurate estimation of the correction terms and the slower convergence of the probability terms give the larger absolute free energy errors when α=0.3.

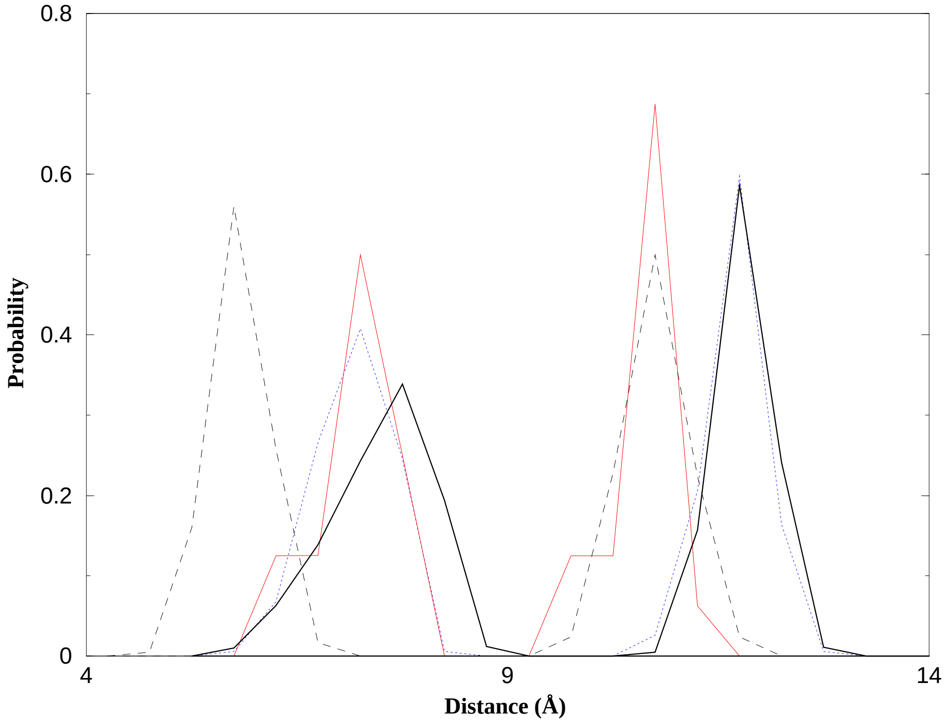


Figure 42. Distributions of the distances from the conventional MD simulation of *Phe* mutant (black solid line), the conventional MD simulation of the wild type (black dashed line), the λ-dynamics trajectory in which the *Phe* mutant occupies the dominant state (red solid line), and CMC/MD trajectory in which the *Phe* mutant occupies the dominant state (blue dotted line). Four distribution plots longer than 9 Å show the distances between C_α of 103rd mutant and C_α of 127th Gly, which represent the distance between the mutation site and the 3rd helix. The other four plots shorter than 9 Å are the distance between C_α of 103rd mutant and 4-position carbon atom of indole belonging to 134th Trp.

6.4 Conclusions

Our results indicate that the λ-dynamics and CMC/MD methods will become a practical tool for suggesting mutations that would stabilize a particular protein with applications in biotechnology. For screening purpose, both methods rapidly identified the favorable mutants.

Furthermore, computational time is almost independent of the number of mutants. For detailed free energy calculations using the iterative technique, both methods produce results in good agreement with experimental. Comparing λ-dynamics with CMC/MD, both methods yielded consistent results and only differed in handling the intermediate states. The additional biasing potentials along the λ coordinates (**Eqn. 74**) can enhance the ratio of the end states and then increase the efficiency of sampling without getting trapped in a local minimum.

By using the restraining potential with scaling parameter α, both methods successfully explore larger conformational space than conventional MD methods. We also found that the proper choice of mass for the λ variables is important for efficient λ-space sampling.

As a limitation, both methods are better suited for the quantitative screening of similar mutants. When sterically different mutant pairs such as *Phe* and *Ala* are considered, it is difficult for both methods to converge quickly. In this case, a large number of cycles of iterations using biasing offsets $\{F\}$ and/or enhanced sampling of the environment atoms around the mutants may be required. Moreover, putative mutants are divided into some groups which only have similar mutants and the λ-dynamics and CMC/MD simulations are carried out on the mutants in each of those groups. By partitioning the mutants into groups such that they have common members, $\Delta\Delta G_{fold}$ for structurally dissimilar mutants belonging to different groups also can be estimated. This strategy may overcome the limitation for both the methods.

Chapter 7

Binding Orientation of Toluene in β-Cyclodextrin

Based on

Shinichi Banba and Charles L. Brooks, III,

“Efficient sampling of guest orientations in β-cyclodextrin using hybrid Monte Carlo / Langevin dynamics method,”

J. Chem. Phys. submitted.

7.1 Introduction

The conformational space of biological systems such as proteins is so large that its complete sampling by conventional molecular dynamics (MD) or Monte Carlo (MC) methods is impossible. These methods tend to get trapped in one of large number of local minimum energy states at low temperatures. In addition, the presence of high-energy barriers between local minima hinders the exhaustive sampling of the system. As a result, these methods usually fail to obtain canonical distributions at low temperatures. The efficient sampling of the configurational space is important for faster convergence in the calculations of thermodynamic parameters. For example, conventional thermodynamic integration (TI) simulations gave incorrect binding free energy for benzylamine with trypsin due to inefficient sampling of benzylamine conformation in the bound state.[181] To overcome the multiple-minima problem, various novel simulation schemes, such as the umbrella sampling technique,[27], multicanonical algorithm,[86-90,182] replica-exchange method,[93,183,184] and simulated tempering method[185] have been proposed and successfully applied to many systems,[88,94,186,187] (For a recent review, see ref. [188]) however, each method has limitations. In multicanonical sampling or simulated tempering, the probability weight factors are not known a priori and have to be determined by iterations of short trial simulations, which is very tedious and time-consuming. In umbrella sampling also, multiple-trajectories have to be generated to compute the free energy difference.

In our previous studies shown above, the λ-dynamics method was demonstrated to be an efficient method for obtaining a reasonable estimation of the binding affinities of the ligands.[39,40] Additionally, these studies also proved to be more efficient for exploring the binding orientations and conformations of the ligands than using conventional MD. These factors motivated us to introduce the Monte Carlo / Langevin dynamics method (MC/LD) to overcome the multiple-minima problem. In the λ-dynamics method[6,13,14,39,40] or CMC/MD[15,16], the different ligands compete to get the relative binding free energy, whereas, in this MC/LD, the replicas compete to generate the canonical ensemble of {x} without becoming trapped in local minima. Although the Monte Carlo method was used for selecting the replica in this study, it is straightforward to apply the λ-dynamics method for this purpose.

7.2 Computational details

The system studied is toluene bound to β-cyclodextrin (β-CD). The sugar forcefield (par_all22.sugar) developed by Guyan Liang and John Brady (Cornell University 09/08/1995) was used to model β-CD. We adopted our starting orientation of toluene in the cavity as suggested in Liu et al.[159] such that the long axis of the guest is perpendicular to the plane of the β-CD and the hydrophobic substituent of the guest resides at the narrow rim of the truncated cone **(Figure 43)**. The crystal structure of β-CD-benzyl alcohol complex taken from the 'DEBGOG' entry of the Cambridge Crystallographic Database was used for preparation of the initial structure for simulations. The solvated host-guest complex model was prepared according to the previous protocols (see **Section 5.2.2.**). The final system contains 2940 environment atoms, and 15 focus atoms.

