# Chapter 1

# Introduction

Despite the increase in the number of studies using molecular dynamics to simulate the dynamic and thermodynamic properties of biomolecules, the effort in improvement of sampling efficiency in the system with rugged energy surface has never been stopped. Our goal of this dissertation is to develop enhanced sampling algorithms and validate these new models for use on biomolecular simulations.

## 1.1 Molecular Dynamics Simulation

Molecular dynamics (MD) is performed by numerically solving the classical Newtonian equations of motions, which can provide detailed picture of how the system evolves with time. In MD simulation, it is also possible to generate any thermodynamic ensembles based on that a variety of thermodynamic averaging properties can be calculated and compared with experimental data[1,2]. Application of MD simulation to biological macromolecules goes back to about three decades ago when McCammon *et. al.* first simulated a protein BPTI in gas phase for 9.2 picoseconds[3]. Since then, along with the considerable progress in X-ray, NMR techniques and computer power, MD simulation has become a very useful tool for understanding the structure and motion of biological systems.

Consider a system consisting of $N$ particles of mass $m_i$ only under the influence of inter-particle interactions, which is specified by an energy function $U(r_1, \ldots, r_N)$, where $r_1, \ldots, r_N$ are the positions of particles. If the forces on the $N$ particles are denoted as $F_1, \ldots, F_N$, then,

$$F_i = -\frac{\partial U(r_1,...,r_N)}{\partial r_i} \tag{1.1}$$

According to Newton's second law, the classic motion of the system is given by,

$$m_i \ddot{r_i} = F_i \tag{1.2}$$

Equations 1.1 and 1.2 therefore completely determine the positions and velocities of the system at time $t$. However, an analytical solution to those equations is almost impossible except in special cases, therefore the MD trajectory is generated by using finite difference method[4].

### 1.1.1 Current Status

One of the current hottest areas of MD research has been and continues to its applications to study properties of biomolecules as reflected by the overwhelming increase in the number of recent publications on this topic. Indeed, the integration of MD simulation and experimental data is now deciphering many complex biological problems such as protein folding[5,6], motor protein function[7], protein channel selectivity[8,9], and enzymatic catalysis[10] etc.

On the other hand, despite the continuing growth of impressive applications, the improvement of force fields and the development of more efficient sampling algorithms remain a particularly active aspect of current MD research. It is no news that the quality of results obtained from MD simulation depends critically on two factors: the energy function must provide an accurate model of the underlying physics of the system and the simulation should adequately sample the important regions of the resulting energy landscape. Numerous studies in this direction with different level of success have been reported[11-16].

### 1.1.2 Look into the Future

Given the striking improvement in simulation methodology and computer power, the next stage of MD simulation will span enormous spaces in terms of both the size and the length of simulation. At one end the simulation will go beyond typical size of thousands or tens of thousands of atoms to even bigger system such as cellular level. Initially even pretty recently this type of studies mainly rely on some coarse-grain models[17,18]. The realistic representation of system and its surroundings in the MD simulation will allow to reproduce more useful details of the system such as, side chain orientation, local electrostatic environment, protonation state etc. At the other the MD simulation duration will be progressing from nanoseconds toward microseconds even milliseconds. The increase in the length of MD simulation will make it possible to generate thermodynamic average quantities with less statistical errors and to directly observe more biological events on much slower time scale.

Another area that will see remarkable growth in the next few years will be MD simulation with a quantum mechanical model to determine the forces as a MD simulation proceeds. As we have already known under many circumstances, such as chemical reaction, ionic solutions or highly coupled system etc, fully quantum mechanical description of the system is necessary. Even though *ab initio* MD has the obvious advantages over empirical force field MD by providing not only classical insight but also many-body effects and properties dependent on electron distribution, such calculation can be extremely time-consuming. An alternative approach based on Car-Parrinello scheme has been proven very useful in many recent studies[19-21].

While MD simulation undoubtedly has been and will be very successful in the future, one shall be cautious in interpretation of the MD results, bearing the limitation of the model in the mind, and compare to any available experimental data if have a chance[1,15].

## 1.2 Force Fields

Force fields are a set of energy functions used to describe the microscopic inter-atomic interactions. It is no doubt that the quality of the MD simulations primarily depends on the accuracy of the force fields. Recent advances in force fields have made it possible to model realistically systems as complex as mouse acetylcholinesterase[22], human water channel aquaporin-1[8], and the mitochondrial membrane protein $F_0F_1$-ATP synthase[7].

Currently, the most common force fields for biomoleular simulations include AMBER[23], CHARMM[24], GROMOS[25] and OPLS-AA[26]. These force fields have been extensively used in many applications and shown in many cases capable of capturing the underlying physics of bimolecular structure and dynamics. These force fields were primarily built and parameterized to match small molecular data, both experimental and theoretical. Parameters were tuned to give accurate fit to quantum mechanical energy barrier, as well as to reproduce enthalpy of vaporization and densities for pure liquids.

The typical force fields used in AMBER take a relatively simple form as follows,

$$\begin{aligned} V(r) &= \sum_{bonds} k_b (b-b_0)^2 + \sum_{angles} k_\theta (\theta-\theta_0)^2 \\ &+ \sum_{torsions} k_\phi [\cos(n\phi+\delta)+1] + \sum_{\substack{nonbond\\pair}} [\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}] \end{aligned} \tag{1.3}$$

The first three summations are over bonds, angles and torsions. The torsion terms also include so-called "improper" torsions, which are used to enforce planarity around $sp^2$ center atoms. The final summation is referred as nonbonded interactions, which describe electrostatic interactions and van der Waals (vdw) interactions via Lennard-Jones 6-12 potentials. The 1-2 and 1-3 interactions are excluded from nonbonded terms while 1-4 electrostatic and vdw interactions are usually scaled by two empirical factors[23].

Although an ever-increasing number of successful applications of current force fields have been reported, the recent protein folding studies revealed some problems of current force fields and indicated that additional improvement is needed[11,14]. One of the particular problems is the over-stabilization of helical structure with AMBER and CHARMM force fields. It is not totally surprising however that there exists some problem in the current force fields since they were traditionally built upon some limited data of small molecules with several major assumptions beyond Born-Oppenheimer approximation, such as fixed-charge, no coupling between different terms of interactions. In fact, after the refinement of charge model using RESP methods in earlier 90's[27,28], the focus of improvement in force fields has recently been shifted to the refitting of torsion parameters. Usually, the torsion parameters were done on the last stage of fitting and were least justified, that leaves room for further improvement. The efforts toward this direction have resulted in several promising applications such as the predictions of the miniprotein Trp-cage structure[13] and the side chain rotamer preference[29].

Furthermore, the importance of many-body and non-classical effects such as polarizability and charge transfer has drawn more and more attention. Over the past decade, dramatic progress has been made in this area[30]. Nevertheless, due to the computational cost, the polarizable force fields have not yet gained popularity in biomolecular simulations. Neither have the extensive comparisons with the current fixed-charge models been compiled in the literature.

## 1.3 Treatment of Solvation Effects

### 1.3.1 Explicit Solvent Model

The choice of solvation models also influences both accuracy and sampling of MD simulations. Computer simulation including explicit treatment of a large number of atomic-scale solvent molecules represents one of the most detailed and realistic approaches to mimic the solvent effects in experiments. The most commonly used water models in current MD simulations are two similar rigid three-site SPC model and TIP3P model[31], which were introduced in early 1980s by Bredensen[32] and Jorgensen[33] respectively. The parameters were obtained by fitting to reproduce the bulk phase structural and thermodynamics properties of liquid water. Recent use of explicit water model together with PME treatment of long-range electrostatic interaction still represents the best though achievable approach for long-time simulation of highly charged DNA and DNA-protein systems[34].

### 1.3.2 Implicit Generalized Born (GB) Model

Although the explicit solvent model has obvious advantages in many aspects, such as satisfactory description of water bridge or water bound motif and implicitly including of hydrophobic effect, the significant computational cost associate with modeling the large number of water molecules makes the explicit MD simulation less efficient. Therefore, a simplified description of solvent effect would be desirable. In most forces fields for biomolecules, the total solvation free energy has been conveniently expressed as a sum of non-polar and electrostatic contributions[35],

$$\Delta G_{sol} = \Delta G_{np} + \Delta G_{elec} \qquad (1.4)$$

where $\Delta G_{np}$ is the free energy change for apolar transfer step, and $\Delta G_{elec}$ is then the electrostatic work of charging the system in solvent versus vacuum. The non-polar part can further be estimated via the solvent accessible surface area of the molecule[36,37],

$$\Delta G_{np} = \gamma(SA) + b \qquad (1.5)$$

Continuum treatment of electrostatic part is based on the Poisson equation or Poisson-Boltzmann (PB) equations with the presence of mobile ions[38]. Although the PB theory has become a standard tool for the investigation of biomolecular electrostatics by solving the PB equation via finite difference or boundary element approaches[39,40], the direct combination of PB model with MD simulation is still not widely used due to its significant computation overhead and the difficulty of calculating energy gradient[41]. Instead, the Born model has been shown to be very efficient to calculate the electrostatic solvation energy of spherical ions in a continuous medium[42]. In 1990, Still *et. al.* extended the Born model by introducing the following formula to approximate the charge-charge interactions in a low dielectric medium[43],

$$\Delta G_{elec} = -\frac{1}{2}\left(\frac{1}{\varepsilon_o} - \frac{1}{\varepsilon_i}\right)\sum_{ij}\frac{q_i q_j}{f_{GB}} \qquad (1.6)$$

where $q_i$ and $q_j$ are atomic partial charges, $\varepsilon_0$, $\varepsilon_1$ are the dielectric constants of the vacuum and solvent respectively, and $f_{GB}$ depends on the effective Born radii $R_i$, $R_j$, and the distance $r_{ij}$ between atoms,

$$f_{GB} = [r_{ij}^2 + R_i R_j \exp(-\frac{r_{ij}^2}{4R_i R_j})]^{1/2} \tag{1.7}$$

As $r_{ij} \rightarrow 0, f_{GB} \rightarrow R_i$, the effective radius that establishes the self-energy of charges that arises from polarization of the surrounding dielectric medium.

When a Debye-Huckel term is incorporated to account for salt effects at low salt concentration, equation 1.6 thus becomes[44,45],

$$\Delta G_{elec} = -\frac{1}{2}(\frac{1}{\varepsilon_o} - \frac{e^{-\kappa f_{GB}}}{\varepsilon_i})\sum_{ij}\frac{q_i q_j}{f_{GB}} \tag{1.8}$$

Even though equation 1.7 is pretty much an empirical relationship, the GB model has been proven capable of capturing the underlying physics of various aspects of the electrostatic contribution to the solvation free energy, yielding reasonable accurate results in comparison to PB calculations. Indeed, due to its relatively simple analytical form, GB model has been increasingly used to study the protein folding[13], DNA stability[46], protein-protein interactions[47] and among many others.

As shown in equation 1.6 - 1.8, other than the function form to account for charge-charge interactions, the accuracy and speed of GB model are very dependent on how the so-called effective Born radii *R* are calculated. In practice, the effective Born radii is calculated from the self-energy, which is the polarization energy of a single point charge surrounded by a high dielectric medium of any geometry. Under the Coulomb field approximation, the self-energy term is given by[48],

$$G_{self}^{Born} \equiv -\frac{q_i^2}{2R_i} \approx \frac{1}{8\pi}[\frac{1}{\varepsilon_i}\int_{V_0}^{V}\frac{q_i^2}{r^4}dV + \frac{1}{\varepsilon_0}\int_{V}^{\infty}\frac{q_i^2}{r^4}dV] \tag{1.9}$$

Originally the integral has been computed by numerical method[43], a pair wise method is subsequently introduced via a summation over pairs of atoms by several groups[48,49]. The first derivatives (forces) and second derivatives of solvation energy can be straightforwardly obtained through the pair wise formula, which makes it suitable for the integration with MD simulation. In current MD simulations with AMBER, the effective Born radii is calculated using the following expression,

$$R_i^{-1} = a_i^{-1} - \sum_j H(r_{ij}, S_j, a_j) \tag{1.10}$$

where effective radii *R* is expressed as a function of the positions and sizes of all atoms in the system, and an additional scaling factor $S_j$ which is an empirical correction first introduced by Hawkins *et. al.* to account for overlaps[50].

## 1.4 Rare Events (Enhanced Sampling) Techniques

Many biologically relevant processes occur on a time scale ranging from microseconds to seconds, which is far beyond the current attainable time scale (several ns duration on 10,000 atoms) provided by MD simulation. Therefore, under these circumstances MD simulation tends to be trapped in or near its initial basin. Due to this reason, the ensemble average properties calculated from the simulation could be totally misleading. Fortunately, a variety of algorithms have been devised to address this difficulty.

### 1.4.1 Generalized Ensemble Method

In the simulation of biomolecules, one is often interested in computing the ensemble average properties such as free energy difference between two states. For these purposes, the exact time dependence is not required. A mean field approximation often referred as Locally Enhanced Sampling (LES) has been proven very useful in smoothing the overall energy surface while keeping the global energy minimum unaltered[51-53]. Another category of methods that has seen a recent increase in use is often referred as generalized ensemble algorithms, including multi-canonical methods[54,55], simulated tempering[56,57] and the replica exchange method (REM)[58,59]. Multi-canonical methods are achieved by replacing the Boltzmann factor exp($-\beta E$) with the multi-canonical probability $n(E)^{-1}$ while the other two methods generally take advantage of higher temperature to accelerate the sampling. An appealing aspect of the generalized ensemble methods is that the canonical ensemble thermodynamics can be recovered over non-canonical samplings.

### 1.4.2 Transition Pathway Sampling

On the other hand, although MD simulation is a good technique to sample the most populated structures in many cases, it is also desired to simulate only a rare event, such as a reaction pathway, giving insights into bimolecular motions on different time scales. To simulate a reaction pathway, one therefore has to manipulate the system and enforce a reactive encounter by using such as a geometric constraint or umbrella potential or by introducing kinetic energy in some translational, rotational or vibrational mode. For instance these methods have been well known as umbrella sampling[60], target MD[61], steered MD[62] and self-guided MD[63] etc. Obviously, this manipulation makes the dynamics of the illustrative reaction pathway less realistic. Whether the found pathway is indeed a representative one can be verified by using the technique known as the transition path sampling (TPS) method developed by Chandler *et. al*[64].

## 1.5 Overview of My Research

In the following section, I briefly summarize four projects included in this dissertation, that mainly focus on developing enhanced sampling techniques and their applications to biomolecular systems.

### 1.5.1 Combined LES with Generalized Born Solvation Model

A strategy was devised that combines the LES technique[51] and GB continuum solvent model to improve the conformational sampling for structure refinement and prediction studies. We applied the resulting method to the simulation of conformational change in an RNA UUCG tetraloop and have shown that the combined GB+LES approach is more efficient than use of either GB or LES alone. We carried out a large number of these simulations and showed in a converged manner that the rate constant for the conformational transition is increased with GB+LES as compared to GB alone. In addition, it was demonstrated that the combined method significantly improves the ability of LES copies to explore independent transition paths as compared to LES simulations with explicit solvation.

### 1.5.2 Energy Barrier Reduction Through LES Approximation

LES method, a mean field approach first introduced by Elber and Karplus[51], reduces the sampling cost of rough energy landscapes by effectively lowering the heights of energy barriers between multiple minima. Even though the strength of this approach has been demonstrated in many studies, there has not been any direct comparison of the kinetic barrier before and after system is modified with multiple copies. In this study, we carried out ensembles of simulations for a conformational transition in an RNA tetraloop, and extracted rate constants for the process with and without LES. Simulations were repeated to obtain rate constants as a function of temperature and the activation energies were obtained from a fit to the data. Therefore, we demonstrate conclusively that the LES method indeed reduces effective energetic barriers for conformational transitions; the transition barrier height was reduced by 74% from 4.6 kcal/mol to 1.2 kcal/mol when a 3-copy LES system was used, which is exactly what Roitberg and Elber had envisioned in their paper[52].

### 1.5.3 New Replica Exchange Techniques

The replica exchange method (REM) has recently been successfully used to study the structure and thermodynamic properties of biomolecules such as peptides and small proteins[58]. For large systems, however, applying REM can be costly since the number of replicas needed increases as the square root of the number of degrees of freedom in the system[65]. Often, enhanced sampling is only needed for a subset of atoms, such as a loop region of a large protein or a small ligand binding to a receptor. For these cases, we derived two variant REM methods, Partial Replica Exchange Method (PREM) and Local Replica Exchange Method (LREM). In both approaches, we assume a weak dependence of the structure of larger region on the instantaneous conformation of the smaller region of interest. The Hamiltonian for the system is then separated, with replica exchange carried out only for terms involving the subsystem of interest while the remainder of the system is maintained at a single temperature. While standard REM simulations are limited by the $f^{1/2}$ increase in number of replicas for $f$ degrees of freedom of the system, our methods permit application to much larger systems with the increase in replicas corresponding to the number of degrees of freedom only in the "focused" region where enhanced sampling is required. These two methods were tested on the loop region of an RNA hairpin model system and it was demonstrated that both methods are able to refine the loop region with dramatic improvement over standard MD approaches. The modified methods are now available in AMBER8 along with our implementation of standard REM.

### 1.5.4 Oxidatively Damaged DNA

One of the most abundant forms of DNA oxidative damage is 8-oxo-7,8-dyhydroguanine(8oxoG)[66]. In the present work, we carried out multiple unrestrained MD simulations of four different DNA 13-mer sequences with G:C, G:A, 8oxoG:C and 8oxoG:A. Our simulation results confirmed the predominance of the normal *anti:anti* form of the 8oxoG:C base pair and the Hoogsteen *syn:anti* form of the 8oxoG:A pair. In the case of 8oxoG:A pair, we observed flipping of the 8oxoG, resulting in a spontaneous *anti*→*syn* transition, in accord with NMR data. 8oxoG:C duplexes were stable in standard Watson-Crick alignment while it adopted a more bended structure as compared with the control structure in the case of G:C pair. In order to gain further insight into the details of this structure transition and local structural fluctuations, we applied our modified REM approaches to the lesion site of the above four DNA systems and obtained probability distributions for alternate base pair conformations for each sequence. The combination of unrestrained dynamics and the thermodynamic data from REM provides new insights into the dynamic behavior of this system and how this behavior is affected by the chemical modifications involved in the oxidative damage.

# Chapter 2

# Improved Conformational Sampling through an Efficient Combination of Mean-Field Simulation Approaches

## 2.1 Introduction

An accurate description of solvation is critical to the success of modeling biological systems[38]. Explicit inclusion of solvent molecules has proven very successful for biomolecular simulations[31,67], particularly when combined with efficient approaches to treat long-range electrostatics, such as particle mesh Ewald (PME)[68].

While explicit solvation may provide an atomic-detail model of solvation, the cost associated with computing forces and integrating equations of motion for the large number of explicit solvent atoms reduces the number of solute conformations that can be evaluated. In addition, frictional and packing effects from solvent may result in a slower time scale for the process of interest, thus requiring calculation of an increased number of time steps to model events of interest using molecular dynamics simulations. Thus, explicit solvation can significantly limit conformational sampling, and obtaining well-converged sampling in explicit solvent is still far from trivial[5,69,70].

An alternative approach to modeling the electrostatic effects of solvation is through a continuum description, such as the generalized Born (GB) model[43,71,72]. One of the key advantages of GB over an explicit solvent model is that it is much more computationally efficient. Only the solute degrees of freedom are considered explicitly, and solvent is approximated as a dielectric medium that influences the behavior of the solute atoms. Furthermore, it has been shown that convergence of biomolecular simulations is accelerated with frictionless implicit solvent models, so that in many cases fewer simulation steps are needed to model a particular transition (as compared to an explicit model).

In simulations reported by Tsui and Case[46], duplex A-form DNA (d(CCAACGTTGG)$_2$) converged to B-form more than 20 times faster in GB than in explicit solvent. Williams and Hall studied the applicability of the GB model to an RNA tetraloop system[73], for which an important structural transition did not occur in standard MD simulations in explicit solvent[74]. With GB solvation, the structural transition became accessible on the nanosecond time scale in otherwise standard MD. These studies strongly imply that molecular dynamics simulation with the GB model can explore phase space much more efficiently than MD with explicit solvent. An additional "bonus" is that the cost per unit simulation time is often reduced because of the smaller system size. Thus, the simulations using GB may extend the effective time scale of processes that we are able to model and permit observation of events that are inaccessible or unaffordable in simulations with explicit solvation, such as the folding of small proteins[13].

While a continuum solvent model may improve sampling in some cases, many barriers to conformational transitions do not arise from the solvent. Locally enhanced sampling (LES) is a mean-field approach that has proven useful in improving sampling through a reduction in internal barrier heights[51,52,75]. LES is effective even when an explicit solvent model is employed[76,77]. The details of the LES approach have been described in detail in the past and further detail is given below. In brief, the LES method provides the opportunity to focus computational resources on the portion of the system of interest by replacing it with multiple copies. The mean-field effect obtained from averaging the interactions among LES copies provides a smoothing effect of the energy landscape[52,53], improving sampling efficiency through reduction of barrier heights. Hornak and Simmerling recently showed that LES could be efficiently used to optimize conformations of proteins loops[78], although in that case a distance-dependent dielectric solvation model was employed since the GB + LES method described here was not available at that time.

Since a major benefit of LES is the ability to simultaneously obtain multiple trajectories for the copied portion of the system, it is desirable to maximize the independence of the replicas during the simulation to increase both the amount of phase space that is sampled and the magnitude of the mean-field smoothing effect. While a major advantage of LES is that it can be successfully employed with an explicit solvent model[76,77], we have observed that the positional variance of solvent-exposed copies is not nearly as large as that obtained during similar simulations in the gas phase. This is likely due to the simultaneous interaction of solvent molecules with all of the copies; copy divergence therefore requires the creation of a larger solvent cavity. This results in a free-energy penalty analogous to the hydrophobic effect and tends to reduce the independence of the copies through an indirect coupling. In addition, the solvent molecules surrounding the group of copies may not be able to simultaneously provide ideal solvation for each of the copies. This issue is discussed in greater detail below.

Since LES and continuum solvation each increase effective transition rates (but through different approaches), we expect that overall sampling with the combined method should exceed that obtained when LES or continuum solvent are employed alone. This combination also permits an approach to solvation that provides greater independence and improved solvation for solvent-exposed copies. We therefore developed a combined GB + LES approach and implemented it in the AMBER suite of programs[79].

Additionally, Cui and Simmerling previously reported that non-LES simulations for a pyrene-substituted DNA system using GB converged to either of two low-energy conformations more rapidly than we observed with explicit solvent[80]. However, transitions between these structures (of similar energy) were not observed even with the continuum solvent. When we employed LES in explicit solvent, interconversions were seen but the simulations were much more computationally demanding because of the explicit solvent. These observations provided additional incentive for the development of the combined GB + LES approach presented here.

As a model system to test this technique, we chose the RNA UUCG tetraloop ($G_1G_2A_3C_4[U_5U_6C_7G_8]G_9U_{10}C_{11}C_{12}$), for which structures have been determined by NMR[81,82]. This makes an excellent model because of its small size and since several previously reported theoretical studies explored the conversion of an incorrect conformation (I) for the loop region into the correct one (C). Standard MD simulation in explicit solvent resulted in no conversion of I to C in several nanoseconds of MD[74]. The use of LES to make multiple copies of the loop region in explicit solvent resulted in reproducible, spontaneous conversion of I to C in about 200 ps[83]. Single-copy GB simulations were also successful in the I→C conversion within about 1200 ps[73]. Thus, each of these approaches to enhanced sampling was successful, and we can compare these results to those from our combined GB + LES algorithm.

The GB + LES simulations that we present converge more rapidly than the single-copy GB or explicitly solvated LES simulations, suggesting that the sampling enhancements provided by these two approaches are complementary. Perhaps more important, however, is the observation that the copies in combined GB + LES are able to sample alternate transition pathways in a single simulation, which was never observed with LES in explicit solvent and is (of course) not possible with standard MD simulations.

## 2.2 Theory

### 2.2.1 Locally Enhanced Sampling

In the current implementation of LES in AMBER, partial charges, Lennard-Jones parameters, bond and angle force constants, and dihedral barriers are all scaled such that the total system energy is equivalent to the average energy of the multiple non-LES "reference" systems. We define a reference system as the single-copy system obtained by combining all of the atoms belonging to one copy with the noncopied atoms. Thus, there are N reference systems for an N-copy LES system. We will refer to these hypothetical reference systems at several points later on, and a specific example for a simple model system is described below. Different copies do not interact with each other during the simulation and interact with the noncopied atoms in an average way. Calculation of energies and forces for this system can be more efficient than the corresponding calculations for all of the separated reference conformations, since interactions involving the noncopied region are only calculated once for all reference conformations when LES is used.

It can be shown that the global energy minimum of the LES system occurs when all copies occupy the position of the global minimum of the non-LES system. This is an extremely valuable property of LES, particularly for structure optimization, and our goal is to maintain this correspondence in our combined GB + LES approach. As described above, the energy function for the LES system is constructed such that the potential energy is the arithmetic average of the energies of the reference systems. As a result, the energies of each copy do not depend on the coordinates of the other copies. Since the copy energies are independent, the global energy minimum of the LES system must occur when each copy is in its own global minimum. In other words, if moving one copy to an alternate location results in lower system energy, then moving the others to the same location would also reduce the energy. The global minimum is therefore a configuration with all copies in identical positions. Since the LES energy is an average of the corresponding single-copy energies, any LES configuration in which the copies have the same coordinates must have the same energy as the corresponding non-LES system. As described above, this is the criterion for constructing the LES energy function. Therefore, the LES global energy minimum is the non-LES configuration with the lowest energy, that is, the non-LES global energy minimum. Other local minima on the original (non-LES) energy landscape have corresponding minima for LES (with all of the copies in the position of the original minimum), but the LES landscape also introduces many new local minima in which copies simultaneously populate different local minima from the non-LES landscape.

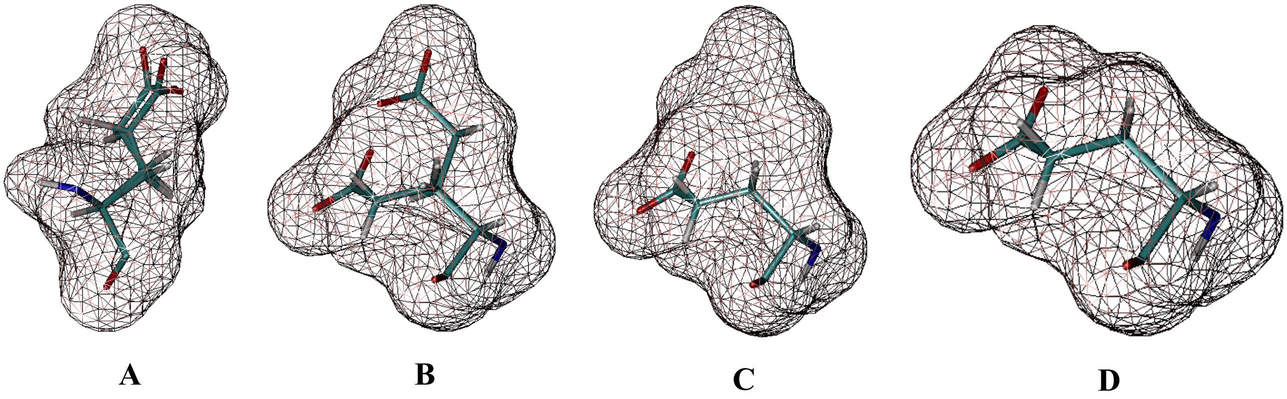


Figure 2.1 A Glu amino acid, shown with solvent-accessible surface area (SASA). (A) Two LES copies of the side chain in explicit solvent (solvent not shown), after equilibration with MD. The penalty associated with creating a larger solvent cavity typically results in only small variance in copy positions. (B) A snapshot of the same system in which the copies are sampling different conformations. The SASA represents the actual solvent cavity when the solvent interacts with both copies. (C) The non-LES reference system corresponding to one of the copies in Figure 2.1B. The SASA is that seen by the copy during LES simulation in explicit solvent, but the larger cavity represents a low-probability solvent configuration at normal T and P. (D). The desired reference system from Figure 2.1C, with correct SASA. The solvent configuration clearly differs from that sampled by the LES system.

An advantage to LES compared to many approaches to improved sampling is that it can be employed with explicit inclusion of solvent molecules[50]. However, when solvent-exposed copies occupy similar but non-identical positions, explicit solvent is excluded from a volume corresponding to the region occupied by any of the copies. As a result, none of the copies samples a truly solvent-exposed state (Figure 2.1). In this case, LES may give efficient sampling of states of the reference system that are possible, but of lower probability than those with the copies in direct contact with solvent. This effect is inversely correlated to that described above; as the copies become more independent, the instantaneous solvation becomes less representative of what would be expected for a non-LES system exploring the same conformations. In the limit of 0K, the problem may disappear if the copies converge to identical positions, but at finite temperatures the thermal fluctuations typically result in varying degrees of this undesirable behavior.

Thus, the LES system with explicit solvent corresponds to a set of reference systems that are not ideally solvated. If LES is combined with a continuum solvent model, it is possible to independently solvate each of the reference systems. Our strategy in deriving this combination is to maintain the correspondence of the LES system energy and the average energy of the reference systems, while providing a more realistic representation of solvation for the individual copies.

### 2.2.2 GB + LES: Difficulties with the Effective Born Radii

In the present study, we will combine the LES approach with the GB method for calculation of the electrostatic component of solvation free energy. The detail of the GB approximation has been given in equation 1.6 – 1.8 of the previous chapter. The main challenge in combining GB and LES arises when calculating the effective Born radii for each atom. In AMBER (without LES), the effective Born radii $\alpha_i$ are calculated via the pairwise descreening approximation[50],

$$\alpha_i^{-1} = \rho_i^{-1} - \frac{1}{2}\sum_j \int_{L_{ij}}^{U_{ij}} dr\left(\frac{1}{r^2} - \frac{r_{ij}}{2r^3} - \frac{1}{2r_{ij}r} + \frac{S_{ij}^2\rho_j^2}{2r_{ij}r^3}\right)$$

$$= \rho_i^{-1} - \frac{1}{2}\sum_j \left[\frac{1}{L_{ij}} - \frac{1}{U_{ij}} + \frac{r_{ij}}{4}\left(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}\right) + \frac{1}{2r_{ij}}\ln\frac{L_{ij}}{U_{ij}} + \frac{S_{ij}^2\rho_j^2}{4r_{ij}}\left(\frac{1}{L_{ij}^2} - \frac{1}{U_{ij}^2}\right)\right] \qquad (2.1)$$

where

$$L_{ij} = \begin{cases} 1 & if \quad r_{ij} + s_{ij}\rho_j \le \rho_i \\ \rho_i & if \quad r_{ij} - s_{ij}\rho_j \le \rho_i < r_{ij} + s_{ij}\rho_j \\ r_{ij} - s_{ij}\rho_j & if \quad \rho_i < r_{ij} - s_{ij}\rho_j \end{cases}$$

and $$U_{ij} = \begin{cases} 1 & if \;\; r_{ij} + s_{ij}\rho_j \le \rho_i \\ r_{ij} + s_{ij}\rho_{ij} & if \;\; \rho_i < r_{ij} + s_{ij}\rho_j \end{cases}$$

In GB, the effective radius of an atom, and therefore the interaction between any pair of atoms, is no longer independent of the coordinates of the rest of the system because of the descreening effects of the other atoms. This is the cause of problems employing the GB model with frozen atoms[84]. If we take the approach that each noncopied atom should be simultaneously descreened by all of the LES copies, the corresponding reference systems would have the individual copies occupying a solvation shell corresponding to the space occupied by all copies-directly analogous to the situation with LES in explicit solvent. Each reference solute would not be fully solvated (Figure 2.1) and would therefore represent less probable, though possible, configurations of the reference systems. This is not the ideal average as this partial desolvation of the copies is one of the problems encountered with LES and explicit solvent that we wished to avoid. Since different LES copies can also occupy the same space, the pairwise descreening approach[50] that is used in AMBER to calculate effective Born radii would not be reasonable since the descreening effects due to the multiple copies would not necessarily be additive because of allowed overlap between different copies.

An additional problem with this approach is that moving any copy would potentially change the effective radii of all atoms and therefore directly affect the energetics of interactions in other copies. Thus, our proof of equivalence of global minima presented above would no longer remain valid. This problem does not arise with explicit solvation, since the nonbonded energy can still be separated into a sum of pairwise terms and thus copy independence is maintained. Even if an alternate formalism were employed in which this coupling was not present, the effective solvation of the system would incorporate the same inaccuracies as encountered with explicit solvent, that is, the effective solvation cavity would surround the set of copies rather than represent individual solvation of each reference system.

We therefore take the approach of explicitly enforcing the correspondence with the average of the reference systems that the LES system represents, including correct reference solvation of each copy. We first imagine separating the copies, combining each with the noncopied region, and calculating effective radii in each resulting system. In this case, every atom in each reference system will require a unique effective radius as compared to the same atom in the other systems: the LES copies of an atom can occupy different positions in space and therefore need different radii, but the non-LES atoms also require multiple radii, one for each of the hypothetical systems, since the descreening of these atoms will differ among reference systems because of the changes in the positions of LES atoms. In the actual simulation, we do not explicitly separate these hypothetical systems, but the interactions involving any pair of non-LES atoms are no longer identical in the reference systems, and obtaining the correct average over the reference systems requires explicit calculation of these interactions multiple times using each of the sets of effective radii.

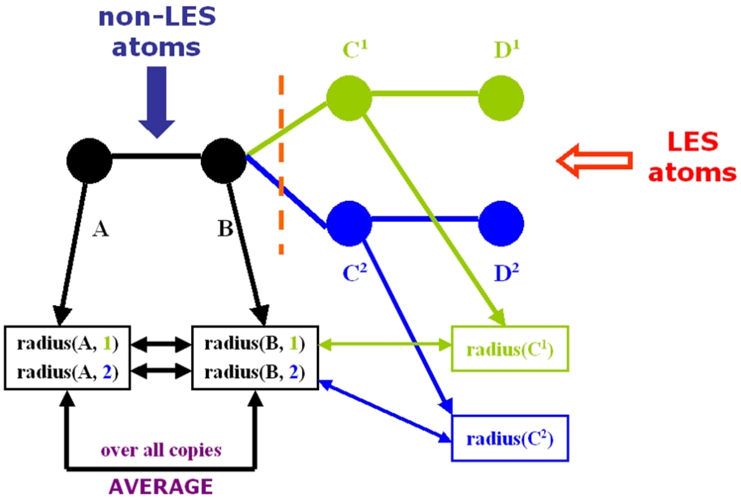


Figure 2.2 A simple model system to clarify the calculation and use of effective radii. In a system with four atoms A, B, C, and D, atoms C and D are replaced with two LES copies. The radii are described in the text.