In this study, we defined the angle $\theta = \mathrm{arcos}\left(\left(\vec{A}\times\vec{B}\right)\bullet\vec{C} \Big/ \left|\vec{A}\times\vec{B}\right|\left|\vec{C}\right|\right)$ as shown in **Figure 43**. The outer product of vector *A* and *B* corresponds to the cavity axis of β-CD directed from the wide rim to narrow rim. The methyl group of toluene is directed to the narrow rim with θ=0° and the wide rim with θ=180°, respectively. All bonds containing hydrogen atoms were constrained to their equilibrium values using SHAKE.[121] The temperature of the system was maintained near 300K by coupling the non-hydrogen atoms to a Langevin heatbath using a frictional coefficient of 50 ps^{-1}. Nonbonded interactions were treated using a cutoff of 12.4Å along with van der Waals switching between 8.5Å and 10.0 Å and electrostatic shifting functions. The time step used in all simulations was 1.0fs. In MC/LD simulations, a Monte Carlo step was carried out at every 10 LD steps. The focus atoms were replicated five times, as consisted of a "selected replica" and four unselected replicas. A 50ps equilibration period was followed by a 200ps production run for the MC/LD simulation. The structural parameter θ was sampled at every 100 LD steps. All calculations were carried out using the CHARMM molecular dynamics package.[119]

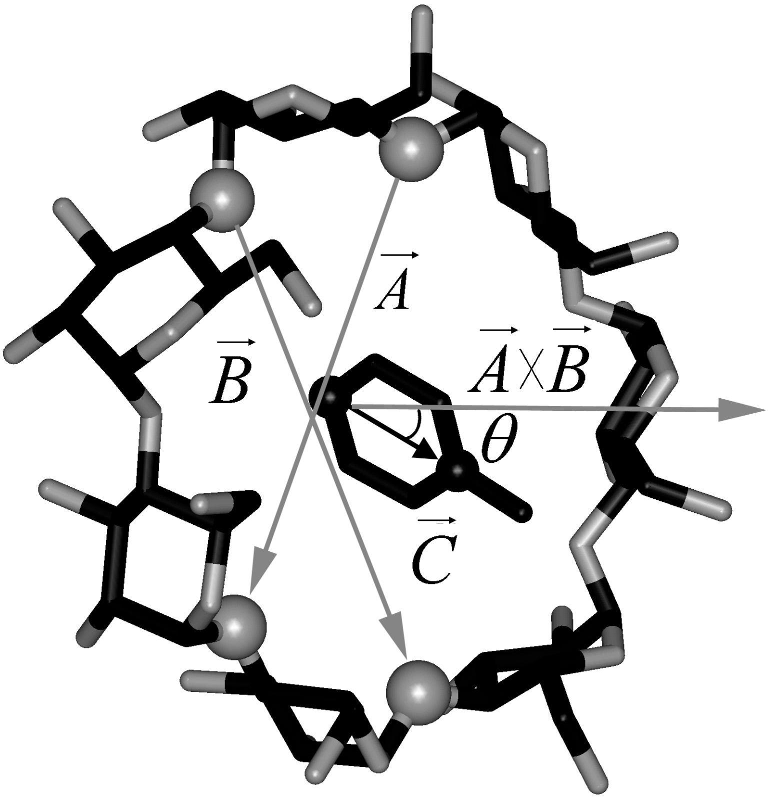


Figure 43. The definition of the structural parameter θ. The carbon atoms and oxygen atoms are shown in black and gray, respectively. The vector *A* and *B* are directed from the first bridging oxygen atom (larger gray spheres) to the fourth one and from the second one to the fifth one, respectively. The vector *C* is directed from the carbon at 4 position to that at 1 position of toluene shown by smaller black spheres.

7.3 Results and discussion

7.3.1 The Monte Carlo / Langevin dynamics simulations

Binding of toluene with β-CD can form two isomeric complexes.[158,159,189] It is difficult to experimentally determine the orientation of the guest molecule in the β-CD. In the initial structure of the complex, we placed the methyl group of toluene face to the narrower rim due to its hydrophobicity and size. When toluene is bound with the initial orientation ($\theta<90°$, "narrow" orientation), the methyl group is immersed in the hydrophobic cavity and has favorable van der Waals interactions with the host, whereas, within the "wide" orientation ($\theta>90°$, "wide" orientation) the dipole moment of toluene is anti-parallel to that of host and gave a favorable electrostatic interaction with host. The trajectory of θ values calculated by the MC/LD method is listed in **Figure 44** along with the conventional constant temperature LD trajectory. Although the conventional method was trapped in the "narrow" orientation near the initial structure, the MC/LD method successfully explored both orientations. The barrier crossing was achieved only by the unselected replicas under "ghost force", so that the frequent transition between two minima was carried out via MC steps. The distributions of the interaction energy V_{focus} (see **Eqn. 66**) of the selected replica and the unselected replicas are shown in **Figure 45**. Although the unselected replicas sampled higher energy states due to the scaled ghost forces, most of the time, one of many unselected replicas was in a state of lower energy enough to compete with the selected replica and resulted in the high acceptance ratio. The average acceptance ratio along the MC/LD trajectory was 9%. These results indicate that larger sampling space with the scaled "ghost force" was compensated by using the multiple replicas.

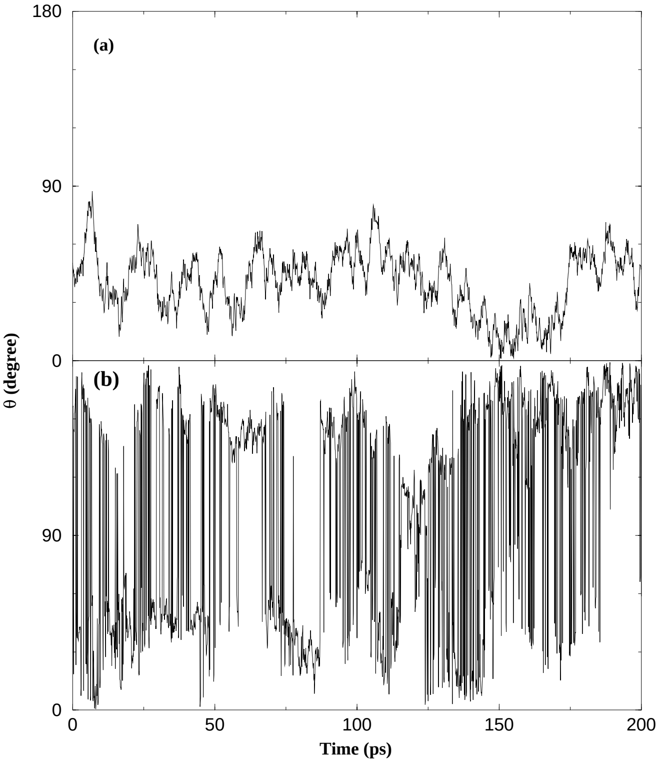


Figure 44. The dynamics trajectory of θ values for (a) the conventional LD simulation, and (b) MC/LD simulation of the selected replica. 50ps equilibration phases are not shown.

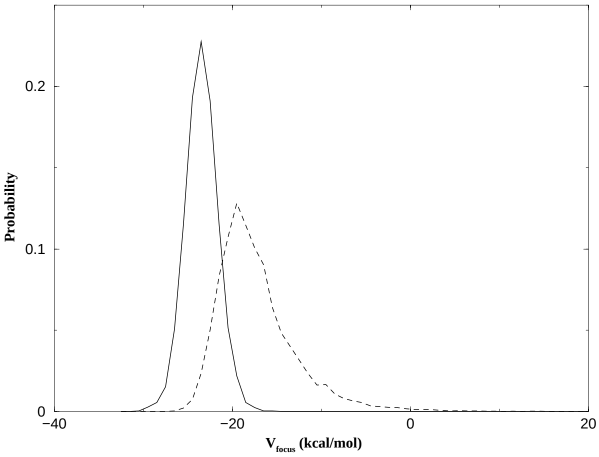


Figure 45. The distribution of interaction energy V_{focus} taken from the MC/LD trajectory. The solid line and dashed line represent the average of the selected replica and the unselected replicas, respectively.

7.3.2 Umbrella sampling simulations

A series of seventeen simulations was used for the construction of the potential of mean force (PMF) along the structural parameter θ, which was restrained around 10, 20, …, 160, and 170° using the harmonic restraining potential with a force constant of 30 kcal/mol/Å^2. Each simulation consisted of 100ps equilibration of the system followed by a 200ps production phase. The structural parameter θ was sampled at every 100 LD steps. The θ-histograms from these trajectories were combined using WHAM to compute the PMF profile along the orientational coordinate. The PMF profile thus obtained is shown in **Figure 46** along with the profile obtained from a 200ps MC/LD trajectory.

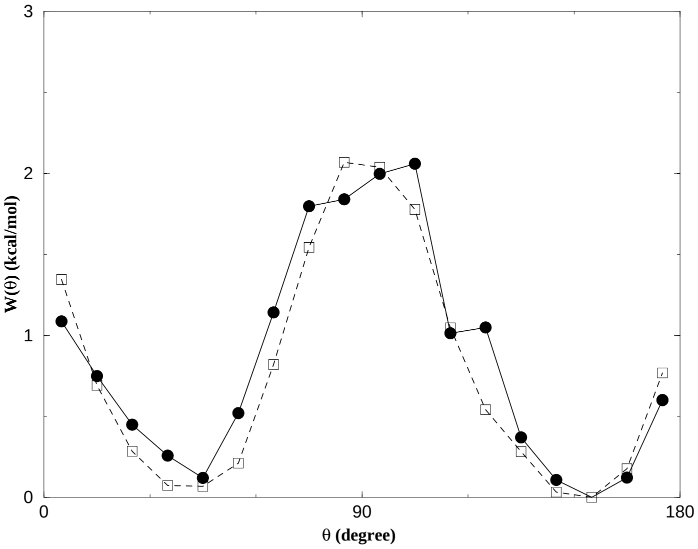


Figure 46. The comparison of free energy surfaces *W(θ)* from the 17 trajectories (dashed line) using umbrella potential combined by WHAM and from a MC/LD trajectory (solid line). 18 grids with 10° interval were used in both methods. The lowest free energy grid centered at 155° was chosen as reference.

Both profiles are in very good agreement. These results validate the approximation that “ghost force” generated the distribution shown in **Eqn. 68**. The MC/LD method yields the PMF profile correctly with about 20 times smaller computational resource than a series of LD trajectories using umbrella sampling technique. This is because the unselected multiple replicas can explore larger space quickly under the scaled potential and MC steps can jump to other local minima without crossing the barrier. Although we got the reasonable PMF profile with the higher free energy states due to their relatively small values (about 2 kcal/mol), it should be emphasized that much higher free energy states are sampled with relatively low resolution even in the MC/LD method since MC/LD trajectory generates the canonical ensemble. To accurately estimate such a higher free energy state, non-Boltzmann sampling methods such as multicanonical ensemble will be required.

The quantitative studies of toluene inclusion complexation of β-CD based on the non-linear regression predicted that the “narrow” orientation is about 0.3kcal/mol more stable than the “wide” orientation,[158,159] however, in our atom-based calculation, the “wide” orientation is about 0.1kcal/mol more stable. Although the relative stability order is different, the small difference (0.5kcal/mol) can be attributed to the inaccuracy of the force field parameters or the regression equation. In their regression, the contribution of hydrophobic interactions of the methyl group favor to occupy the “narrow” orientation, whereas its electron-donating contribution favors the “wide” orientation to make the guest dipole antiparallel to the host dipole. The van der Waals force of the methyl group contributes equally to both orientations. To further compare our result with their regression model, average interaction energy (van der Waals interaction: $\left\langle V_{vdW} \right\rangle$, electrostatic interaction: $\left\langle V_{elect} \right\rangle$) were calculated at each orientation using a 200ps MC/LD trajectory. The average interaction energy differences, $\left\langle V_{vdW}^{wide} \right\rangle - \left\langle V_{vdW}^{narrow} \right\rangle$ and $\left\langle V_{ele}^{wide} \right\rangle - \left\langle V_{ele}^{narrow} \right\rangle$, between the selected guest and the environment atoms are -0.05 and -0.72kcal/mol, respectively. The van der Waals interactions contribute mostly equal to each orientation, which is consistent with the previous regression model. Electrostatic interactions between guest and environment atoms stabilize the “wide” orientation, which also agrees with the regression model. The average interaction energy differences within the selected guest are 0.0kcal/mol for both van der Waals and electrostatic interactions. Calculating the average interaction energy within the environment atoms results in stabilizing the ”narrow” orientation with 0.4 kcal/mol. This 0.4kcal/mol stability at “narrow” orientation comes from the interaction energy within water molecules. Thus, the contribution of hydrophobic interactions of the methyl group at their regression model can be explained partially by the stability of water molecules at the “narrow” orientation. From our investigation using the average interaction energyies the MC/LD trajectory successfully explains the regression model..

7.4 Conclusions

In this chapter, the MC/LD method was demonstrated to yield the canonical ensemble, without being entrapped in local minima as happens with the conventional LD simulation. One MC/LD simulation succeeded in not only detecting the binding orientations of toluene bound to β-CD, but also yielding the correct PMF profile. These results clearly show that MC/LD efficiently explores the free energy surface and not a deformed potential energy surface.

As mentioned above, the MC/LD method can restrict the enhanced sampling region manually such as the guest molecule in this study. But absence of such enhancement over whole system might hinder adequate sampling of the full potential energy surface when the overlap of the interaction energy distributions is small between the selected replica and the unselected replica placed in other local minima at a given coordinates of environment atoms. In such a case, additional intermediate states may be required to get a high acceptance ratio in a MC step.

The MC/LD method will be applied for many purposes such as investigating the binding orientations of toluene in this study. For example, when the ligand has multiple binding modes, the MC/LD method can be predicted the physical properties more accurately. The chemical potential could be calculated more efficiently than by the conventional insertion methods.[190] Furthermore, the converged MC/LD trajectory could also be applied for the reference trajectory used for the free energy estimation methods such as FEP[12] or MM/PBSA. [41,42,44,59] Furthermore, the incorporation of the continuum solvent model[65,67,76,83] will have a great potential in its applications for the protein structure analysis such as the folding studies, the loop search or the conformational search of the side chain.