To clarify this issue, we describe a simple model system with four atoms, A, B, C, and D (Figure 2.2). We replace atoms C and D with two copies, $C^1/C^2$ and $D^1/D^2$ where the superscript denotes the copy number. Since different copies of a region do not interact, $C^1$ does not interact with $C^2$ or $D^2$, and so on. The energy of this system is an average of the energies of the two reference conformations represented by LES: $ABC^1D^1$ and $ABC^2D^2$. In standard LES, the interaction between A and B does not depend on the coordinates of C and D, and this interaction is factored out of the average and calculated only once. Thus, the form of the LES averaging in this case is given by the following equations:

$$V_{real} = V(A,B,C,D) = V_{non-LES}(A,B) + V_{non-LES,LES}(A,B,C,D) + V_{LES}(C,D) \quad (2.2)$$

$$\begin{aligned} V_{LES} = V_{non-LES}(A,B) + \frac{1}{2}[V_{non-LES,LES}(A,B,C^1,D^1) + V_{LES}(C^1,D^1) + \\ V_{non-LES,LES}(A,B,C^2,D^2) + V_{LES}(C^2,D^2)] \end{aligned} \qquad (2.3)$$

where, $V = \Sigma\,(E_{bond} + E_{angle} + E_{tor} + E_{ele} + E_{vdw} + E_{GB})$

The first step in a GB calculation is to determine effective Born radii for all atoms. As stated above, both LES and non-LES atoms require a separate effective radius for each of the reference systems. For example, radii radius(A,1), radius(B,1), radius(C,1), and radius(D,1) are calculated within reference conformation 1 {ABC$^1$D$^1$} and radius(A,2), radius(B,2), radius(C,2), and radius(D,2) are calculated within conformation 2 {ABC$^2$D$^2$}. For this calculation, the increase in effort required is smaller than the number of copies since some duplicate calculations can be avoided. For example, the descreening contribution of atom A because of B is the same in both systems because all of the systems have this same atom pair at the same separation distance. For small LES regions, this can result in large improvements to the efficiency of calculation of radii (analogous to factoring of duplicated nonbonded interactions in traditional LES). Not all descreening contributions are the same; the descreening of atom A because of C differs in the two systems since C has different coordinates.

Calculation of the atomic forces due to the GB solvation with LES proceeds in an analogous manner to non-LES GB. For interactions between an atom in LES copy i and a non-LES atom, the energy and forces are calculated using the single radius for the LES atom and radius i for the non-LES atom. Only one calculation is performed, and the difference due to LES is just the selection of the radius for the non-LES atom that was calculated using the descreening from the corresponding LES copy. For example, the interaction between A and C1 is calculated using the first copy of the effective radius for atom A, since atom C1 belongs to LES copy 1. For the interaction of A and B, we calculate the interaction twice: once with radii set 1 and again with radii set 2, each corresponding to alternate conformations of the LES region that affect the screening of this interaction. The computational effort required for this part of the calculation (the pairwise electrostatics) increases by a factor of the number of LES copies employed. This is in contrast to the calculation of the multiple sets of effective radii, which may not require significantly more effort since most of the contributions to the descreening were shared among the sets.

In the limit of a very small LES region, the number of effective radii and nonbonded interactions that need to be calculated will increase by a factor of the number of reference conformations represented by the LES copies. This differs dramatically from the non-GB LES approach we implemented in AMBER, in which interactions in the noncopied region were identical among all of the reference systems and were therefore factored out of the average and calculated only once. This factoring of duplicated terms (along with the mean-field effect) is the source of the computational efficiency of using LES as compared to multiple simulations of the entire system, especially when the LES region is a small fraction of the system.

This procedure provides an exact average of the energies and forces of the separated and properly solvated reference systems but is less computationally efficient. However, the mean-field effect of LES is still obtained, and dynamics on the smoother energy landscape may provide greater efficiency than multiple individual trajectories. In addition, we present reasonable approximations below that avoid the majority of this computational overhead.

### 2.2.3 Approximations

The forces and energies calculated using the approach described above represent an exact average of the reference systems, and no approximations have been introduced beyond those already inherent in LES and GB. In this case, the pairwise interactions using GB + LES with N copies take N times as long as a corresponding non-LES calculation. For each pairwise interaction in the original system, there exist N times as many interactions in the LES system: N interactions involving the N copies of each LES atom, and N components in the average for each fully non-LES pair. The calculation of effective Born radii can be simplified, since the fully non-LES pairwise descreening contributions are identical in all reference systems and can be shared when calculating the multiple radii of the non-LES atoms. Other drawbacks to the introduction of multiple effective radii include larger memory requirements and increased communication overhead in parallel implementations, especially important when using PC clusters with low network bandwidth. We therefore investigated approximations that would retain a high level of accuracy while reducing the number of effective radii that were required.

The source of the need for multiple radii for non-LES atoms is that the descreening effects arising from the LES region may differ for different copies, since the copy conformations may differ. When the atoms are far from the LES region, this effect is reduced, and the variation in effective Born radii due to the change in copy conformation is usually very small. This suggests that we can ignore the differential descreening effects of copied atoms at long distances, and the calculation of the radii would thus be faster.

This type of neglect of the effect of conformational changes for atoms at long distances is the essence of the reported combination of GB with frozen atom approximation[84], in which atoms far from the moving region do not have their radii updated each step. In the present case, however, all of the atoms are moving and thus need to have their radii recalculated regardless of whether LES is used. We therefore do not use a distance cutoff for calculation of the effective radii. It is important to keep in mind where the extra work is needed: not in the calculation of the effective radii (since most terms involve fully non-LES pairs and therefore introduce no extra effort as compared to standard GB), but in using them. The key challenge is to reduce the number of non-LES atoms that need multiple radii to avoid the explicit enumeration of all elements in the average interaction with other non-LES atoms.

We thus introduce a cutoff for the permissible deviation among the multiple radii of a given atom. We approximate that if the radii are similar, the average of the N interactions using the sets of radii can be approximated by one of the elements in the average. Therefore, in our implementation a threshold value was introduced to reduce the calculation time. When the differences among the N copies of effective radii of a non-LES atom are less than this threshold, it is reasonable to use a single effective radius. As a consequence, N pairs of calculations involving such atom pairs are reduced to one per pair, a factor of N speedup. Moreover, using one effective radius saves memory and reduces communication overhead since there are fewer effective radii to distribute among multiple computing nodes.

To test the feasibility of this approximation, we calculated the energy and force errors using various radii difference threshold (RDT) values. The test involves calculation of the atomic energies and forces for the LES system, and comparing these data to an exact value obtained from explicitly separating and averaging the N corresponding non-LES reference conformations. This test also indicates the relative computational efficiency compared to multiple non-LES simulations.

Since the non-LES region in our small RNA tetraloop model system is small, only a relatively small percentage of atoms are not copied and therefore most atoms are affected by differences in the conformations of the LES copies. The gain in efficiency due to the cutoff is therefore not as apparent as might be observed in a typical, larger system with a smaller LES fraction. We thus carried out the efficiency tests for *triose phosphate isomerase* (TIM), a protein composed of 247 residues. Five LES copies of loop 6 (the active site lid, residues 165-175) atoms were employed, and short GB + LES MD simulation was carried out to obtain a configuration with non-identical coordinates for the LES atoms.

Table 2.1 Comparison of the Effects of Various RDT values

| RDT (Å) | $N_{RDT}/N_{tot}$[a] | $\Delta E$[b] (kcal·mol$^{-1}$) | RMSDf[c] (kcal·mol$^{-1}$·Å$^{-1}$) | $\Delta F_{max}$[d] (kcal·mol$^{-1}$·Å$^{-1}$) | Speedup[e] |
|---|---|---|---|---|---|
| 0.1 | 0.99 | 0.70 | 0.0069 | 0.131 | 2.89 |
| 0.01 | 0.94 | 0.12 | 0.0013 | 0.022 | 2.46 |
| 0.001 | 0.78 | 0.03 | 0.0002 | 0.003 | 1.71 |

[a] $N_{RDT}/N_{tot}$ indictes the fraction of non-LES atoms that can use a single effective radius instead of 5; [b] $\Delta E$ is the unsigned energy difference between calculations with/without RDT; [c] RMSDf is the root mean square deviation of forces between calculations with/without RDT; [d] $\Delta F_{max}$ is the maximum difference in atomic force components between calculations with/without RDT; [e] Speedup is the time required for evaluation of nonbonded energy and forces without RDT divided by the time with RDT.

In Table 2.1 we provide the results of energy and force errors and relative performance using LES + GB, with various RDT values, for five copies of the active site loop in TIM. We observed that the energy and force errors are generally small enough for stable multi-ns MD simulation when the RDT is set to 0.01 Å (which means that the average effective radius for a non-LES atom is used if the difference among the multiple radii is less than 0.01 Å). In this case, ~94% of the atoms use a single effective radius, and the calculation of nonbonded interactions is nearly 3 times faster than without use of the RDT approximation (the speedup is not a factor of 5 since LES adds additional overhead besides that reduced by the RDT). Thus, it appears to be a reasonable approach, since LES is already an approximate method and the "exact" effective Born radii calculated through pairwise approximation are inherently imperfect and become the dominant source of error for solvation free-energy calculations[85].

If one desires a more accurate trajectory, an RDT value of 0.001 Å results in an energy difference (compared to the exact calculation) of ~0.03 kcal/mol, and average atomic force deviations less than 0.0002 kcal·mol$^{-1}$·Å$^{-1}$. Even with this small RDT value, nearly 80% of non-LES atoms employ a single effective radius, and the calculation requires less than 60% of the time required without the RDT approximation.

Even with this approximation, the calculation using GB + LES is somewhat more computationally intensive and requires more memory than the non-LES calculations. However, we show below that the increase in efficiency of the LES simulations is much greater than this additional expense. The simulations not only converge more rapidly than corresponding non-LES simulations, but also show multiple transition pathways and time scales in single simulations, thus providing an improvement over non-LES approaches and LES simulations in explicit solvent.

## 2.3 Simulation Details

### 2.3.1 System setup

Simulations were carried out using AMBER6 package with the modification to include our more rigorous GB+LES algorithm. The original incorrect and correct RNA tetraloop NMR models were used as starting structures[81,82]. The AMBER module ADDLES was used to construct the LES systems for simulation. All LES copies of individual atoms were initially assigned identical coordinates but unique velocities, and therefore diverged with propagation of time. The time step was 1 fs, and SHAKE was applied to all bonds involving hydrogen[86]. No nonbonded cutoff was used. The AMBER ff94 force field[23] was used in all calculations with either the GB continuum model or a simple distance dependent (1/r) dielectric treatment, as noted when the simulation is described. The Born radii were adopted from Bondi with modification of hydrogen[46], and the scaling factors for Born radii were taken from the tinker modeling package[87]. An offset of 0.09 Å was used for the radii.

The crystal structure of TIM (PDB code 1YPI) was used for single-point energy and force calculations[88]. The same set of GB parameters was used for TIM as for the RNA tetraloop. To compare the effects of different RDT values on resulting forces and energies, we generated a TIM structure with five loop copies with loop region RMSD about 1.0 Å from each other. Of the 3778 atoms in the original TIM system, 150 loop atoms were replaced by five copies. The resulting LES system was composed of 3628 non-LES atoms and 750 LES atoms.

As a test of the program code, we investigated the accuracy of the average energies and forces calculated for the LES copies. First, we calculated energy and forces for a LES system in which the coordinates of each copy differed, and then divided the LES copies into the five reference (non-LES) systems, each using the same coordinates for the non-LES region. Energies and atomic forces were calculated for each system, and the averages of these values were compared to those obtained directly for the LES system. The values from each of these approaches to the averaging were identical, suggesting that the code was robust. This also confirms that the LES calculation is properly accounting for individual solvation of each copy, one of our main goals in the development of the algorithm.

Since the combination of GB and LES is nontrivial, we also tested the behavior of a distance-dependent dielectric treatment of solvation since this approach is both straightforward and efficient. However, simulations starting from incorrect and correct structures for the RNA tetraloop did not show behavior comparable to those observed in otherwise identical GB simulations. The correct structure was unstable and the incorrect structure did not convert to the correct structure. In both cases, RMSD values compared to the C conformation were ~3-4 Å, and the LES copies did not converge to a single conformation. These findings are consistent with similar instability for the tetraloop reported by Hall et al. for single-copy simulations with a distance-dependent dielectric.

### 2.3.2 Evaluating Convergence

The force fluctuation metric introduced by Thirumalai *et. al.* was used as a rigorous measure of the rate of sampling[89-91]. Following their definition, the average force on the *i*th atom for the fluctuation metric is defined as

$$f_i^a(t) = \frac{1}{t}\int_0^t dsF_i^a(s) \tag{2.4}$$

where *a* indicates that the average is calculated over the *a*th trajectory. Given two independent trajectories, the time-dependent force metric can be defined as the difference between the averages calculated over the N frames of the pair of trajectories *a* and *b*:

$$d(t) = \frac{1}{N}\sum_{i=1}^{N} | f_i^a(t) - f_i^b(t) |^2 \tag{2.5}$$

The AMBER program was modified to report these values during standard GB and GB + LES simulations.

Secondly, the approximate I→C transition rate is also used as a measure of sampling efficiency. The calculation of rate constant is introduced in the following section.

### 2.3.3 Calculation of Rate Constant and Energy Barrier

Rates for the I→C transition were obtained by collecting first passage times for an ensemble of simulations initiated from the incorrect conformation. Random number seeds were varied among the simulations to provide different initial velocity distribution and divergent behavior. The time dependence of the fraction of the ensemble that remained in the incorrect conformation was fit to a single exponential, assuming first order kinetic behavior. By collecting first passage times, this procedure directly provides forward rate constants, rather than the sum of forward and reverse rate constants that are obtained from a traditional experiment. A loop region heavy atom RMSD value of 1.0 Å from the correct NMR structure was used as the threshold for determination of the transition event.

### 2.3.4 MM-GB Calculation

The relative free energy of the two alternate conformations of the RNA tetraloop was estimated through the MM-GB (molecular mechanic energy + GB solvation) method[92,93]. For each conformation, 2000 equally spaced snapshots were collected from 2ns explicit solvent simulation. The MM energy was calculated as the average of the sum of all bonded and non-bonded interactions using the ff94 force fields. The electrostatic contribution to the solvation free energy of each conformation was calculated with generalized Born model. The solvent-accessible surface area term was neglected since the two conformers have almost identical surface areas.

## 2.4 Results and Discussions

The major difference between the two experimental structures[81,82] of the RNA tetraloop is the hydrogen bond pattern between the bases in residues U5 and G8 (Figure 2.3). As described above, previous standard MD simulations of this RNA tetraloop in explicit solvent with PME did not result in interconversion between the two structures. When LES was employed in explicit solvent, the correct conformation was stable, but the incorrect underwent a rapid transition to the correct form within 200 ps. Transition from the incorrect to correct structure was also observed in about 1200 ps using the GB implicit solvent model without LES. We therefore concluded that both GB model and LES sampling method were able to enhance the sampling of phase space in RNA tetraloop simulations.

We investigated whether the combined GB + LES approach was able to provide stable simulations of the correct structure under conditions that also resulted in spontaneous conversion of incorrect to correct structure. In addition, we investigated whether any advantage is gained by using GB + LES as compared to GB alone. In other words, are the enhancements provided by GB and LES complementary?

The results of all simulations with various temperature and solvent models from both I and C structures are summarized in Table 2.2. All root-mean-square deviation (RMSD) calculations include non-hydrogen atoms in the UUCG tetraloop (residues 5-8) except the base atoms of U6, which does not form specific contacts and shows higher mobility. This RMSD selection was chosen to be consistent with the LES/PME study of this system[83]. The results of each simulation are described in further detail below.

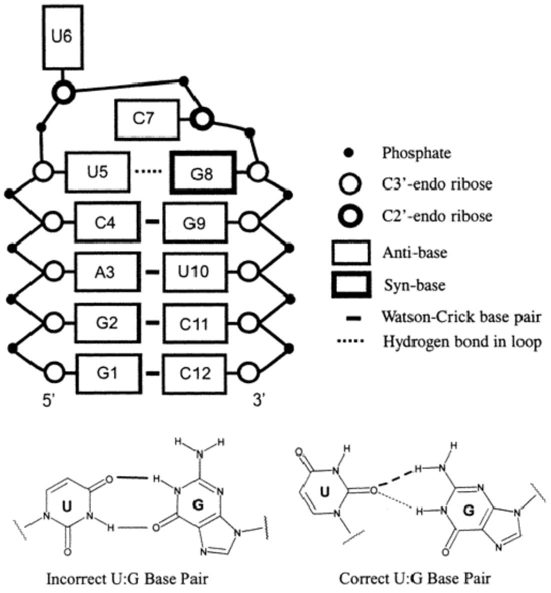


Figure 2.3 A schematic diagram of the topology of the RNA tetraloop system being studied. The hydrogen bond patterns for the U5:G8 base pair in the incorrect and correct NMR structures are shown in the lower figure. Solid lines are used for the incorrect hydrogen bonds, and dashed lines are used for the correct ones.

Table 2.2 Summary of Results Obtained from Various Simulations for RNA tetraloop[a]

| | # LES copies | Temp (K) | # total atoms | Starting structure | Time of I→C transition (ps) | Final RMSD to C |
|---|---|---|---|---|---|---|
| PME+LES[b] | 5 | 300 | 7358 | I | 200 | 0.8 |
| GB non-LESa | 1 | 300 | 382 | C | n/a | 0.9 |
| GB non-LESb | 1 | 300 | 382 | I | 1100 | 1.0 |
| GB non-LESc | 1 | 300 | 382 | I | 1600 | 1.0 |
| GB non-LESd | 1 | 300 | 382 | I | 200 | 1.0 |
| GB+LESa | 3 | 200 | 632 | C | n/a | 1.0 |
| GB+LESb | 3 | 200 | 632 | I | 40-370 | 1.0 |
| GB+LESc | 3 | 200 | 632 | I | 160 | 1.0 |
| GB+LESd | 3 | 130 | 632 | I | 470-1340 | 1.0 |
| GB+LESe | 3 | 100 | 632 | I | n/a | 3.0 |

[a] The details of each simulation are described in the main text. Transition times for LES simulations are given as a range when the different copies showed significantly different transition times. [b] Reference[83].

We first made the same choice as Simmerling *et. al.* did in the previous LES/PME study of this RNA tetraloop[83] and used five LES copies of the entire UUCG loop. Each of these copies was attached to the stem, and the stem interacted with these copies in an average way. We initiated a simulation at 300 K from the correct structure, with all LES copies having identical initial coordinates. This simulation resulted in fully extended conformation of RNA within 400 ps. The heavy atom RMSD rose to 4.5 Å and all base pairs were lost during the simulation. Similarly undesirable results were obtained when starting from the incorrect conformation. We hypothesized that this might arise from too great a weakening of the Watson-Crick hydrogen bonds because of the scaling of partial charges and Lennard-Jones well depth parameters. It has also been shown that the behavior of the LES system corresponds to a non-LES system of higher temperature[94]. Moreover, lack of solvent friction likely makes the dynamic behavior of the RNA more sensitive than that with LES in explicit solvent.

We therefore empirically reduced the temperature and number of copies that were used and evaluated the dynamics of the correct structure. We found that three LES copies at 200K (GB + LESa in Table 2.2 and right graph in Figure 2.4) resulted in a stable simulation with similar fluctuations to those observed in non-LES GB simulation at 300K (GB non-LESa and left graph in Figure 2.4). In both cases, the structure was stable and the loop RMSD values fluctuated about 1 Å. We also obtained stable trajectories of the correct conformation with three LES copies at temperatures of 150K and 130K (data not shown). The use of LES has therefore not affected the ability of the simulation to maintain a stable correct conformation (although with a reduced temperature).

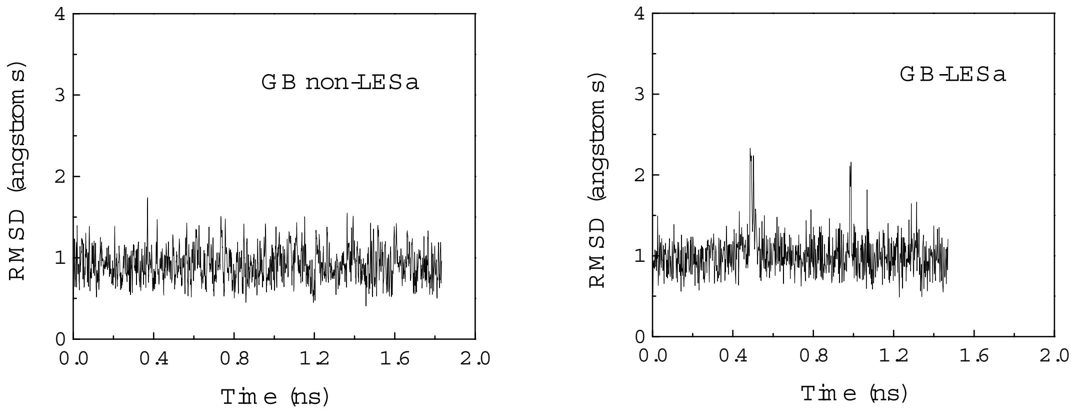


Figure 2.4 The loop RMSD as a function of time for simulations starting with the correct conformation. On the left is the GB simulation (GB non-LESa in Table 2.2); on the right is the GB + LES simulation (GB-LESa in Table 2.2). Only average LES RMSD is displayed for clarity.

This scaling of temperature makes the use of GB + LES (and LES in general, although GB + LES appears to be particularly sensitive) somewhat more complex than standard MD. Others have pointed out this difficulty with LES, and we do not address this aspect of the LES method here; rather, we are interested in combination of the method with an efficient solvation model. As a general guide to evaluating temperature when predicting unknown structure, we expect that the LES copies should converge to a single conformation regardless of initial structure. When optimization protocols such as simulated annealing are used, multiple predictions should be compared to ensure that the final structures are not the result of kinetic trapping. One should always demonstrate insensitivity of results to the initial coordinates for well-converged simulations of any type. With LES, we can compare the predictions being provided by each LES copy as the simulation proceeds. This provides us with continuous measures of the precision of the prediction by exploiting the built-in convergence test available with LES, especially when using alternate initial conformations for the copies. One can thus obtain conformational "error bars" in a single LES simulation.

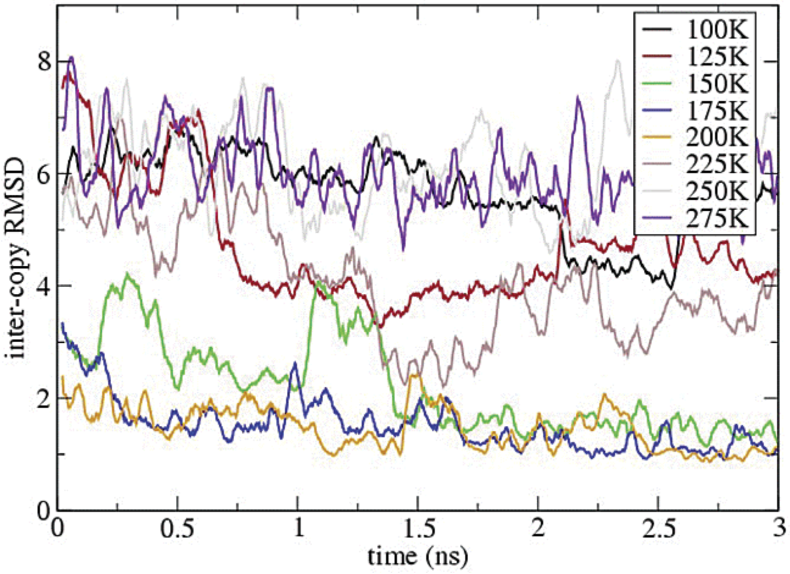


Figure 2.5 Average all-atom deviations between the three LES copies during GB + LES simulation at various temperatures. All simulations were initiated with the same set of alternate LES conformations and the initial deviation between the copies was ~4.5 Å. All of the copies converge to a single family of structures in 1-1.5 ns at 150, 175, and 200K.

To demonstrate this approach, we initiated simulations with three LES copies of the loop region and assigned each copy a different initial structure chosen at random from high-temperature dynamics. We calculated the average all-atom deviation of the copies from each other during the simulation (fit to the noncopied stem), and present the average RMSD of the copies as a function of time in Figure 2.5. This procedure was repeated for temperatures in the range of 100-275K in increments of 25K. This gives us a temperature-dependent measure of how quickly the copies converge to a similar set of structures, regardless of what that structure may be. At low temperatures, such as 100K, the deviation remains high because the barriers are still too large to overcome on this time scale even with LES. At 150, 175, and 200 K, all of the copies converge to a set of structures that differ from each other by <2 Å for all atoms, even though they started from different conformations. At higher temperatures, the deviation remains large not because of trapping, instead the copies show large motions but do not sample a single structure. This suggests that a useful range for obtaining a converged prediction is 150-200 K.

Without LES, one must repeat the simulations from different initial conditions for each optimization protocol to get a measure of its reliability, although we can and do carry out independent LES simulations to gain further confidence in their convergence as well. We are currently investigating alternate approaches to aid in selecting the optimal temperature or the use of replica exchange approaches[58] to avoid the need to select a single temperature.

Next, simulations were carried out starting from the incorrect conformation under conditions in which the correct conformation was stable. In the non-LES GB simulation (GB non-LESb, left graph in Figure 2.6), the RNA underwent conformational change at about 1100 ps, a time scale very similar to the 1200 ps reported for non-LES GB by Williams and Hall. The hydrogen bond geometry started to reorganize in less than 400 ps, but the successful transition was not achieved until about 1 ns (as measured by RMSD). The critical conformational change involved rotation of the U5 N1-C1' torsion angle and took place over a short time scale (<10 ps, as indicated by the sudden drop of RMSD from 2.2 Å to about 1 Å).

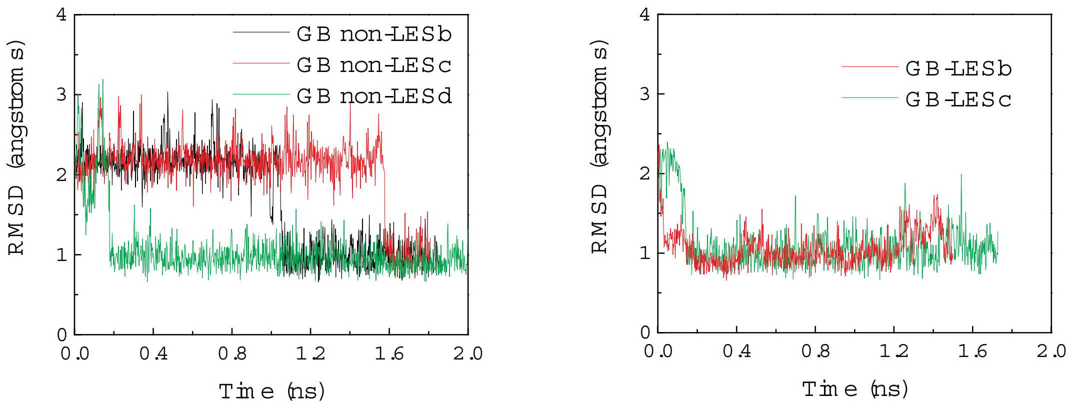


Figure 2.6 The loop RMSD (compared to the correct structure) as a function of time. On the left are three independent non-LES GB simulations (GB non-LESb, c, d). On the right are two GB + LES simulations (GB-LESb and GB-LESc). Three LES copies were employed for the entire UUCG loop. The transition from incorrect to correct structure occurs much more rapidly in the LES simulations.

To gain a better understanding of the pathways of this transition, we performed three additional simulations that differed only in the assignment of initial velocities. Analysis of the trajectory data indicates that two of these simulations (GB non-LESc and d) showed similar transition pathways (but different time scales, shown in the left graph of Figure 2.6). However, the other non-LES simulation failed to convert to the correct structure even after 10 ns (data not shown). During that simulation, a large change in backbone conformation was observed near C7, and the U5 base partially flipped out of the loop.

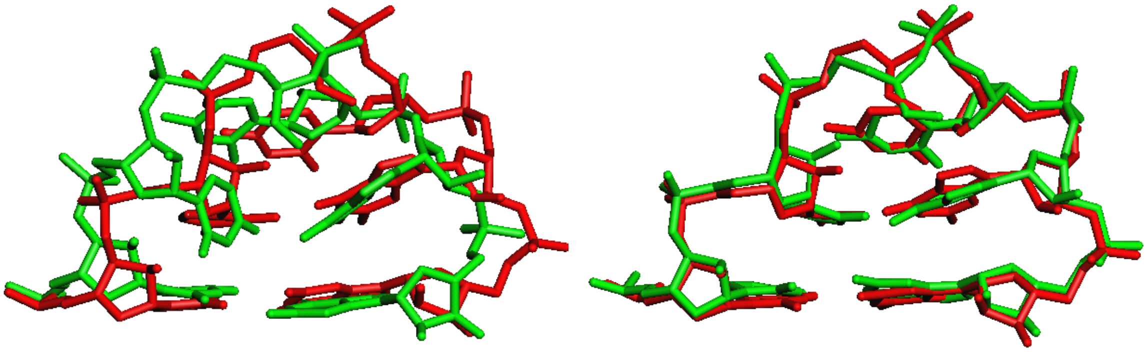


Figure 2.7 Comparison of the loop regions in the correct NMR structure (red) and the average LES MD structure (green) from GB simulations using three LES copies of the UUCG loop region. Only the loop and the first base pair of the stem are shown (residues 4-9 except the mobile U6 base). The left image shows the initial (incorrect) structure. In addition to the base pair hydrogen bond differences shown in Figure 2.3, there is severe buckling of the U5:G8 base pair, as well as other significant differences in backbone conformation on the 5' end of the loop. The right image shows the same comparison after 500 ps of GB + LES molecular dynamics simulation (GB-LESb). All of the major differences have been corrected.

In the right graph of Figure 2.6, results from two alternate GB + LES simulations at 200K are shown (GB-LESb and c). Similar to that observed for non-LES GB, a reduction in RMSD from 2.3 to 1.0 Å occurs in both simulations, demonstrating that these GB + LES simulations achieved the transition from incorrect to correct conformation. However, the transition occurs on a substantially shorter time scale with LES than for the single-copy simulations. These RMSD values represent an average value for the entire three-copy LES system; each of the three copies converted to the correct NMR structure, and the details of the transition for each copy will be discussed in further detail below. Figure 2.7 shows the comparison of the loop regions in the correct NMR structure and the average LES MD structure from GB + LES simulation after 500 ps of molecular dynamics simulation. All of the major differences from the initial (incorrect) structure have been corrected. Similar to observations based on PME + LES and standard GB simulations, the U6 base samples multiple conformations even when the remainder of the stem-loop system samples the correct geometry.

We further examined the sensitivity of these results to the simulation temperature. We expected and observed that at a lower temperature (130K, GB-LESd) the transition takes place on a longer time scale than at 200K. However, when the temperature was reduced to 100K, the C conformation remained stable for the LES copies but the I conformation was also stable, failing to convert to the C conformation during the 4 ns simulation (GB-LESe). This is consistent with our previous observation that the alternate LES conformations did not converge to a single structure at this temperature (Figure 2.5). Thus, the structure was kinetically trapped under these conditions even with LES.

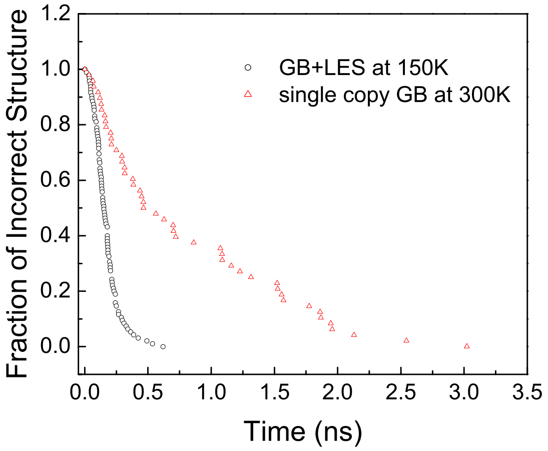


Figure 2.8 RNA tetraloop folding kinetics characterizing the conversion of an ensemble of incorrect structures to the correct one. The red triangles denote 48 independent single-copy GB simulations at 300K. The black circles represent 96 GB + LES trajectories at 150K.

One must be cautious in the use of the time of a single observed transition to represent the actual rate that would be obtained for an ensemble of such events. Likewise, comparison of times obtained from relatively few events using different methodologies may not reflect the actual differences in barrier crossing rates. We therefore performed more statistically significant comparisons of the rate of I→C transition from standard GB and GB + LES simulations.

A series of simulations were carried out for the incorrect structure, differing only in initial velocity assignments. For standard GB, 48 independent trajectories were obtained at 300 K. Similarly, 32 GB + LES simulations were performed at 150K, each using three copies, thus providing 96 loop trajectories. In each case, the time dependence of first passage (I→C) was collected until the entire ensemble had undergone the I→C transition (Figure 2.8). This ensures that differences in rates are due solely to reduction in free-energy barriers and are not the result of variance in transition times because of poor statistical sampling. The process is clearly more rapid in GB + LES than for standard GB, with estimated rate constants of 6.5 $ns^{-1}$ and 1.2 $ns^{-1}$, respectively.

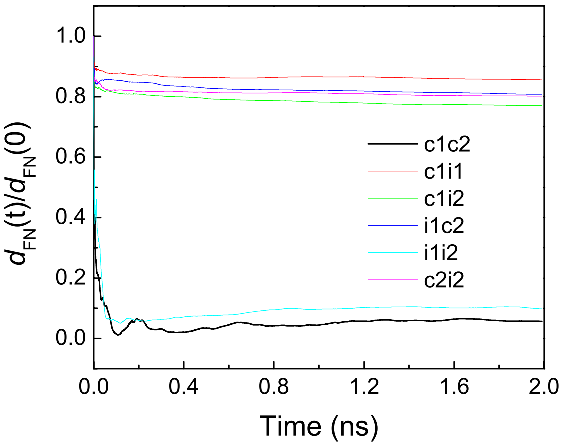


Figure 2.9 Plots of the normalized nonbonded force metric, $d_{FN}(t)/d_{FN}(0)$, as a function of time for the RNA tetraloop during standard GB simulations at 300 K. Data are calculated for pairs of trajectories, denoted in the legend. c1 and c2 started from the correct structure; i1 and i2 started from the incorrect structure. The self-averaging metric approaches zero on this time scale only for trajectories initiated in the same structure.

As an additional test of convergence, we calculated the force metrics (equation 2.5) for the nonbonded forces acting on the tetraloop atoms as a function of time. These values should approach zero for pairs of trajectories that are sampling the same regions of phase space. In the standard simulations (Figure 2.9), we see subnanosecond convergence for pairs of trajectories that were initiated from the same structure (both incorrect or both correct). However, convergence is not obtained for pairs of simulations started from two different structures. This indicates that these non-LES simulations remain confined to their initial conformational substates and are unable to cross the intervening barriers on this time scale.

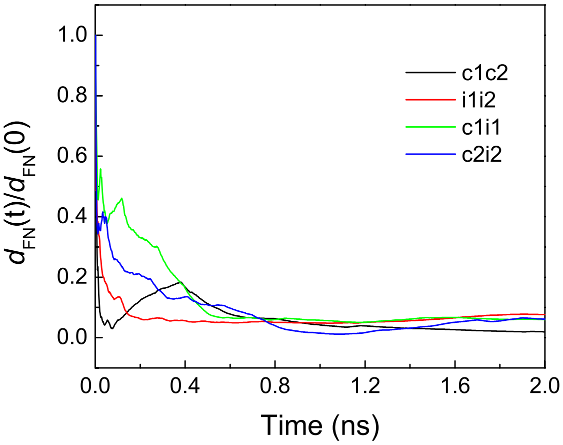


Figure 2.10 Plots of the normalized nonbonded force metric, $d_{FN}(t)/d_{FN}(0)$, as a function of time for the RNA tetraloop during GB + LES simulations at 150 K. Data are calculated for pairs of trajectories, denoted in the legend. c1 and c2 started from the correct structure; i1 and i2 started from the incorrect structure. In contrast to non-LES simulations, the metric approaches zero on this time scale for all pairs of trajectories.

A remarkable difference can be seen from the force metrics for the GB + LES simulations. In Figure 2.10, all of the nonbonded force metrics decay to 0 after ~500 ps, irrespective of the initial conformations. These results provide additional evidence that GB + LES can explore different conformational substates on the nanosecond time scale more efficiently than standard GB simulations.

One advantage of LES is the ability to accelerate conformational transitions. Another potential advantage is that the copies may explore alternate regions in phase space, thus providing multiple transition events in a single simulation at a reduced computational cost as compared to multiple non-LES simulations. In our previous investigation of this I→C transition in explicit solvent[83], the instantaneous backbone RMSD values between pairs of different copies were less than 0.2 Å throughout the entire simulation. This indicates that the copies not only were unable to explore alternate transition pathways, but a time coupling was also present and therefore only a single time scale for the event could be sampled.