Chapter 8

Summary and Outlook

In this thesis, we have described the newly developed "free energy based screening methods". These methods may be used either to rapidly identify ligands with the most favorable binding free energy or to estimate specific changes in free energy within a congeneric series. Since λ-dynamics and the related family of methods work based on the binding free energy of the ligands instead of the interaction energy, they provide a more accurate assessment of binding affinity. Species whose binding free energies differ by more than a few kcal/mol from the most favorable binder can be rapidly screened out within a few tens of picosenconds of simulation because they do not compete for interactions with the receptor. The total computation time is not expected to increase with the total number of ligands because only the few favorable binders are able to compete for the $\lambda^2=1$ state. This situation is in contrast to that of conventional free energy calculation methods, where a typical calculation of relative binding free energy between two ligands takes hundreds of picoseconds of simulation time and increases in proportion to the number of ligands. Although the intrinsic problems of the FEP method such as requiring proper overlap of the important configurations, still exists in λ-dynamics-based methods, they can be minimized by using umbrella sampling and/or the iterative procedure with WHAM. Moreover, iterative procedures with WHAM may also be applied to yield quantitative free energy differences for all ligands.

The λ-dynamics-based methods also provide a means to explore the binding orientations and conformations of the ligands much better than does conventional MD. This attractive feature in these methods removes the restriction that the initial orientation of the ligand inside the binding pocket must be close to its true bound orientation in order to get a reasonable estimate of binding free energy - a prerequisite for other free energy calculation methods. Furthermore, the MC/LD method, which is a variant of CMC/MD and specific to overcoming the multiple-minima problem, yields the canonical ensemble without being entrapped in local minima. Thus, MC/LD can be expected to be a useful tool for many purposes such as docking study and loop search.

Free energy based screening methods should fill the gap between empirical methods and theoretically rigorous but computationally intensive methods such as FEP and TI. For example, they can be applied to design a combinatorial library or funnel down the large number of hits detected by the empirical methods. The incorporation of continuum solvent representations such as the generalized Born model into free energy based screening methods accelerates the computational screening process and has a promising future for drug lead optimization and protein design. Given this renewal of effort in “computational alchemy” and the encouraging findings from early studies, we can anticipate that rational free energy based computational approaches to drug and protein design will re-emerge from the tool chest and move to the desktop of the computational medicinal chemist.

Appendix A Consistent change of the Born radius from the surface to the interior

The analytic GB approximation has been known to underestimate the interaction between the interior region atoms as the system increases in size.[65,144] Some modifications of the methods have already been suggested to address this problem.[80,83,191] In a previous study, different parameters were used for molecules of the different size and composition.[83] When we used the parameterized GB model specifically for proteins, MD simulations with the GB energy using the CHARMM 22 all atom force field[138] resulted in continuous expansion of the system for the large molecules like trypsin. This occurred even when we set the hydrogen radii in the param22 parameters to 0.8Å, as suggested in other studies.[83,192] Our analysis of the MD trajectory revealed that the relatively small change in the effective Born radii of the hydrogen atoms on the surface and interior as compared with the heavy atoms gave the expanded states lower overall energies. **Figure 47** shows the effective Born radius as a function of neighbor solute atoms. When the atoms are buried in the interior, the effective Born radius for the heavy atom, whose van der Waals radius is 2.0Å, increases rapidly, whereas for the hydrogen atom the effective Born radii remain small. This inconsistent estimation for the buried atoms' Born radius gives an underestimate of electrostatic interactions related to the buried hydrogen atoms, which always have non-negative partial charges. As a consequence, a repulsive force coming from the interaction among buried negatively charged atoms makes the protein expand continuously. Increasing the van der Waals radii of the hydrogen atoms used for calculation of the effective Born radius alleviated this problem. Even with the larger radii, the correlation of the GB energy with PB energy, calculated by finite difference PB calculations, remained good as shown in **Figure 48**. Since increasing the radii of hydrogen atoms mainly affects the effective Born radii of the deeply buried hydrogen atoms and electrostatic interactions including these hydrogen atoms, the small size molecules that do not have the buried hydrogen, or the atoms on the surface of the large molecules, are influenced little by this modification. Constant temperature (300K) MD simulations with the modified GB energy gave a radius of gyration almost the same as that of the initial X-ray structure in trypsin-benzamidine complex. Using the default setting of the radius resulted in continuous expansion of the protein (see **Figure 49**). Stable trajectories are also obtained in cytochrome c peroxidase and HIV protease by increasing the radii of hydrogen atoms (data not shown). These results also imply that the methodology used to estimate of the effective Born radius may have room for improvement in the case of large molecules.[162]

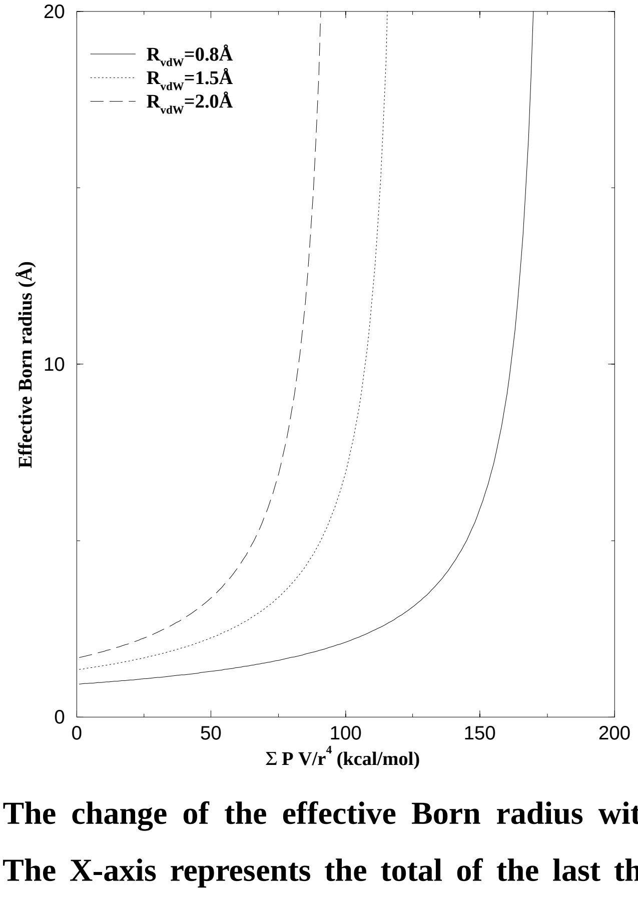


Figure 47. The change of the effective Born radius with the different van der Waals radious (R_{vdW}). The X-axis represents the total of the last three terms ($\Sigma PV/r^4$) in Eqn. 51. $\Sigma PV/r^4$ increases as the atom is buried inside the solute.