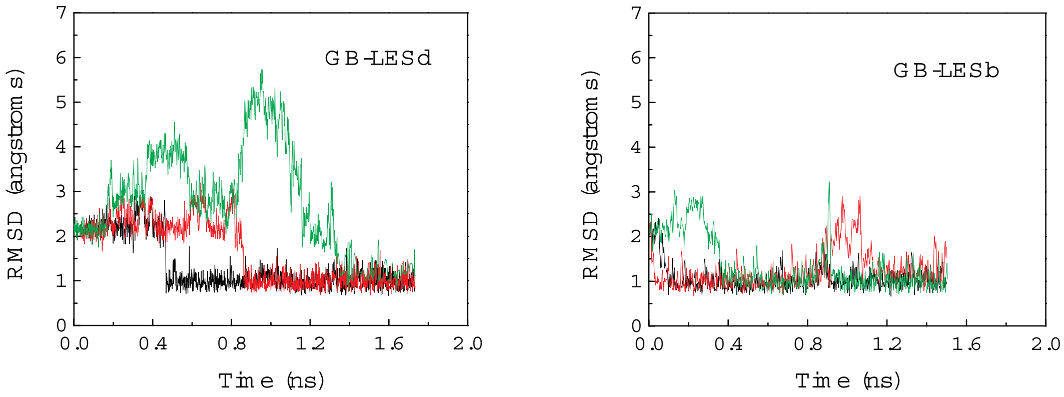


Figure 2.11 The loop RMSD (compared to the correct structure) as a function of time for each of the three LES copies. Both simulations were initiated from the incorrect structure. The copies undergo the transition to the correct structure at different times. The same behavior is seen during simulation at 130K (left) and 200K (right).

As described above, our goal during our development of the GB + LES model was to overcome this weakness of LES through individual solvation of each copy, avoiding the caging effect associated with a single explicit solvent cavity for all copies. This directly affects the ability of the LES simulation to model alternate transition pathways in a single simulation. We therefore investigated whether our approach was successful in overcoming it. In Figure 2.11, we show the RMSD value for each of the copies as a function of time (in contrast to Figure 2.6 in which the average RMSD for all copies was shown) for GB + LES simulation at 130K (GB-LESd) and 200K (GB-LESb). It is apparent that each copy undergoes the I→C transition. In contrast to the simulation with explicit solvent, however, the copies show much greater independence and undergo the transition at significantly different times. Similar results were described above in which the deviations in copy conformations were large (Figure 2.5), but in that case the divergence was a product of the initial coordinate generation and in the present case the copies are able to spontaneously explore alternate pathways after starting from the same conformation.

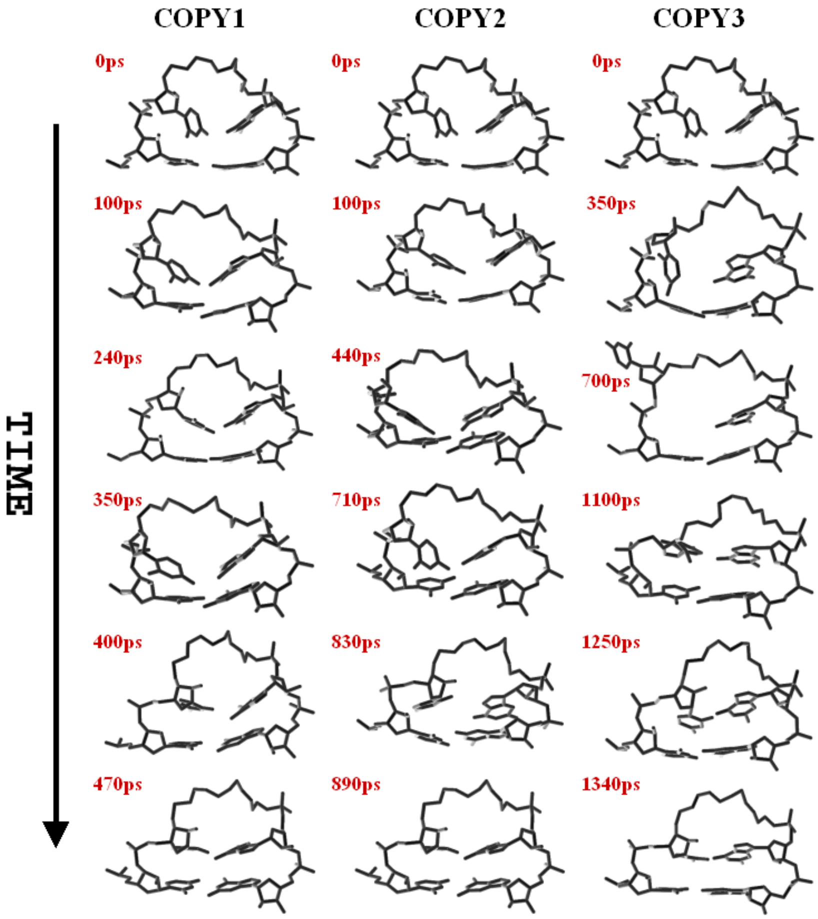


Figure 2.12 Snapshots of the U5:G8 base pair during simulation GB + LESd that employed three LES copies of the UUCG loop. Also shown is the C4:G9 base pair in the stem, which has a stacking interaction with the U5:G8 pair. The three columns correspond to snapshots of each of the three sets of copies. For clarity, the rest of the system is not drawn. The details of the transitions are provided in the main text.

The difference in peak RMSD values for each copy suggests that different transition pathways are explored in this single simulation. We examined the events for each copy in simulation GB + LESd in detail and show representative snapshots in Figure 2.12. In the first pathway (left column in Figure 2.12), the most direct interconversion is observed. At 100 ps, the U5-G8 base pair partially lost the original hydrogen bond pattern because of the rotation of the G8 base pair about N9-C' dihedral. U5 is then observed to flip back and forth via rotation of N1-C1' dihedral. At 400 ps, this flipping motion results in formation of partially correct structure. Meanwhile, the breaking of the N3-O6 and O4-N1 reverse-wobble hydrogen bonds and formation of the bifurcated pattern involving O2-N1 and O2-N2 also are achieved, as shown by the hydrogen bond plot (Figure 2.13). At this point, the hydrogen bond pattern is correct while the relative orientation of these bases and the stacking of the UG pair against the stem CG pair differ from that found in the correct conformation. At 470 ps, both the correct hydrogen bond pattern and stacking are attained. This conformation is retained throughout the remainder of the simulation.

The middle column of Figure 2.12 shows a similar transition pattern for the second copy. However, the reorganization of the two U5:G8 bases after partial separation takes ~600 ps, and the base pair reformed with the correct hydrogen bond pattern but with stacking against the stem that is similar to the incorrect conformation. The correct stacking pattern and relative orientation were attained within the next 60 ps. In this case, the most significant structural transition was also achieved by rapid rotation of base U5 N1-C1' dihedral angle, but the backbone near C4 and U5 shows significantly greater distortion during the transition as compared to the first pathway.

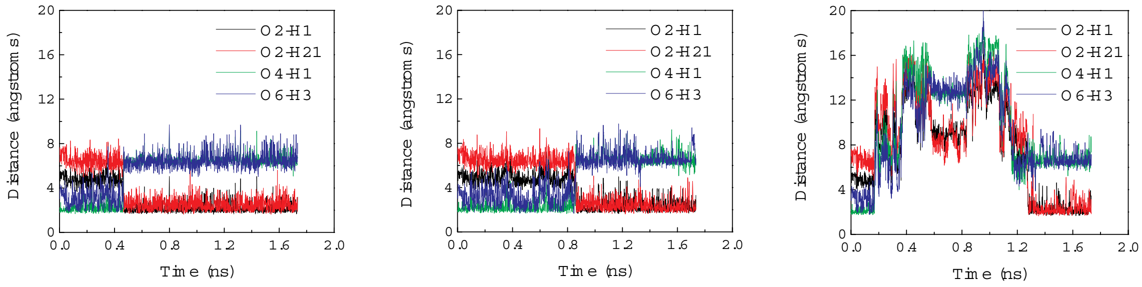


Figure 2.13 Distances between hydrogen and oxygen atoms corresponding to the hydrogen bonds in the U5:G8 base pair shown in Figure 2.3. Green and blue are distances

of the initial incorrect hydrogen patterns; red and black are distances of the final correct hydrogen patterns. The three figures represent the three loop copies and correspond to the three columns in Figure 2.12.

In contrast, the right column shows a dramatically different interconversion sampled by the third copy. At 100 ps, the U5-G8 bases separate and fluctuate back and forth for about 600 ps. At 700 ps, the U5 base flips out of the loop, remaining in this location for more than 400 ps, finally shifting back toward the loop through a rotation of the C3'-C4' dihedral. The base is then positioned to form the correct hydrogen bonds and converts to the correct structure through shifts in stacking that occur over the next 100 ps.

Detailed comparison of the pathways sampled by the LES copies with those sampled during non-LES GB simulations revealed that the first two LES pathways are both remarkably similar to those observed in all three successful GB non-LES simulations, despite the difference in the time scale of the transitions. The third LES pathway shown in Figure 2.12 was not observed in the non-LES GB simulations. However, a flipping motion of the U5 base similar to that which initiated this transition was observed in the non-LES GB simulation described above that did not convert to the correct structure. A similar flipping motion of the U5 base was observed in standard MD simulation of this system in explicit solvent (Miller, J., pers. comm.), suggesting that the process may be involved in a pathway of lower probability.

That these different transitions occur as three independent events observed in a single LES simulation is quite remarkable and further demonstrates the advantages of this combined methodology: most current simulation methods do not allow for even a single observation of such events and therefore cannot provide rapid insight concerning the existence and nature of alternate transition pathways.

## 2.5 Energetic Barrier Reduction by LES

Many previous studies have explored formal properties of the potential energy surface and kinetic energy of the LES approximation[51,53,94]. These theoretical explorations established the foundation for the use of the LES method. Roitberg and Elber also showed intuitively how and why the energy barrier was reduced by using LES[52]. If let $U_B$ be the value of the transition state connecting two energy minima $U_I$ and $U_C$, then the energy barrier for moving I to C in the real system is simply $U_B$ - $U_I$. The height of barrier for LES system would be the same if all copies were placed on the same transition state. However, this is not a transition state for the LES system. The corresponding transition state will involve only one copy on the transition sate while the rest of copies will be at their minimum energy states, which leads to a barrier height of $(1/N)(U_B - U_I)$, where N is the number of copies for the LES system, if no strong coupling between the LES region and the non-LES region is assumed. Although this approximate expression demonstrated the LES yields a lower energy barrier, the rigorous derivation for the relation between the barrier of real system and that of LES system is still impossible. For a complete understanding of the thermodynamics and the dynamics of LES system, only the information of minima is not adequate, the quantitative properties of the barrier heights are needed as well[95].

The main idea of this work is to use real statistics to directly calculate the energy barriers for both GB and GB + LES simulations and further to show energy barrier reduction by LES approximation. Through large–scale calculations, which would be formidable without the dramatic recent increase in computer power, we are able to provide direct evidence of significant barrier reduction by LES in a RNA tetraloop system.

We first performed GB and GB + LES simulations at different temperatures. For each temperature, a series of simulations were carried out starting from the incorrect structure I[81], differing only in initial velocity assignment. For standard GB, 60 independent trajectories were obtained at each temperature ranging from 300K to 400K with 25K intervals. Similarly, 60 GB + LES simulations were performed at each temperature ranging from 100K to 300K, each using 3 LES copies for the UUCG loop, thus providing 180 loop trajectories. In each case, the time dependence of first passage (I→C) was collected until the entire ensemble had undergone the I→C transition.

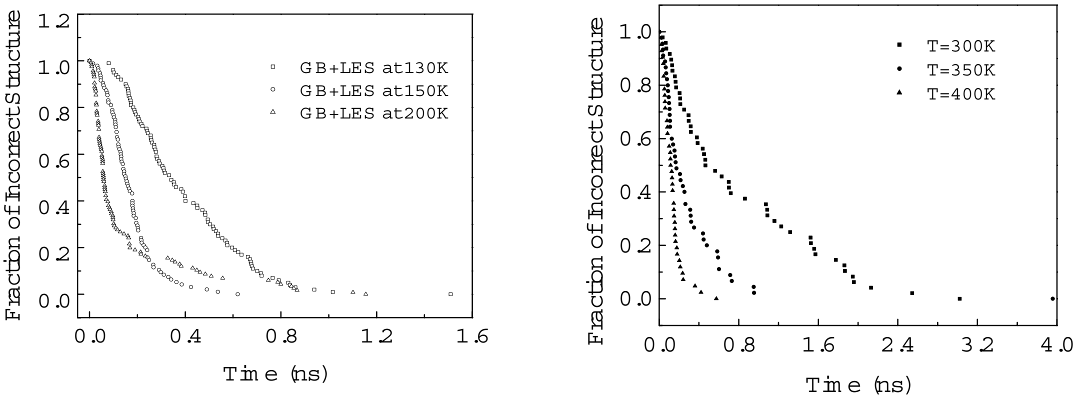


Figure 2.14 The fraction of incorrect structure as a function of time in GB and GB+LES simulations. For GB simulations each curve represents 60 molecular dynamics trajectories; for GB+LES each curve was obtained from 60 LES simulations, splitting to 180 single copy trajectories. Both curves can fit very well by exponential decay. The fitting parameters are used to calculate the folding rate in Figure 2.15.

The time dependence of the fraction of the ensemble that remains in the incorrect conformation was shown in Figure 2.14 and was further fit to a single exponential (*fraction of the incorrect% = exp(-kt)*), assuming the first order kinetic behavior. All of the curves can fit very well by exponential decay, and each fitting parameter *k* therefore gives the transition rate for each ensemble of simulations at one particular temperature.

The common aspect of small molecule reaction is Arrhenius-like temperature dependence. This requires that the rate constant can be written to a good approximation in the following form[96]

$$k = A(T)\exp(-\Delta E^{\neq} / RT) \qquad (2.6)$$

where $\Delta E^{\neq}$, the activation energy, is approximately independent of temperature and *A(T)* is the pre-exponential factor with only a weak temperature dependence. Given the above equation, we have

$$\frac{d\ln k}{d(1/T)} \cong -\frac{\Delta E^{\neq}}{R} \qquad (2.7)$$

a plot of *lnk* versus *1/T* approximates a straight line. The $ln\tau_{1/2}$ ( $ln\tau_{1/2}$ = *-lnk + constant*) as a function of inverse temperature for GB and GB+LES simulations is shown in Figure

2.15. A straight line can fit the data very well for both cases. Based on Arrhenius approximation the calculated apparent activation energy barrier for I→C transition in GB+LES is about 1.2±0.4 kcal/mol while the corresponding value in GB is about 4.6±0.1 kcal/mol. This result clearly shows that differences in rates are due solely to reduction in free energy barrier. In fact, the enhanced sampling through smoothing the energy barrier is quite significant in the RNA I→C transition.

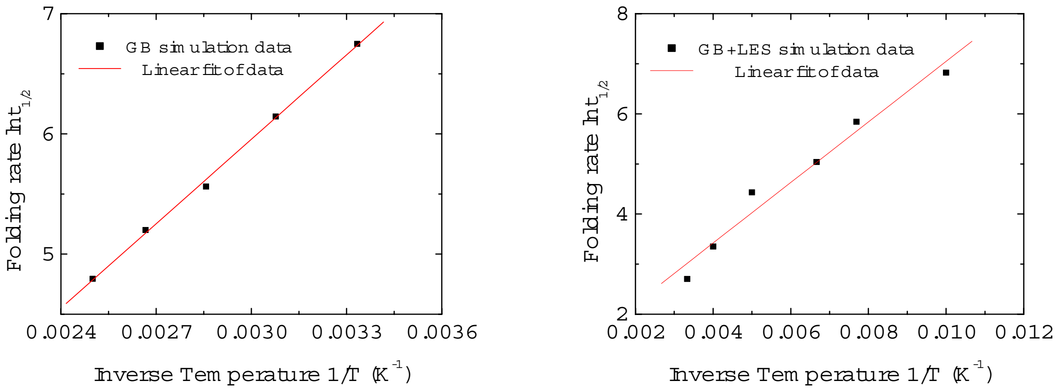


Figure 2.15 The ln$\tau_{1/2}$ as a function of inverse temperatures for GB (left) and GB+LES (right) simulations. A straight line can fit the data very well. The fitting parameters are used to calculate the apparent activation energy barriers.

To completely understand the free energy surface involving transition between I and C structures, only the knowledge of barrier height is not enough. We also need information of the energy difference between two minima. We investigate this free energy difference through an approximate MM-GB approach[92,93]. In this method, 2000 structures for each conformation were collected from explicit solvent simulations, stripped of water, and used in molecular mechanic and GB energy evaluation. The energy distribution probability as shown in Figure 2.16 indicates that there is an obvious preference for the correct structure; the estimated free energy difference is as large as 7.5±10.47 kcal/mol. This relatively large energy difference explains why the I structure is not highly populated at equilibrium. It's of interest to note that two loop conformations have almost identical internal energy; the difference is mainly from electronic solvation term, which favors the correct structure by about 6.2±9.5 kcal/mol. By combining the energy difference calculated from MM-GB method with the barrier height results obtained in previous multiple simulations we are able to construct the free energy profile for the conversion between I and C structures (Figure 2.17).

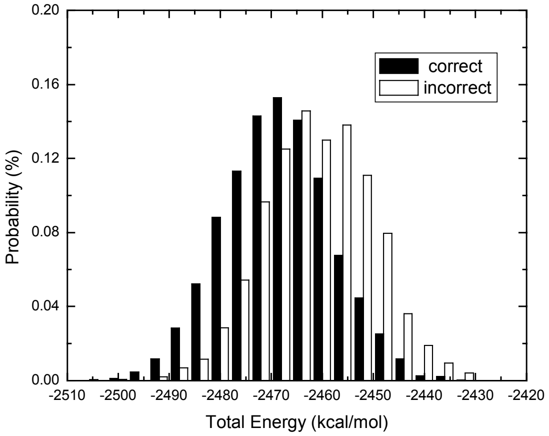


Figure 2.16 MM-GB energy calculations for correct (black) and incorrect (white) conformations.

Although the correct structure is much more stable than the incorrect structure, the relatively high barrier between them hampers the conversion from I to C structures. That explains why most of explicit solvent simulations get trapped in their initial conformation if started from incorrect structure. When the LES is employed to enhance the sampling, the energy barrier is reduced to 1.2 kcal/mol, enabling the I→C transition to occur more easily. The actual transition may involve several frog-leap like steps; each time only one copy does the I→C transition. This scenario is in agreement with our observation in GB+LES simulations. The transition is usually transient at higher temperatures. RMSD plots always show a one-time transition. However, at lower temperatures, the transition of each copy is temporally separated. The one-by-one jumping pattern can be easily seen from our previous RMSD plots at 130 K as shown in left graph of Figure 2.11. It is worthy of noting that the one copy a time C→I backward jumping is made easier as well, but moving all copies from C to I will involve multiple higher-energy jumps, making the C→I transition much more energetically unfavorable and therefore a rare event to occur during the GB+LES simulation.

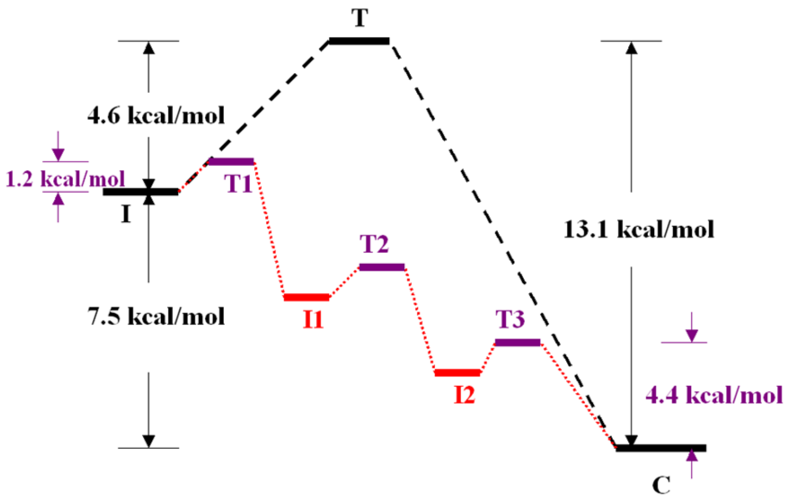


Figure 2.17 A schematic diagram of effective energetic barriers for RNA tetraloop system with GB and GB+LES methods. Black lines are used for single copy GB method; and red lines are used for GB+LES method. The purple lines represent a hypothetical step-by-step jumping. (I stands for incorrect NMR structure, C stands for correct NMR structure and T stands for transition state; T1-T3 are proposed transition states for LES system with only one loop conformation sitting on the corresponding transition state for the real system; I1 and I2 are proposed intermediate structures with 1 and 2 loop conformations in correct structure respectively.)

Proteins, DNA and RNA are complex systems so that their reactions typically have characteristics not often observed in small molecules[97]. However, in our simulation, we observed nearly perfect Arrhenius-like temperature dependence for the reaction rates. This is somehow out of our expectation. The possible explanation for this is that the transition is relatively “simple” – only four bases involved in the process thus no multiple time-scale motions involved in contrast to protein folding problem. Consequently, the barrier is more energetic whereas the entropy contribution to the barrier is not significant. The use of multiple copies in LES tends to amplify the entropy effect. If the two states have disparate entropy contributions, the transition probability might be wrong when using LES. Fortunately, this is not the case here.

The LES approximation is found to efficiently enhance the sampling while the energy minimum is the same as compared to the real system. We have used real statistics to provide direct evidence of the significant energy barrier reduction by this mean field approximation. The present simulation addressed successfully the enhanced sampling is through the energy barrier reduction not from the increased local temperature. This has been an issue in many previous simulations[98], especially when implicit solvent model is used where the simulation becomes extremely sensitive to the selected temperatures due to the lack of the friction term[46]. Moreover, although the implicit solvent model has been used to approximate the solvation throughout the simulations, we believe that the general conclusion of barrier reduction by LES holds true for other solvation model as well.

## 2.6 Conclusions

We have shown for the first time that the GB model can be used to provide a proper implicit solvation treatment for a simulation employing locally enhanced sampling. We derived an exact approach to combining the two algorithms in which each LES copy is individually solvated, rather than the solvation of the ensemble of copies that is obtained with explicit solvation. Our approach does lead to increased calculations in the non-LES region (not seen with LES/PME) but a reasonable approximation reduced this overhead by ~70%. This approach has been implemented in the AMBER suite of programs (version 8).

We carried out tests of the combined approach using a well-studied RNA UUCG tetraloop system. Previous work had shown that the use of either GB solvation or LES in explicit solvent was able to successfully model the conversion of an incorrect to correct conformation[73,83]. We have shown that the GB + LES simulations have the same ability to reproduce the correct conformational change, but with greater computational efficiency than obtained with either GB or LES alone. This improved convergence was demonstrated by (1) comparison of rate constants for the conversion obtained from a large set of independent simulations and (2) calculation of force metrics that clearly show improved sampling in the LES simulations.

The reduction in barrier heights provided by LES results in simulations that are much more efficient than single-copy GB simulations. However, these combined GB + LES simulations are more sensitive to the temperature and number of copies than corresponding non-LES GB simulations or LES simulations carried out in the explicit solvent. We showed that one way to evaluate these parameters for a system in which the correct structure is unknown is to monitor the convergence of the conformations of the alternate copies after assigning them different initial coordinates.

Finally, we make the remarkable observation that a single LES simulation can provide multiple (and qualitatively different) instances of key transitions. LES simulations in explicit solvent for this system were unable to provide independent transitions for the copies because of a caging effect arising from the sharing of a single solvent cavity for all LES copies. We believe that this approach is likely to be an important component of all-atom structure refinement of biomolecular systems, particularly when a portion of the structure, such as a loop region, is poorly determined and requires additional local sampling.

# Chapter 3

# A Modified Replica Exchange Simulation Method for Local Structure Refinement

## 3.1 Introduction

The potential energy surfaces of biological systems have long been recognized to be rugged[99-101], which hampers the efficiency of conformational transitions between various local minima. Due to this property of the energy landscape, efficient computational approaches to searching for low-energy minima in these complex systems present a great challenge. This sampling problem can preclude success even when the correct Hamiltonian of the system is used in the simulations. Thus, numerous algorithms have been developed to improve the sampling of phase space for molecular simulations[100,102].

A general approach to surface flattening is obtained by the application of mean-field theory. Among these, the LES method[51,83,94,103,104] is of particular interest for structure optimization due to the equivalence of the LES global energy minimum to that of the original system[52,53], thus avoiding cumbersome mapping procedures. The LES method has been successfully applied to many biomolecular problems such as structure prediction[52,83,105,106], free energy calculations[107], and ligand design[108].

Another category of methods that has seen a recent increase in use is often referred as generalized ensemble algorithms, including multi-canonical methods[54,55], simulated tempering[56,57] and the replica exchange method (REM)[58,109,110]. Multi-canonical methods and simulated tempering improve the sampling by replacing the Boltzmann factor $exp(-\beta E)$ with the multi-canonical probability $\mathrm{n}(E)^{-1}$. This allows the system to sample freely as a one-dimensional random walk in energy or temperature space. A difficulty in applying these two methods is in determining the multi-canonical probability functions *a priori*[57].

In REM, several non-interacting copies (replicas) are independently and simultaneously simulated at different temperatures. At intervals during the simulations, conformations of the system being sampled at different temperatures are exchanged based on a Metropolis-type criterion that considers the probability of each conformation being sampled at the alternate temperature. In this way, REM is hampered to a lesser degree by the local minima problem, since the low temperature simulations (replicas) have the potential to escape kinetic traps by jumping to minima that are being sampled by the higher temperature replicas. On the other hand, the high energy regions of conformational basins often sampled by the high-temperature replicas can be relaxed in a way similar to the temperature annealing method. Moreover, the transition probability is constructed such that the canonical ensemble properties are maintained during the simulation, thus providing potentially useful information about conformational probabilities as a function of temperature. Due to these advantages, REM has been widely applied to studies of peptide and small protein folding[58,69,110,111].

For large systems, however, application of REM can require significant computational resources, thus limiting its advantages. It has been shown that the number of replicas needed to cover a given temperature range increases as $O(f^{1/2})$ for a system with $f$ degrees of freedom[65]. Several promising techniques have been proposed to deal with this apparent disadvantage to REM. In a method proposed by Sugita, REM was coupled to the multi-canonical method to take the advantage of both techniques[59,112]. Pak proposed a method combining REM with the generalized effective potential method[113]. Takada developed an alternative REM, called Hamiltonian REM[65]. In their method, the Hamiltonian was separated into two parts; one was assumed to be tightly coupled to temperature space and the other was not. REM was then performed on the former part of Hamiltonian, including torsion angle and repulsive vdw terms, which are mainly responsible for the rugged energy surface. By doing so, the number of replicas needed is reduced by excluding the other part of degrees of freedom such as solvent, bond length and bond angle terms. Two of their examples, called scaled hydrophobicity REM and phantom chain REM, clearly demonstrated the strength of the model.

Motivated by our desire to focus the enhanced sampling on a subset of the degrees of freedom in the system, we propose a different partitioning of the Hamiltonian for a large system. We define a subset of the system as the "sampling desired" region. The remainder is often the part of system that we are not specifically interested in, but which must be present during the simulation. A typical example of such a division would be into a protein core built on a homology model and a surface loop for which experimental structural data is unavailable.

In one of our proposed methods, partial replica exchange method (PREM), the system is divided into two regions. Under the assumption of no strong coupling between them, the two regions are coupled to separate temperature baths. This is often a reasonable approximation for many biological applications. For example, DNA experiments and simulations show that flipping base has minor effects on the global DNA stability, and the structural change is usually localized to the flip site. Once this two-temperature-bath system is built, then only the desired subspace is simulated over a range of temperatures, with periodic exchanges performed using a standard REM approach while the other part maintains the same temperature during the simulation. Acceptable exchange ratio can thus be obtained with fewer replicas than with traditional REM.

In contrast to PREM, replicas in our local replica exchange method (LREM) are made of only a portion of the system under the scheme of our LES implementation in the AMBER program[83]. We note that replica has common features with LES. Also this is an extension of our weak coupling assumption, so we can approximate that the "large" portion will have identical coordinates in each replica. This means we do not have to recalculate the energies and forces for terms fully in the large portion. This would be fine if it was frozen, but we can assume weak coupling (rather than none) and allow the region to move using the LES approximation. On the other hand, one of the major problems with LES is how to choose a single temperature for LES copies as described previously[78,98]. Therefore we prefer an LES method without the temperature selection problem. For the above two reasons, we devised a second modified REM method called LREM, which will employ certain aspects of both LES and PREM. In this method, these sub-system replicas interact with the non-replicated region in an average manner. Each LES copy is coupled to a separate temperature bath. Standard replica exchange is then carried out, with the exception that we need exchange only a small portion of the system, not the entire system. LREM represents a further approximation beyond PREM method. But compared to PREM, the efficiency of LREM is improved by sharing the same non-replicated region. Moreover, the reduction in barrier heights intrinsically provided by the LES component of the method also suggests that a smaller temperature range may be sufficient to achieve adequate sampling of the relevant conformations. Besides, the range of temperatures used in the LREM method avoids the difficulty of choosing an appropriate single temperature for LES simulations.

## 3.2 Theory

### 3.2.1 Replica Exchange Method (REM)

In standard REM, the simulated system consists of ***n*** non-interacting copies (replicas) at ***n*** different temperatures[58]. The positions, momenta and temperature for each replica are denoted by $\{q^{[i]}, p^{[i]}, T_M\}$, $i = 1,\ldots, n$; $M = 1,\ldots, n$. The equilibrium probability for this generalized ensemble is

$$W(p^{[i]}, q^{[i]}, T_M) = \exp\{-\sum_{i=1}^{n} \frac{1}{k_B T_M} H(p^{[i]}, q^{[i]})\} \qquad (3.1)$$

where Hamiltonian $H(p^{[i]}, q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and potential energy $E(q^{[i]})$. For convenience we denote $\{p^{[i]}, q^{[i]}\}$ at temperature $T_M$ by $x_M^{[i]}$ and further define X = $\{x_1^{[i(1)]}$, …, $x_M^{[i(M)]}\}$ as one state of the generalized ensemble. We now consider exchanging a pair of replicas. Suppose we exchange replicas *i* and *j*, which are at temperatures $T_m$ and $T_n$ respectively,

$$X = \{\ldots; x_m^{[i]}; \ldots; x_n^{[j]}; \ldots\} \rightarrow X' = \{\ldots; x_m^{[j]}; \ldots; x_n^{[i]}; \ldots\} \qquad (3.2)$$

In order to maintain detailed balance of the generalized system, microscopic reversibility has to be satisfied, thus gives,

$$W(X)\,\rho(X \to X') = W(X')\,\rho(X' \to X) \qquad (3.3)$$

where $\rho(X \rightarrow X')$ is the exchange probability between two states *X* and *X'*. For canonical ensemble, the potential energy *E* rather than the total Hamiltonian *H* will be used simply because the momentum part can be integrated out. Inserting equation 3.1 to equation 3.3, we obtain the following equation for the exchange probability,

$$\frac{\rho(X \to X')}{\rho(X' \to X)} = \exp\{(\frac{1}{k_B T_m} - \frac{1}{k_B T_n})(E(q^{[i]}) - E(q^{[j]}))\} \qquad (3.4)$$

The above condition is usually satisfied by the use of Metropolis criterion[114]. In practice, several replicas at certain target temperatures are simulated simultaneously and independently for certain MD steps, then a pair of replicas at neighboring temperatures are tested to exchange with the probability of $\rho$ calculated by equation 3.4. If the exchange is accepted, the temperatures of two replicas will be swapped, and the velocities will be scaled accordingly to match the new temperatures. Otherwise, if the exchange is rejected, each replica will proceed with its own trajectory.

As have been mentioned before, one of the major limitations of REM is the number of replicas grows proportionally to the square root of degrees of freedom. Based on equation 3.4, we provide a crude justification for this. In Figure 3.1, we show potential energy fluctuations of two replicas sampling at the target temperatures $T_n$ and $T_{n-1}$. The instantaneous energy fluctuation $\delta E$ in a given simulation at temperature *T* is proportional to $\sqrt{f}\,T$, and the average energy gap $\Delta E$ between two neighboring replicas is proportional to $f\Delta T$. Thus ΔE/δE is proportional to $\Delta T\sqrt{f}/T$. In order to keep $\Delta E$ and $\delta E$ comparable to reach a reasonable acceptance ratio for any system size, $\Delta T\sqrt{f}/T$ has to be roughly the same magnitude, which leads to $\Delta T \sim 1/\sqrt{f}$, indicating the temperature gap has to decrease when the system size increases for a given acceptance ratio.

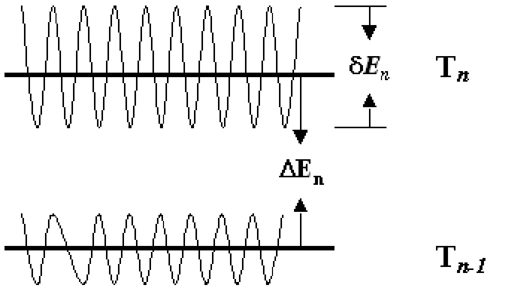


Figure 3.1 A schematic diagram illustrating the energy fluctuations for simulations at two temperatures for neighboring replicas. In order to obtain high exchange probabilities, the energy fluctuations δE in each simulation should be of comparable magnitude to the mean energy difference ΔE.

Importantly, as noted in equation 3.4, the average exchange probability $P_{acc}$ is proportional to $exp(-\Delta T^2/T^2)$, which implies that greater acceptance ratio will require smaller temperature gap $\Delta T$ to reach. On the other hand, in order to enhance the overall sampling, $\Delta T$ should be as large as possible to span the greatest temperature range by using a fixed number of replicas.

### 3.2.2 Partial Replica Exchange Method (PREM)

In PREM, the system is divided into a "bath" and a smaller, more interesting "focused" region. Under the weak coupling assumption between these two regions, they are coupled to separate temperature baths. For the new system composed of bath atoms of type $A$ and focused atoms of type $B$, the kinetic energy is taken as,

$$K = \frac{1}{2}\sum_{A}\frac{|p_A|^2}{m_A} + \frac{1}{2}\sum_{B}\frac{|p_B|^2}{m_B} \qquad (3.5)$$

and the potential energy is

$$E = E_{AA} + [E_{AB}(q_A, q_B) + E_{BB}(q_B)] \qquad (3.6)$$

These two equations are essentially the same as that of the original system. However, the instantaneous temperature is related to the kinetic energy *K* in a different way as follows

$$K = \frac{k_B}{2}(3N_A T_A + 3N_B T_B) \qquad (3.7)$$

where $T_A$ and $T_B$ are temperatures, $N_A$ and $N_B$ are degrees of freedom for the bath and focused regions respectively. Once separate temperature baths are applied to two regions, the PREM system consists of *M* non-interacting replicas of the original system with focused regions at *M* different temperatures while all the bath regions maintain the same temperature. Let $X = \{ x_1^{[i(1)]}, \ldots, x_M^{[i(M)]} \}$ stand for a "state" in this generalized ensemble. The state *X* is specified by *M* sets of configuration $q^{[i]}$ and momentum $p^{[i]}$ of <u>all</u> of the atoms in replica *i* at temperature $T_m$: $x_m^{[i]} = (p^{[i]}, q^{[i]}, T_m)$, $T_m$ is however the target temperature of focused region only.