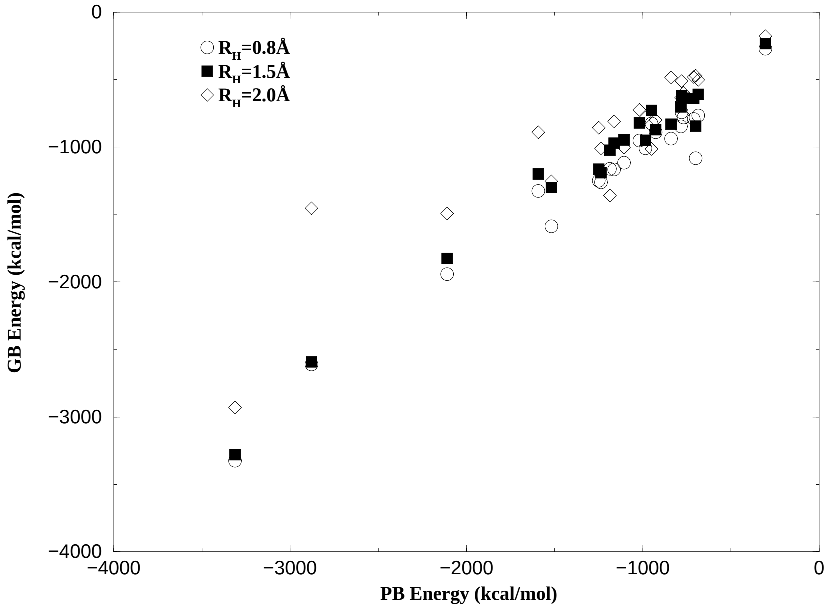


Figure 48. The effect of the van der Waals radii (R_H) of the hydrogen atoms used for the estimation of the effective Born radii, with respect to molecular solvation energies for 22 proteins. PB energies calculated by a finite difference PB method with R_H=0.8Å are taken from Dominy and Brooks.[83] The correlation coefficient between PB and GB energy for R_H=0.8Å, 1.5Å, and 2.0Å are 0.917, 0.943, and 0.715, respectively.

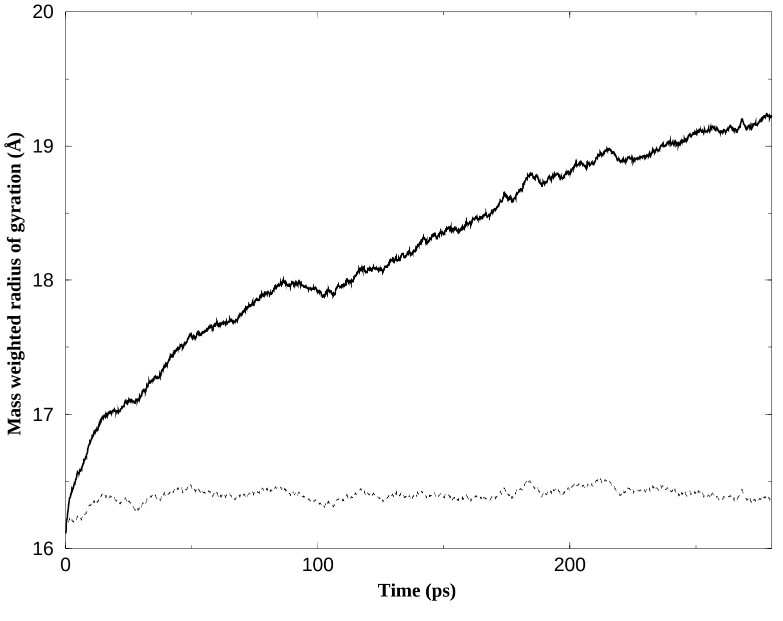


Figure 49. The mass weighted radius of gyration of the trypsin-benzamidine complex. Constant temperature (300 Kelvin) molecular dynamics simulations were carried out with the GB energy from the minimized X-ray crystal structure. All conditions except for the van der Waals radius (R_H) of hydrogen atoms are same in two simulations. The results of R_H = 0.8Å and 1.5Å are shown by the solid line and dashed line, respectively.

Appendix B Refitting the generalized Born parameters

The analytical GB model is known to underestimate the effective Born radii of deeply buried atoms. Onufriev et al. introduced a single parameter to give accurate solvation energies both for small compounds and large macromolecules, however, deeply buried atoms still have some errors in the estimation of pKa shifts. [80] Brian and Brooks used the different values of van der Waals scaling parameter, λ_α, (see **Eqn. 51**) for mono amino acids, dipeptides, and proteins to compensate for the systematic errors in the GB model.[83] For example, refitting λ_α for protein database resulted in λ_α=0.705, which is significantly different from λ_α=0.797 for mono amino acid database. In this study, λ_α is modified to include the influence of molecular size by using switch function:

$$\lambda_\alpha = P_6 + \frac{P_7}{1+N^{P_8}}, \qquad (75)$$

where N represents the number of solute atoms. The switching function is used to maintain a function value of one for isolated atom, and a smooth decay as molecular size is increased. Since the linearized form as shown in **Eqn. 51** was used, λ_α should not reach one when an isolated atom is calculated. The parameters, $P_6 \sim P_8$, were optimized using a systematic search on a grid to minimize the unsigned error between G_{pol} obtained from the GB equation and that obtained from the finite difference PB corrected reaction field against the previously produced composite databases. These databases consisted of 22 globular proteins, 22 nucleic acid strands, 20 amino acids, 210 dipeptides and 15 dinucleotides.[83] The rest of parameters ($P_1 \sim P_5$) were fixed to the values decided previously.[83] The results using optimized parameter sets ($P_6 \sim P_8$) are shown in **Table 12** with those used the previous determined constant van der Waals scaling parameter.[83]

Although number of the parameter is increased, the non-constant λ_α yields smaller unsigned errors over all size of molecules.

Table 12. Average unsigned errors for computed solvation energies of component databases.[a]

database	%error λ_α=0.793	%error $\lambda_\alpha = 1.08 - 0.44\frac{1}{1+N^{-0.25}}$
single amino acids	5.5	5.5
dipeptides	10.5	4.9
proteins	25.8	7.9
single nucleotides	13.4	6.5
dinucleotides	0.6	6.6
nucleic acid strands	20.1	3.0
total	9.7	5.2

[a] $error = \frac{1}{n}\sum_{i=1}^{n}\frac{|GB_i - PB_i|}{|PB_i|}$.

References

(1) J. M. Blaney and J. S. Dixon, *Perspect. Drug Disc. & Design* **1993**, *1*, 301.

(2) I. D. Kuntz, E. C. Meng, and B. K. Shoichet, *Accounts Chem. Res.* **1994**, *27*, 117.

(3) R. Rosenfeld, S. Vajda, and C. DeLisi, *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 677.

(4) D. L. Beveridge and F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *92*, 18.

(5) S. H. Fleischman and C. L. Brooks, III, *Proteins: Structure, Function, and Genetics* **1990**, *7*, 52.

(6) Z. Guo and C. L. Brooks, III, *J. Am. Chem. Soc.* **1998**, *120*, 1920.

(7) J. W. Essex, D. L. Severance, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. B* **1997**, *101*, 9663.

(8) J. Novotny, R. E. Bruccoleri, M. Davis, and K. A. Sharp, *J. Mol. Biol.* **1997**, *268*, 401.

(9) I. D. Kuntz, J. M. Blaney, S. Oatley, R. Langridge, and T. Ferrin, *J. Mol. Biol.* **1982**, *161*, 269.

(10) R. L. DesJarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, *J. Med. Chem.* **1986**, *29*, 2149.

(11) T. Hansson and J. Aqvist, *Protein Engineering* **1995**, *8*, 1137.

(12) R. W. Zwanzig, *J. Chem. Phys.* **1954**, *22*, 1420.

(13) X. Kong and C. L. Brooks, III, *J. Chem. Phys.* **1996**, *105*, 2414.

(14) Z. Guo, C. L. Brooks, III, and X. Kong, *J. Phys. Chem. B* **1998**, *102*, 2032.