If assuming all the bath regions are relatively rigid and not highly dependent on the conformations of focused region at various temperatures, the potential energy of bath region would approximately be cancelled among replicas. In other words, this part of degrees of freedom is not relevant to temperature space during replica exchanges. We now introduce a new Hamiltonian for each replica,

$$\begin{aligned} H &= H_{prem}(\mathrm{r, s})_{\{\mathrm{Tm}\}} + H_{bath}(\mathrm{s})_{\{\mathrm{T0}\}} \\ &= [E_{\mathrm{BB}}(\mathrm{q_B}) + E_{\mathrm{A,B}}(\mathrm{q_A, q_B})]_{\{\mathrm{Tm}\}} + E_{\mathrm{AA}}(\mathrm{q_A})_{\{\mathrm{T0}\}} \qquad (3.8) \end{aligned}$$

where *r* and *s* stand for coordinates and momenta of the focused and bath parts respectively, and we assume all replicas have similar bath energy $H_{bath}$(s). The effective temperature $T_m$ only applies to the degrees of freedom *r* for the focused region; instead the bath region *s* coordinates are coupled with the same temperature $T_0$ for all replicas. Consequently, the appropriate spacing in temperature ($\Delta T$) will only be determined by the number of degrees of freedom for *r*. If *r* is a relatively small number compared to the total degrees of freedom, the number of replica needed is reduced in the PREM method.

Based on the above weak coupling assumptions, the weighting factor of state *X* consisting of *M* replicas can be taken as follows, only depending on the subset of the system,

$$W_{PREM}(X) = \exp\left[\sum_{i=1}^{M} H_{prem}(r^{[i]}, s^{[i]}) / k_B T_m\right] \qquad (3.9)$$

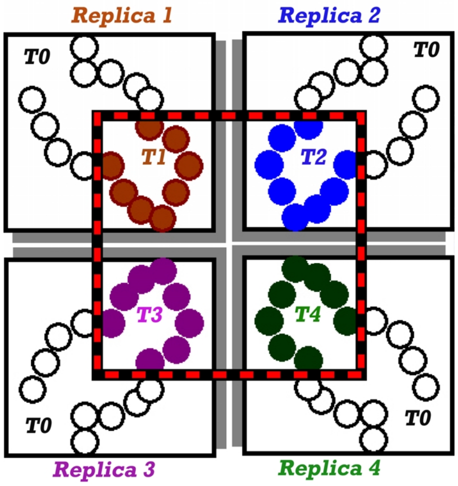


Figure 3.2 A schematic illustration of PREM method, with four replicas. Each maintains the same temperature of $T_0$ for bath region and $T_1$, $T_2$, $T_3$, and $T_4$ for focused region respectively. During the simulation, target temperatures for focused region are swapped with a predefined frequency based on a Metropolis-type criterion.

Thus by following the treatment of Sugita[58], the exchange probability *ρ(X→X')* is given by,

$$\rho(X \to X') = \min[1, \exp(-\Delta)] \qquad (3.10)$$

where $\Delta = (\frac{1}{k_B T^{[i]}} - \frac{1}{k_B T^{[j]}})(H_{prem}(r^{[i]}, s^{[i]}) - H_{prem}(r^{[j]}, s^{[j]}))$

As shown in Figure 3.2, during the simulation, only the focused region is simulated over a range of temperatures, with periodic exchanges performed using standard approach while the bath region is maintained at the same temperature. In this way, we reduced the degrees of freedom taken into account by excluding irrelevant part of the system. Acceptable exchange ratio can thus be obtained with fewer replicas than the conventional REM. Furthermore, as shown by many previous studies[78], heating the bath may not be desirable for local optimization. By introducing local temperatures, we are able to enhance the sampling only in the desired part of the system and meanwhile restrain the other part at lower temperature to keep the integrity of the whole system.

### 3.2.3 Local Replica Exchange Method (LREM)

Motivated by the similarities in the “exchange part of the Hamiltonian” variant of replica exchange to LES approach[51], we further extend the PREM method to a LREM method. In LREM, instead of replicating the entire system, we use LES to construct a new system in which only a subset of atoms is replicated. Each of these replicas interacts with the non-replicated remainder of the system. Similar to standard REM, the sub-system replicas do not interact with each other. For the LES system composed of non-replicated atoms of type A and ***n*** LES copies for atoms of type B, the kinetic energy is taken as

$$K = \frac{1}{2}\sum_{A}\frac{|p_A|^2}{m_A} + \frac{1}{2}\sum_{i=1}^{n}\sum_{B}\frac{|p_{iB}|^2}{m_B} \tag{3.11}$$

and the potential energy is

$$E = V_{AA} + \frac{1}{n}\sum_{i=1}^{n}[E_{AB}(q_A, q_{Bi}) + E_{BB}(q_{Bi})] \tag{3.12}$$

The instantaneous temperature is related to the kinetic energy *K* as follows

$$K = \frac{k_B T}{2}(3N_A + 3n \cdot N_B) \tag{3.13}$$

where $N_A$ and $N_B$ are total degrees of freedom for the non-LES atoms and LES atoms (single copy) respectively. The resulting Hamiltonian for the new LES system is

$$\begin{aligned} H = K + E &= \frac{1}{2}\sum_{A}\frac{|p_A|^2}{m_A} + \frac{1}{2}\sum_{i=1}^{n}\sum_{B}\frac{|p_{iB}|^2}{m_B} + E_{AA} + \frac{1}{n}\sum_{i=1}^{n}[E_{AB}(q_A, q_{Bi}) + E_{BB}(q_{Bi})] \\ &= \sum_{i=1}^{n}\{\frac{1}{2}\sum_{A}\frac{|p_A^*|^2}{m_A} + \frac{1}{2}\sum_{B}\frac{|p_{iB}|^2}{m_B} + E_{AA}^* + \frac{1}{n}[E_{AB}(q_A, q_{Bi}) + E_{BB}(q_{Bi})]\} \end{aligned} \tag{3.14}$$

which can be written as ***n*** full copies of the reference system by using the effective momentum $p_A^*$ and potential energy $E_{AA}^*$ for the non-LES region. We define a reference system as the single-copy system obtained by combining all of the atoms belonging to one LES copy with the non-copied atoms. Up to this point, we have described essentially a standard LES system. For LREM, we extend this treatment by allowing coupling of non-LES and each LES region into different temperature baths (as is done for standard REM). The following equation therefore holds

$$K = \frac{k_B}{2}(3N_A T_A + 3N_B \sum_{i=1}^{N} T_{Bi}) \tag{3.15}$$

where $T_A$ and $T_{Bi}$ are temperatures for the non-LES and various LES regions respectively. If we assume there is no strong coupling between the LES and non-LES atoms, this system represents a set of simulations for the full reference systems over a range of temperatures.

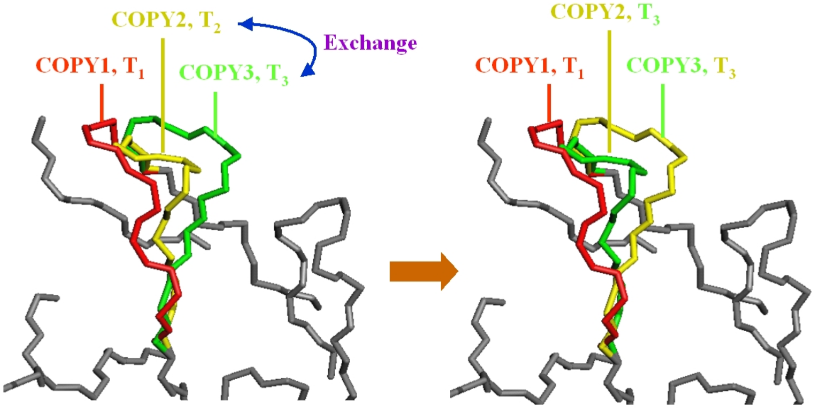


Figure 3.3 A simple model system to clarify the exchange scheme of the LREM, with one non-LES region (grey) and 3 LES copies (red, yellow and green). Each copy maintains its own temperature $T_1$, $T_2$ and $T_3$. After a predefined period of time, the target temperatures of 2 copies (in this case copy 2 and copy 3) are swapped based on a Metropolis-type criterion.

Since all the copies have the same non-LES atoms, the potential energy of non-LES region would be cancelled among copies. Similarly as in PREM, we introduce a new Hamiltonian for the new LREM system,

$$\begin{aligned} H &= \sum_{i=1}^{n} H_{LREM}(r^{[i]}, s)_{\{T_m\}} + H_{non-LES}(s)_{\{T_0\}} \\ &= \sum_{i=1}^{n} [E_{BB}(q_{Bi}) + E_{A,B}(q_A, q_{Bi})]_{\{T_m\}} + E_{AA}(q_A)_{\{T_0\}} \end{aligned} \tag{3.16}$$

where $r^{[i]}$ and $s$ stand for the coordinates and momenta of $i$th LES copy and non-LES part respectively, and we assume all copies have the same effective non-copy energy $H_{non\text{-}LES}$(s). The effective temperature $T_m$ only applies to the degree of freedom $r$ for LES part, and the non-LES part $s$ coordinates are coupled with $T_0$ instead.

Finally, we define the transition probability as the following,

$$W_{LREM}(X^{*}) = \exp\left[\sum_{i=1}^{n} H_{LREM}(r^{[i]},s)/k_B T_m\right] \tag{3.17}$$

where $X^{*} = \{\, x_1^{[1]},\ \ldots,\ x_M^{[M]} \}$ composed of only a subset of the system, stands for a "state" in this generalized ensemble. The state $X^{*}$ is specified by $M$ sets of configuration $q_B{}^{[i]}$ and momentum $p_B{}^{[i]}$ of LES atoms in replica $i$ at temperature $T_m$: $x_m^{[i]} = (p_B^{[i]}, q_B^{[i]}, T_m)$. In the similar manner, the acceptance probability is given by,

$$\rho(X \to X') = \min[1, \exp(-\Delta)] \tag{3.18}$$

where $\Delta = (\frac{1}{k_B T_m} - \frac{1}{k_B T_n})(H_{LREM}(r^{[i]},s) - H_{LREM}(r^{[j]},s))$

During the simulation, LES copies are simulated over a range of temperatures, with periodic exchanges performed using standard REM approach.

In both PREM and LREM, replica exchange is actually performed on part of the system, which reduces the number of replicas significantly. Since only degrees of freedoms of the focused or LES regions are coupled to the REM temperature space, the total number of replicas needed for two variant REM systems is much fewer than that of the original system due to the fact that the number of replicas ~ O($f^{1/2}$). For example, in a protein system of 4000 atoms where only a loop region of less than 100 atoms are of interest, then the replicas needed for PREM or LREM system should be about 6 times less than that of the original system. LREM is based on the idea of LES approximation, thus representing a further approximation beyond PREM. However, it is worthy of noting the Hamiltonian of LES system is scaled as compared to the original one, through which the energy surface is flattened to improve overall sampling, therefore it is only possible for LREM to obtain a qualitative picture of the probability distribution for any thermodynamic quantities. Otherwise, compared to PREM, LREM has several additional advantages. Firstly, since the entire replica shares the same non-copy region, this part of interactions only need to be calculated once, which greatly increases the efficiency. Another advantage of LREM is inherited from the LES method. Since the energy surface is flattened through the average among various copies. It becomes easier for each copy to overcome the energy barrier. This has significant implication in the practical use of LREM, that is, a smaller temperature range can be used with LREM than that in REM or

PREM, which will result in even less replicas in order to achieve the same sampling ability.

## 3.3 Simulation Details

We demonstrate the strength of the modified replica exchange methods by testing on the RNA tetraloop system ($G_1G_2A_3C_4[U_5U_6C_7G_8]G_9U_{10}C_{11}C_{12}$), for which structures have been determined by NMR[81,82]. This makes an excellent model due to its small size, and since several previously reported theoretical studies explored the conversion of an incorrect conformation (I) for the loop region into the correct one (C). Most importantly, we recently observed improved loop conformational sampling in our combined GB+LES method[98]. Application of the modified REM approaches to the same system will allow us to directly compare with standard REM as well as our previous GB+LES simulations.

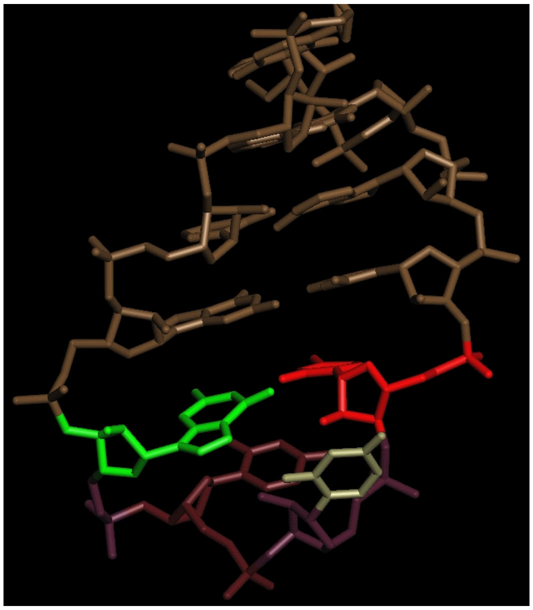


Figure 3.4 The correct structure of the RNA tetraloop derived from NMR studies, with U5 in red, G8 in green, the flexible U6 ring in grey, and the remainder of the loop region in purple.

In Figure 3.4, the correct structure of the RNA tetraloop derived from NMR studies is shown, with U5 in red, G8 in green, the flexible U6 ring in grey, and the remainder of the loop region in purple. The 12 bases of this single-stranded RNA fold back to form a double-helical stem capped by a UUCG loop. The major difference between the two experimentally determined structures of the RNA stem-loop system is the hydrogen bond pattern between the bases U5 and G8 in the tetraloop region. A bifurcated hydrogen bond is present between one of the U5 carboxyl oxygen atoms and the imino and amino groups of the G8 in the correct structure (C). In contrast, this base pair forms a reverse wobble pattern in the incorrect structure (I).

### 3.3.1 REM Setup

The standard REM simulations were run by using the REM facility implemented in AMBER(version 8)[79]. The following 8 temperatures, 266K, 282K, 300K, 318K, 338K, 359K, 381K and 405K were used. These temperatures were optimized to give a uniform and optimal exchange acceptance ratio of about 10%. The original incorrect and correct RNA tetraloop NMR structures were used as starting structures for the simulations[81,82]. The time step was 1 fs, and SHAKE was applied to all bonds involving hydrogen[86]. Berendsen temperature coupling[115] is used with a relaxation time of 0.5 $ps^{-1}$. No cutoff on nonbonded was used. The AMBER ff94 force field was used for all calculations[23]. Solvation effects were included through use of the GB continuum model as implemented in AMBER[43]. The Born radii were adopted from Bondi with modification of hydrogen[46], and the scaling factors for Born radii were taken from the TINKER modeling package[87]. Before running productions, each replica was equilibrated at its target temperature for 100 ps. The replica exchange was attempted every 500 ps, and the data was saved after each exchange for later analysis.

### 3.3.2 PREM Setup

For PREM simulation, the entire UUCG loop of RNA tetraloop was defined as the focused region and the other part of the molecule as the bath region. And the simulation was run with our modified AMBER REM module whereas five replicas can cover the similar temperature range as covered by eight replicas in standard REM. The five target temperatures used for the simulation are 266K, 295K, 327K, 363K and 403K, which were also designed to give an exchange acceptance ratio of about 10%. The bath region temperature was maintained at 300K. The same temperature coupling constant 0.5 $ps^{-1}$ was applied to both the bath and the focused regions. All of the other control parameters used by PREM are the same as described above for standard REM.

### 3.3.3 LREM Setup

For LREM simulation, firstly the AMBER module ADDLES was used to construct the LES system[79]. The entire UUCG loop of RNA tetraloop was replaced by five LES copies. All LES copies of individual atoms were initially assigned identical coordinates but unique velocities according to their target temperatures. We used temperatures of 80K, 88K, 99K, 108K, and 120K for the LES copies respectively whereas the non-LES region was maintained at 100K. Similarly, the temperatures for LES copies were optimized to give the desired exchange ratio of about 10%. LREM was run with our modified AMBER GB+LES module. And all of the other control parameters for the simulation are the same as described above.

## 3.4 Results and Discussions

In both modified approaches we tried to apply temperature coupling separately to multiple regions in one single molecule, such as the focused region and the bath region in PREM, and the non-LES region and each LES region in LREM. This could potentially generate a problem of unequal distribution of energies due to the interactions between regions with different temperatures. In order to investigate whether the target temperature could be properly maintained for each temperature subspace, we first performed an equilibration simulation on an RNA tetraloop LES system. The LES system was constructed by replacing the entire UUCG loop with five LES copies. During the simulation, the non-LES and each LES regions were coupled to different temperature baths while no exchange was attempted. In Figure 3.5, the temperature fluctuations during the simulation are shown. The left is for the non-LES atoms and the right shows temperature fluctuations for 5 LES copies. The non-LES region target temperature was assigned to 300K and the LES region target temperatures were assigned ranging from 200K to 360K, evenly distributed in space. The plot clearly shows that both the non-LES and LES regions have the ability to maintain the desired temperatures, which is essential as the first step for both modified REM approaches.

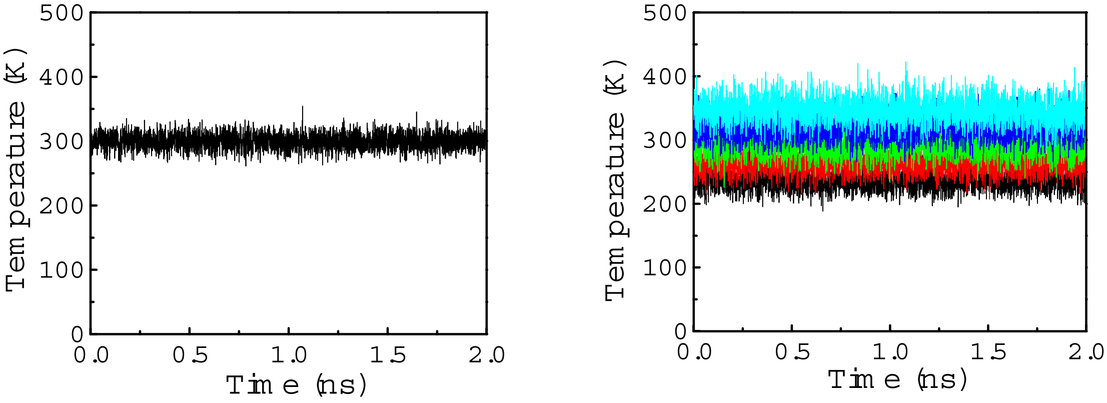


Figure 3.5 Time series of temperature fluctuations during a modified LES simulation with separate temperature baths applied to the non-LES region and each LES region. The left shows temperature fluctuation for the non-LES region and the right shows temperature fluctuations for 5 LES regions.