(15) J. Pitera and P. A. Kollman, *J. Am. Chem. Soc.* **1998**, *120*, 7557.

(16) M. A. L. Eriksson, J. Pitera, and P. A. Kollman, *J. Med. Chem.* **1999**, *42*, 868.

(17) J. C. Owicki and H. A. Scheraga, *J. Am. Chem. Soc.* **1977**, *99*, 8392.

(18) W. L. Jorgensen and C. Ravimohan, *J. Chem. Phys.* **1985**, *83*, 3050.

(19) P. A. Kollman, *Chem. Rev.* **1993**, *93*, 2395.

(20) U. C. Singh, F. K. Brown, P. A. Bash, and P. A. Kollman, *J. Am. Chem. Soc.* **1987**, *109*, 1607.

(21) B. L. Tembe and J. A. McCammon, *Comput. Chem.* **1984**, *8*, 281.

(22) C. L. Brooks, III and S. H. Fleischman, *J. Am. Chem. Soc.* **1990**, *112*, 3307.

(23) J. J. McDonald and C. L. Brooks, III, *J. Am. Chem. Soc.* **1991**, *113*, 2295.

(24) J. J. McDonald and C. L. Brooks, III, *J. Am. Chem. Soc.* **1992**, *114*, 2062.

(25) L. D. Landau and E. M. Lifschitz *Statistical Physics, 2nd ed.*; Pergamon: New York, 1960.

(26) J. P. Valleau and G. M. Torrie, in *Statistical Mechanics, Part A*, B. J. Berne, Ed.; Plenum Press: New York, 1977.

(27) G. M. Torrie and J. P. Valleau, *J. Comput. Chem.* **1977**, *23*, 187.

(28) D. J. Tobias, S. F. Sneddon, and C. L. Brooks, III, *J. Mol. Biol.* **1990**, *216,* 783.

(29) S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.* **1992**, *13*, 1011.

(30) S. Kumar, P. W. Payne, and M. Vasquez, *J. Comput. Chem.* **1996**, *17*, 1267.

(31) E. M. Boczko and C. L. Brooks, III, *J. Phys. Chem.* **1993**, *97*, 4509.

(32) A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **1988**, *61*, 2635.

(33) A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **1989**, *63*, 1195.

(34) I. Muegge and Y. C. Martin, *J. Med. Chem.* **1999**, *42*, 791.

(35) R. S. DeWitte and E. I. Shakhnovich, *J. Am. Chem. Soc.* **1996**, *118*, 11733.

(36) H.-J. Bohm, *J. Comput.-Aided Mol. Design.* **1994**, *8*, 243.

(37) H.-J. Bohm, *J. Comput.-Aided Mol. Design.* **1994**, *8*, 623.

(38) M. D. Eldridge, C. W. Murray, T. R. Auton, and G. V. Paolini, *J. Comput.-Aided. Mol. Design* **1997**, *11*, 425.

(39) S. Banba and C. L. Brooks III, *J. Chem. Phys.* **2000**, *113*, 3423.

(40) S. Banba, Z. Guo, and C. L. Brooks, III, *J. Phys. Chem. B* **2000**, *104*, 6903.

(41) B. Kuhn and P. A. Kollman, *J. Am. Chem. Soc.* **2000**, *122,* 3909.

(42) I. Massova and P. A. Kollman, *J. Am. Chem. Soc.* **1999**, *121*, 8133.

(43) H. Schaefer, W. F. van Gunsteren, and A. E. Mark, *J. Comput. Chem.* **1999**, *20*, 1604.

(44) J. Srinivasan, T. E. Cheatham, III, P. Cieplak, P. A. Kollman, and D. A. Case, *J. Am. Chem. Soc.* **1998**, *120*, 9401.

(45) R. J. Radmer and P. A. Kollman, *J. Comput. Chem.* **1997**, *18,* 902.

(46) R. J. Radmer and P. A. Kollman, *J. Comput. -aided Mol. Design* **1998**, *12*, 215.

(47) D. A. Pearlman, *J. Med. Chem.* **1999**, *42*, 4313.

(48) J. Aqvist, C. Medina, and J.-E. Samuelsson, *Protein Engng.* **1994**, *7*, 385.

(49) J. Aqvist, *J. Comput. Chem.* **1996**, *17*, 1587.

(50) T. Hansson, J. Marelius, and J. Aqvist, *J. Comput. -aided Design* **1998**, *12*, 27.

(51) J. Marelius, T. Hansson, and J. Aqvist, *J. Comput. Chem.* **1998**, *69*, 77.

(52) H. A. Carlson and W. L. Jorgensen, *J. Phys. Chem.* **1995**, *99,* 10667.

(53) M. L. Lamb, J. Tirado-Rives, and W. L. Jorgensen, *Bioor & Med. Chem.* **1999**, *7,* 851.

(54) J. Wang, R. Dixon, and P. A. Kollman, *Proteins: Structure Function, and Genetics* **1999**, *34,* 69.

(55) H. Liu, A. E. Mark, and W. F. van Gunsteren, *J. Phys. Chem.* **1996**, *100*, 9485.

(56) T. Z. Mordasini and J. A. McCammon, *J. Phys. Chem. B* **2000**, *104*, 360.

(57) J. W. Pitera, N. R. Nunagala, C. C. Wang, and P. A. Kollman, *Biochemistry* **1999**, *38*, 10298.

(58) L. Wang, D. L. Veenstra, R. J. Radmer, and P. A. Kollman, *Proteins: Structure Function, and Genetics* **1998**, *32*, 438.

(59) T. E. Cheatham, III, J. Srinivasan, D. A. Case, and P. A. Kollman, *J. Bioml. Struct. Dyn.* **1998**, *16*, 265.

(60) A.-S. Yang and B. Honig, *J. Mol. Biol.* **1995**, *252*, 351.

(61) K. Osapay, W. S. Young, D. Bashford, C. L. Brooks, III, and D. A. Case, *J. Phys. Chem.* **1996**, *100*, 2698.

(62) J. Shen and J. Wendoloski, *J. Comput. Chem.* **1996**, *17*, 350.

(63) J. Shen and F. A. Quiocho, *J. Comput. Chem.* **1995**, *16*, 445.

(64) M. Schaefer and M. Karplus, *J. Phys. Chem.* **1996**, *100*, 1578.

(65) D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, *J. Phys. Chem. A* **1997**, *101*, 30005.

(66) I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig, *Proteins: Structure Function, and Genetics* **1986**, *1*, 47.

(67) G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, *Chem. Phys. Lett.* **1995**, *246*, 122.

(68) G. D. Hawkins, C. J. Cramer, and C. D. Truhlar, *J. Phys. Chem.* **1996**, *100*, 19824.

(69) S. R. Edinger, C. Cortis, P. S. Shenkin, and R. A. Friesner, *J. Phys. Chem. B* **1997**, *101*, 1190.

(70) B. Tidor, *J. Phys. Chem.* **1993**, *97*, 1069.

(71) J. Ji, T. Cagin, and B. M. Pettitt, *J. Chem. Phys.* **1992**, *96*, 1333.

(72) S. Nosé, *J. Chem. Phys.* **1984**, *81*, 511.

(73) N. R. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087.

(74) C. H. Bennett, *J. Comp. Phys.* **1976**, *22*, 245.

(75) C. Jarque and B. Tidor, *J. Phys. Chem. B* **1997**, *101*, 9362.

(76) W. C. Still, A. Tempczyk, R. C. Hawley, and T. F. Hendrickson, *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(77) B. Jayaram, Y. Liu, and D. L. Beveridge, *J. Chem. Phys.* **1998**, *109*, 1465.

(78) B. Jayaram, D. Sprous, and D. L. Beveridge, *J. Phys. Chem. B* **1998**, *102*, 9571.

(79) R. Luo, M. S. Head, J. Moult, and M. K. Gilson, *J. Am. Chem. Soc.* **1998**, *120*, 6138.

(80) A. Onufriev, D. Bashford, and D. A. Case, *J. Phys. Chem. B* **2000**, *104*, 3712.

(81) X. Zou, Y. Sun, and I. D. Kuntz, *J. Am. Chem. Soc.* **1999**, *121*, 8033.

(82) C. S. Rapp and R. A. Friesner, *Proteins: Structure Function, and Genetics* **1999**, *35*, 173.

(83) B. N. Dominy and C. L. Brooks, III, *J. Phys. Chem. B* **1999**, *103*, 3765.

(84) S. Banba, K. V. Damodaran, and C. L. Brooks, III, *submitted* **2002**.

(85) M. N. Rosenbluth and A. W. Rosenbluth, *J. Chem. Phys.* **1955**, *23*, 356.

(86) B. A. Berg and T. Neuhaus, *Phys. Lett. B* **1991**, *267*, 249.

(87) U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.* **1993**, *14*, 1333.

(88) Y. Okamoto and U. H. E. Hansmann, *J. Phys. Chem.* **1995**, *99*, 11276.

(89) U. H. E. Hansmann, Y. Okamoto, and F. Eisenmenger, *Chem. Phys. Lett.* **1996**, *259*, 321.

(90) N. Nakajima, H. Nakamura, and A. Kidera, *J. Phys. Chem. B* **1997**, *101*, 817.

(91) X. Wu and S. Wang, *J. Phys. Chem. B* **1998**, *102*, 7238.

(92) X. Wu and S. Wang, *J. Chem. Phys.* **1999**, *110*, 9401.

(93) Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.* **1999**, *314*, 141.

(94) N. Nakajima, *Chem. Phys. Lett.* **1998**, *288*, 319.

(95) E. Elber and M. Karplus, *J. Am. Chem. Soc.* **1990**, *112*, 9161.

(96) W. Nowak, R. Czerminski, and R. Elber, *J. Am. Chem. Soc.* **1991**, *113*, 5627.

(97) A. Roitberg and R. Elber, *J. Chem. Phys.* **1991**, *95*, 9277.

(98) C. Simmerling, T. Fox, and P. A. Kollman, *J. Am. Chem. Soc.* **1998**, *120*, 5771.

(99) H. Li and R. Elber, *J. Biol. Chem.* **1993**, *268*, 17908.

(100) A. Ulitsky and R. Elber, *J. Chem. Phys.* **1993**, *98*, 3380.

(101) A. Ulitsky and R. Elber, *J. Phys. Chem.* **1994**, *98*, 1034.

(102) J. E. Straub and M. Karplus, *J. Chem. Phys.* **1991**, *94*, 6737.

(103) S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, *Phys. Lett. B* **1987**, *195*, 216.

(104) H. Senderowitz, F. Guarnieri, and W. C. Still, *J. Am. Chem. Soc.* **1995**, *117*, 8211.

(105) H. Senderowitz and W. C. Still, *J. Phys. Chem. B* **1997**, *101*, 1409.

(106) H. Senderowitz and W. C. Still, *J. Comput. Chem.* **1998**, *19*, 1736.

(107) D. A. Pearlman, *J. Phys. Chem.* **1994**, *98*, 1487.

(108) P. H. Axelsen and D. Li, *J. Comput. Chem.* **1998**, *19*, 1278.

(109) C. Mattos, B. Rasmussen, X. Ding, G. A. Petsko, and D. Ringe, *Nature Struct. Biol.* **1994**, *1*, 55.

(110) G. M. Morris, D. S. Goodsell, R. Huey, and A. J. Olson, *J. Comput.-Adied Mol. Des.* **1996**, *10*, 293.

(111) G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727.

(112) M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470.

(113) B. K. Schoichet, R. M. Stroud, D. V. Santi, I. D. Kuntz, and K. M. Perry, *Science* **1993**, *259*,

1445.

(114) E. Rutenber, E. B. Fauman, R. J. Keenan, S. Fong, P. S. Furth, P. R. Ortiz de Montellano, E. Meng, I. D. Kuntz, D. L. DeCamp, R. Salto, J. R. Rose, C. Craik, and R. M. Stroud, *J. Biol. Chem.* **1993**, *268*, 15343.

(115) M. Vieth, J. D. Hirst, A. Kolinski, and C. L. Brooks III, *J. Comput. Chem.* **1998**, *19*, 1612.

(116) M. Vieth, J. D. Hirst, B. D. Dominy, H. Daigler, and C. L. Brooks III, *J. Comput. Chem.* **1998**, *19*, 1623.

(117) M. M. Fitzgerald, D. E. McRee, M. J. Churchill, and D. B. Goodin, *Biochemistry* **1994**, *33*, 3807.

(118) M. M. Fitzgerald, M. L. Trester, G. M. Jensen, D. E. McRee, and D. B. Goodin, *Protein Sci* **1995**, *4*, 1844.

(119) B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.* **1983**, *4*, 187.

(120) W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.

(121) J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327.

(122) C. L. Brooks, III, A. T. Brünger, and M. Karplus, *Biopolymers* **1985**, *24*, 843.

(123) Z. Liu and B. J. Berne, *J. Chem. Phys.* **1993**, *99*, 6071.

(124) Y. C. Martin, *Perspect. Drug Discovery Des.* **1997**, *7/8*, 159.

(125) A. C. Good and W. G. Richards, *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 321.

(126) M. J. Mitchell and J. A. McCammon, *J. Comput. Chem.* **1991**, *12*, 271.

(127) M. K. Gilson and B. Honig, *Proteins: Structure Function, and Genetics* **1988**, *4*, 7.

(128) B. Honig, K. Sharp, and A.-S. Yang, *J. Phys. Chem.* **1993**, *97*, 1101.

(129) B. Honig and A. Nicholls, *Science* **1995**, *268*, 1144.

(130) H. R. Resat, T. J. Marrone, and J. A. McCammon, *Biophys. J.* **1997**, *72*, 522.

(131) C. L. Brooks, III and M. Karplus, *J. Chem. Phys.* **1983**, *79*, 6312.

(132) C. L. Brooks, III and M. Karplus, *J. Mol. Biol.* **1989**, *208*, 159.

(133) G. King and A. Warshel, *J. Chem. Phys.* **1989**, *91*, 3647.

(134) D. Beglov and R. Benoit, *J. Chem. Phys.* **1994**, *100*, 9050.

(135) D. Beglov and B. Roux, *Biopolymers* **1995**, *35*, 171.

(136) A. H. Juffer and H. J. C. Berendsen, *Mol. Phys.* **1993**, *79*, 623.

(137) V. Lounnas, K. L. Ludemann, and R. C. Wade, *Biophys. Chem.* **1999**, *78*, 157.

(138) A. D. MacKerell, Jr., et al., *J. Phys. Chem. B* **1998**, *102*, 3586.

(139) J. Weiser, P. S. Shenkin, and W. C. Still, *Biopolymers* **1999**, *50*, 373.

(140) J. Weiser, A. A. Weiser, P. S. Shenkin, and W. C. Still, *J. Comput. Chem.* **1998**, *19*, 797.

(141) J. Weiser, P. S. Shenkin, and W. C. Still, *J. Comput. Chem.* **1999**, *20*, 688.

(142) J. Weiser, P. S. Shenkin, and W. C. Still, *J. Comput. Chem.* **1999**, *20*, 586.

(143) T. Simonson and C. L. Brooks, III, *J. Am. Chem. Soc.* **1996**, *118*, 8452.

(144) J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case, *Theor. Chem. Acc.* **1999**, *101*, 426.

(145) F. W. Lichtenthaler and S. Immel, *Liebigs Ann. Chem.* **1996**, 27.

(146) I. Sanemasa and Y. Akamine, *Bull. Chem. Soc. Jpn.* **1987**, *60*, 2059.

(147) I. Sanemasa, T. Takuma, and T. Deguchi, *Bull. Chem. Soc. Jpn.* **1989**, *62*, 3098.

(148) K. A. Connors, *Chem. Rev.* **1997**, *97*, 1325.

(149) W. Saenger, J. Jacob, K. Gessler, T. Steiner, D. Hoffmann, H. Sanbe, K. Koizumi, S. M. Smith, and T. Takaha, *Chem. Rev.* **1998**, *98*, 1787.

(150) K. Fujita, C. W.-H., D.-Q. Yuan, Y. T. K. Nogami, T. Fujioka, K. S. I. Mihashi, and F. W. Lichtenthaler, *Tetrahedron: Asymm.* **1999**, *10*, 1689.

(151) W. Saenger, *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 344.

(152) K. Ukekami, H. F., and T. Irie, *Chem. Rev.* **1998**, *98*, 2045.

(153) K. B. Lipkowitz, *J. Org. Chem.* **1991**, *56*, 6357.

(154) B. Mayer, G. Marconi, G. Klein, G. Kohler, and P. J. Wolschann, *J. Incl. Phenom. Mol. Recog. Chem.* **1997**, *29*, 79.

(155) J. E. H. Koehler, W. Saenger, and W. F. van Gunsteren, *J. Mol. Biol.* **1988**, *203*, 241.

(156) J. E. H. Koehler, W. Saenger, and W. F. van Gunsteren, *Eur. Biophys. J.* **1988**, *203*, 153.

(157) J. E. H. Koehler, W. Saenger, and W. F. van Gunsteren, *Eur. Biophys. J.* **1987**, *15*, 211.

(158) L. Liu and Q.-X. Guo, *J. Phys. Chem. B* **1999**, *103*, 3461.

(159) L. Liu, W.-G. Li, and Q.-X. Guo, *J. Incl. Phenom. and Macro. Chem.* **1999**, *34*, 413.

(160) B. H. Besler, K. M. Merz, Jr., and P. A. Kollman, *J. Comput. Chem.* **1990**, *11*, 431.

(161) K. V. Damodaran, S. Banba, and C. L. Brooks, III, *J. Phys. Chem. B* **2001**, *105*, 9316.

(162) M. S. Lee, F. R. Salsbury, Jr, and C. L. Brooks, III, *J. Chem. Phys.* **2002**, *116*, 10606.

(163) J. B. Jones and G. Desantis, *Acc. Chem. Res.* **1999**, *32*, 99.

(164) M. W. W. Adams and R. M. Kelly, *Chem. Eng. News* **1992**, *73*, 32.

(165) A. E. Eriksson, W. A. Baase, X.-J. Zhang, D. W. Heinza, M. Blaber, E. P. Baldwin, and B. W. Matthews, *Science* **1992**, *255*, 178.

(166) D. E. Otzen and A. R. Fersht, *Biochemistry* **1995**, *34*, 5718.

(167) D. E. Otzen, E. M. Rheinnecker, and A. R. Fersht, *Biochemistry* **1995**, *34*, 13051.

(168) A. D. Robertson and K. P. Murphy, *Chem. Rev.* **1997**, *97*, 1251.

(169) B. D. Dominy, D. Perl, F. X. Schmid, and C. L. Brooks, III, *J. Mol. Biol.* **2002**, *319*, 541.

(170) R. Palma and P. M. G. Curmi, *Protein Sci.* **1999**, *8*, 913.

(171) G. Colombo and K. M. Merz, Jr., *J. Am. Chem. Soc.* **1999**, *121*, 6895.

(172) B. Tidor and M. Karplus, *Biochemistry* **1991**, *30*, 3217.

(173) M. Saito and R. Tanimura, *Chem. Phys. Lett.* **1995**, *236*, 156.

(174) M. Saito, H. Kono, H. Morii, H. Uedaira, T. H. Tahirov, K. Ogata, and A. Sarai, *J. Phys. Chem. B* **2000**, *104*, 3705.

(175) H. Kono, M. Saito, and A. Sarai, *Proteins: Structure Function, and Genetics* **2000**, *38*, 197.

(176) S. Banba, K. V. Damodaran, and C. L. Brooks, III, *J. Chem. Phys.* **2002**, *Submitted*.

(177) J. W. Pitera and P. A. Kollman, *Proteins* **2000**, *41*, 385.

(178) T. Graf, *Curr. Opin. Genet. Dev.* **1992**, *2*, 249.

(179) K. Ogata, C. Kanei-Ishii, M. Sasaki, H. Hatanaka, A. Nagadoi, M. Enari, H. Nakamura, Y. Nishimura, S. Ishii, and A. Sarai, *Nature Struct. BIol.* **1996**, *3*, 178.

(180) H. Morrii, H. Uedaira, K. Ogata, S. Ishii, and A. Sarai, *J. Mol. Biol.* **1999**, *292*, 909.

(181) N. Ota, C. Stroupe, J. M. S. Ferreira-da-Silva, S. A. Shah, M. Mares-Guia, and A. T. Brünger, *Proteins: Structure Function, and Genetics* **1999**, *37*, 641.

(182) J. Lee, *Phys. Rev. Lett.* **1993**, *71*, 211.

(183) K. Fukushima and K. Nemoto, *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.

(184) Y. Sugita, A. Kitao, and Y. Okamoto, *J. Chem. Phys.* **2000**, *113*, 6042.

(185) A. P. Lyubartsev, A. A. Martinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, *J. Chem. Phys.* **1992**, *96*, 1776.

(186) M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.* **1994**, *98*, 4940.

(187) N. Nakajima, J. Higo, A. Kidera, and N. Nakajima, *Chem. Phys. Lett.* **1997**, *278*, 297.

(188) A. Mitsutake, Y. Sugita, and Y. Okamoto, *Biopolymers (Peptide Science)* **2001**, *60*, 96.

(189) M. V. Rekharsky and Y. Inoue, *Chem. Rev.* **1998**, *98*, 1875.

(190) B. Widom, *J. Chem. Phys.* **1963**, *39*, 2802.

(191) A. Ghosh, C. S. Rapp, and R. A. Friesner, *J. Phys. Chem. B* **1998**, *102*, 10983.

(192) D. Mohanty, B. N. Dominy, A. Kolinaki, C. L. Brooks, III, and J. Skolnick, *Proteins: Structure, Function, and Genetics* **1999**, *35*, 447.