In Figure 3.6 we also show a desired linear relationship between average potential energy and temperature, where temperature is the target temperature for each LES copy, and potential energy is the mean potential energy of each copy taken as $\{E_{AB}(q_A, q_{Bi}) + E_{BB}(q_{Bi})\}$. It's worthy of noting that the interactions between one particular LES copy and the non-LES atoms are fully included as the potential energy for that copy. All the energetic and temperature data was extracted from the single modified LES simulation with five LES copies assigned with different target temperatures.

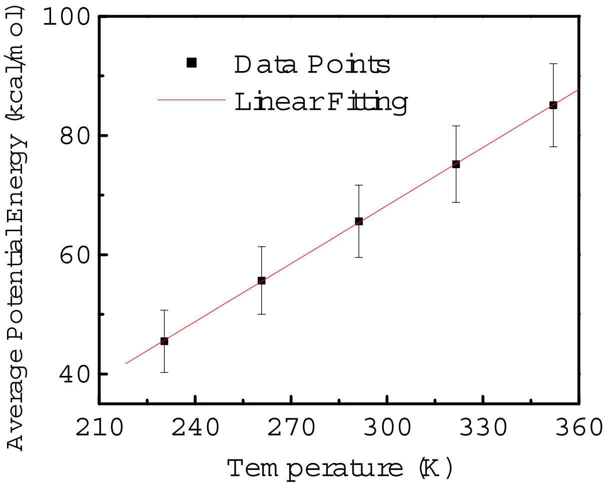


Figure 3.6 Average potential energy vs. target temperature for each copy of LES atoms during a modified LES simulation with separate temperature baths applied to the non-LES region and each LES region.

In an attempt to check whether the replica exchange simulation indeed performed properly, we monitored potential energies and temperatures of replicas during the simulations. Firstly, in order to have sufficient replica exchanges between neighboring replicas, the distribution probability of potential energy should have enough overlap. As shown in Figure 3.7, the distribution probability of potential energy for each target temperature appears to have the desired behavior with higher mean value and broader distribution for higher target temperature, and moreover there is enough overlap between each adjacent temperature pair, which will allow sufficient exchange to occur between replicas.

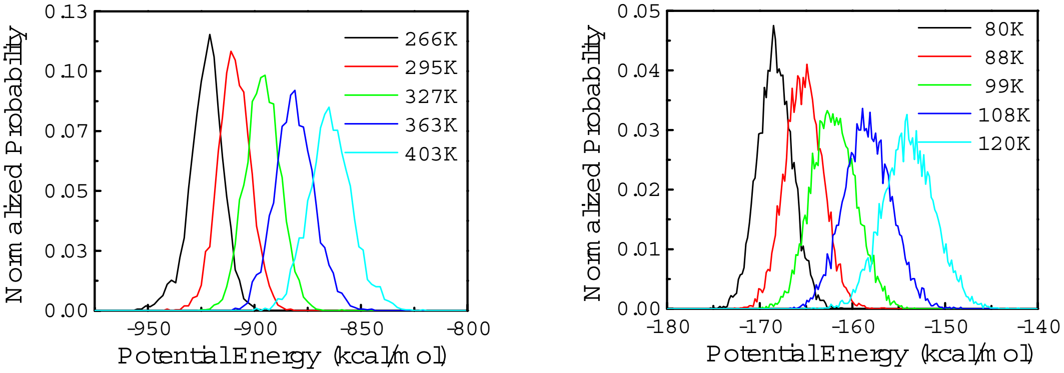


Figure 3.7 The probability distribution of potential energy for each target temperature. The left shows the distribution sampled from PREM simulation; the right from LREM simulation.

The left plots of Figure 3.8 and 3.9 show the time series of replicas jumping for one of the target temperature; and the right shows the time series of target temperature evolving for one of the replicas. We do observe free random walk in the temperature space in both PREM and LREM simulations. This confirms that both modified REM simulations are performed properly, since efficient temperature exchange is indispensable for the molecule to escape the local energy trap so as to enhance the overall sampling ability.

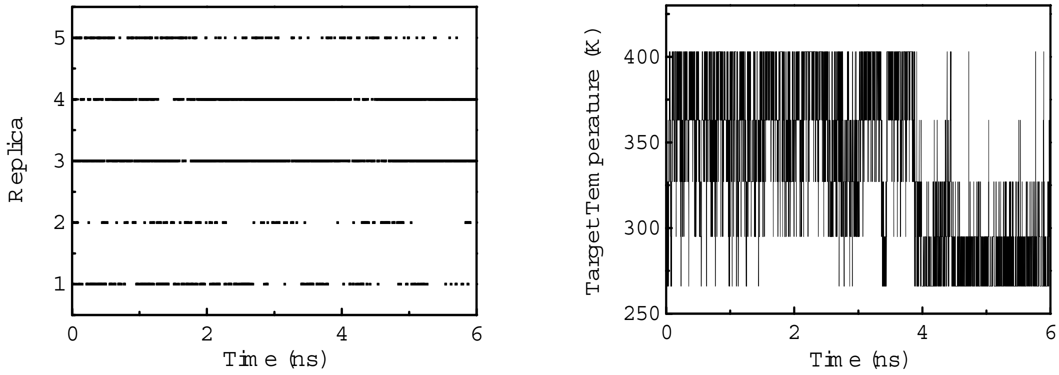


Figure 3.8 Time series of replicas at 295K (left) and target temperatures of replica 2 (right) for PREM simulations of RNA tetraloop.

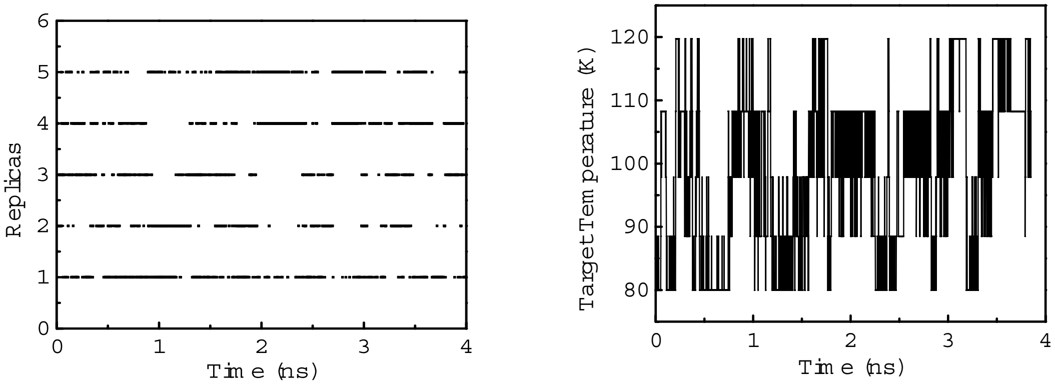


Figure 3.9 Time series of replicas at 99K (left) and target temperatures of replica 2 (right) for LREM simulations of RNA tetraloop.

The following Table 3.1 and 3.2 give the acceptance ratio during PREM and LREM simulations respectively. The overall acceptance ratio was designed to be around 0.1, which would be appropriate to keep the balance between as many as possible exchanges and enough relaxation time upon exchanges. In the current case the exchange is attempted every 500 steps of MD simulation. Each adjacent temperature pair has a chance to swap their target temperatures and the velocities based on the Metropolis-type criterion. These exchange ratio values are consistent with the observations in the above plots; we do have uniform and large enough exchanges (all about 10%) occurred in the simulations.

Table 3.1 Acceptance ratios for the PREM simulation of RNA tetraloop

| Temperature Pairs | Exchange Acceptance Ratio |
|---|---|
| 266K ↔ 295K | 0.11 |
| 295K ↔ 327K | 0.11 |
| 327K ↔363K | 0.11 |
| 363K ↔403K | 0.12 |

Table 3.2 Acceptance ratios for the LREM simulation of RNA tetraloop

| Temperature Pairs | Exchange Acceptance Ratio |
|---|---|
| 80K↔ 88K | 0.10 |
| 88K ↔ 99K | 0.13 |
| 99K ↔108K | 0.13 |
| 108K ↔120K | 0.10 |

Table 3.3 REM related variables for various RNA tetraloop systems

| System | Degree of freedom | $\bar{E}_{pot}/T$ | $\delta^2T$ at 300K | $\delta^2E_{pot}$ at 300K | $N_{replica}$ 200-500K | Accept ratio |
|---|---|---|---|---|---|---|
| RNA 2b (GB) | 192 | 0.18 | 29.44 | 3.94 | 3 | 0.12 |
| RNA 4b (GB) | 375 | 0.35 | 22.06 | 6.73 | 5 | 0.11 |
| RNA (GB) | 1146 | 1.13 | 9.28 | 12.98 | 8 | 0.13 |
| RNA (in Water) | 27432 | 26.82 | 2.57 | 69.61 | 38 | - |

* All example systems are built on RNA tetraloop with different methods. *RNA 2b* stands for 2 nucleotides as focused region for PREM simulation; *RNA 4b* stands for 4 nucleotides are focused for PREM simulation; *RNA* stands for the whole RNA tetraloop for standard REM simulation; *RNA in water* stands for the RNA solvated with explicit water for REM simulation.

- The dash lines in *Accept ratio* column for *RNA in water* means data not available because the systems are too large to run the test with our current computer facility.

Since only degrees of freedom of part of the system are coupled to the effective temperature space in both modified REM approaches, the total number of replicas needed would be much fewer than that of the conventional REM. We have completed several test simulations of the RNA tetraloop with the standard REM and our modified REM approaches. Results list in the following Table 3.3 clearly show that our modified REM methods are much more efficient to cover the same temperature range of about 266-403K with the similar exchange ratio of about 0.10 as compared to conventional REM method. For example, it requires about 8 replicas to run the whole RNA tetraloop with GB continuum solvation model using the standard REM method and many more required for the RNA in explicit water. However, the number of replicas will be reduced to 5 if replica exchange is only focused on four nucleotide of the UUCG loop by using our modified PREM method.

Table 3.4 Comparison of most populated structure sampled in various simulations

| System | NMR correct | MD/water | MD/GB | REM | PREM | LREM |
|---|---|---|---|---|---|---|
| RMSD (Å) | - | 0.9 | 1.0 | 1.0 | 1.0 | 1.1 |
| O2-H1Distance (Å) | 2.7 | 2.8 | 2.9 | 2.9 | 2.9 | 3.0 |
| O2-H21 Distance (Å) | 3.7 | 3.7 | 3.4 | 3.4 | 3.4 | 4.0 |
| U5 $\chi$ angle (Degree) | -157 | -161 | -153 | -152 | -150 | -144 |
| G8-C7 stacking (Å) | 5.9 | 6.2 | 6.7 | 6.9 | 7.2 | 7.8 |

* U5 $\chi$ angle is the glycosidic angle defined as the torsion angle formed by O4'-C1'-N-C2; G8-C7 stacking is calculated as the distance between mass centers of the two bases.

To verify that our modified REM approaches lead to the same conformation prediction as compared to conventional REM method, we have examined all the most populated structures sampled from various simulations started from the I conformation. The detailed results are shown in Table 3.4. Experimentally, the single-stranded RNA was found to have a well-defined hairpin structure with a double-helical stem capped by a UUCG loop. A bifurcated hydrogen bond between U5 and G8 is present in the loop region. All the REM simulations started from the incorrect structure (I). Overall, the most populated conformations reached by various simulations show good agreement with the experimentally determined native structure (C). The similarity between structures is measured by the heavy atom root mean square deviation (RMSD), several featured hydrogen bond distances along with two other geometry variables. All RMSD calculations include non-hydrogen atoms in the UUCG tetraloop (residues 5-8) except the base atoms of U6, which does not form specific contacts and shows higher mobility. This RMSD selection was chosen to be consistent with the previous theoretical studies of this system. The above results clearly confirmed that the same global minimum could be reached by PREM method. And moreover, the global energy minimum is not changed either in the LREM approach using a mean field potential.

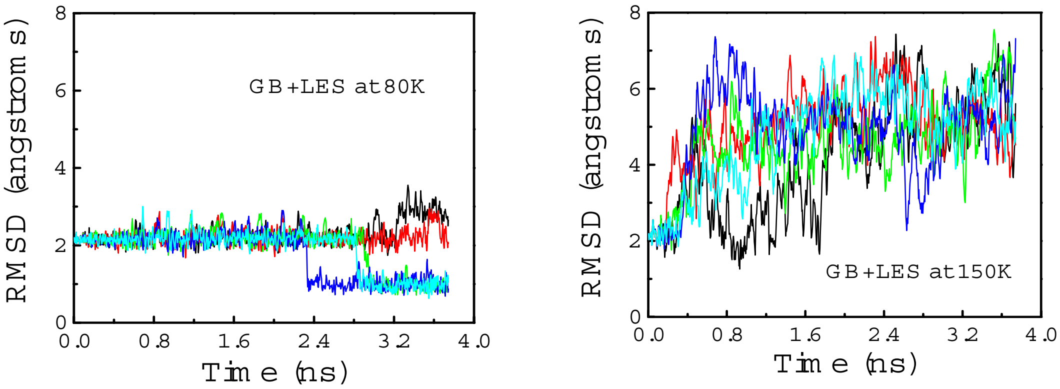


Figure 3.10 The loop RMSD compared to the correct structure as a function of time for two standard LES simulations, at 80K (Left) and at 150K (Right).

Since LREM is built on top of LES topology and designed to overcome its temperature problem, it would be useful to compare these two methods directly. For this purpose, we initiated two GB+LES simulations from the incorrect structure, with all 5 LES copies having identical initial coordinates. As shown in Figure 3.10, at 80K the loop appears to be rather stable as monitored by RMSD. In about 2.4ns one of the five copies converted to the correct structure. However, this entire I→C transition is very slow and even until 4ns there are two copies still staying in the original incorrect basin. Then we elevated the temperature to about 150K for all LES copies. In this case the simulation resulted in fully extended conformation of RNA within 600ps. The RMSD rose to 4.0 Å and all base pairs were lost during the simulation. Similarly undesirable results were obtained when starting from the correct conformation. The possible explanation for this is that this might arise from too great a weakening of the Watson-Crick hydrogen bonds due to the scaling of partial charges and Lennard-Jones well depth parameters. It has also been shown that the behavior of the LES system corresponds to a non-LES system of higher temperature. Moreover, lack of solvent friction likely makes the dynamic behavior of the RNA more sensitive than with LES in explicit solvent. Therefore, the choice of appropriate temperature for LES copies in the GB+LES simulations has been a major difficulty of applying this efficient mean field method to many other systems.

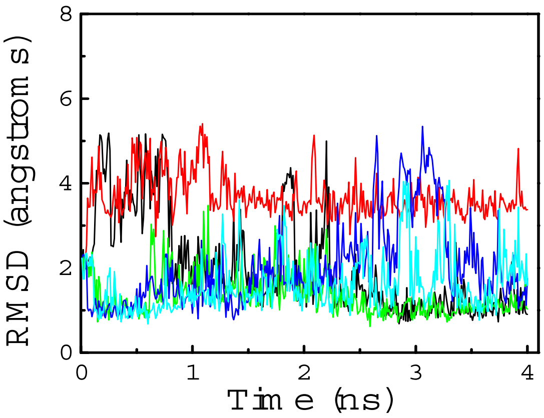
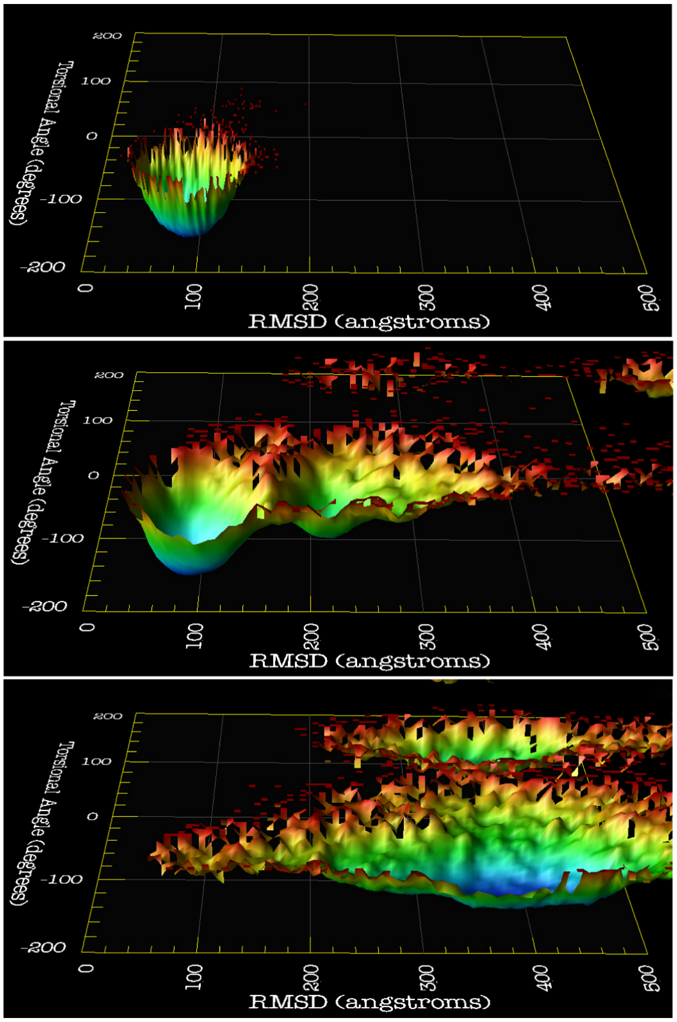


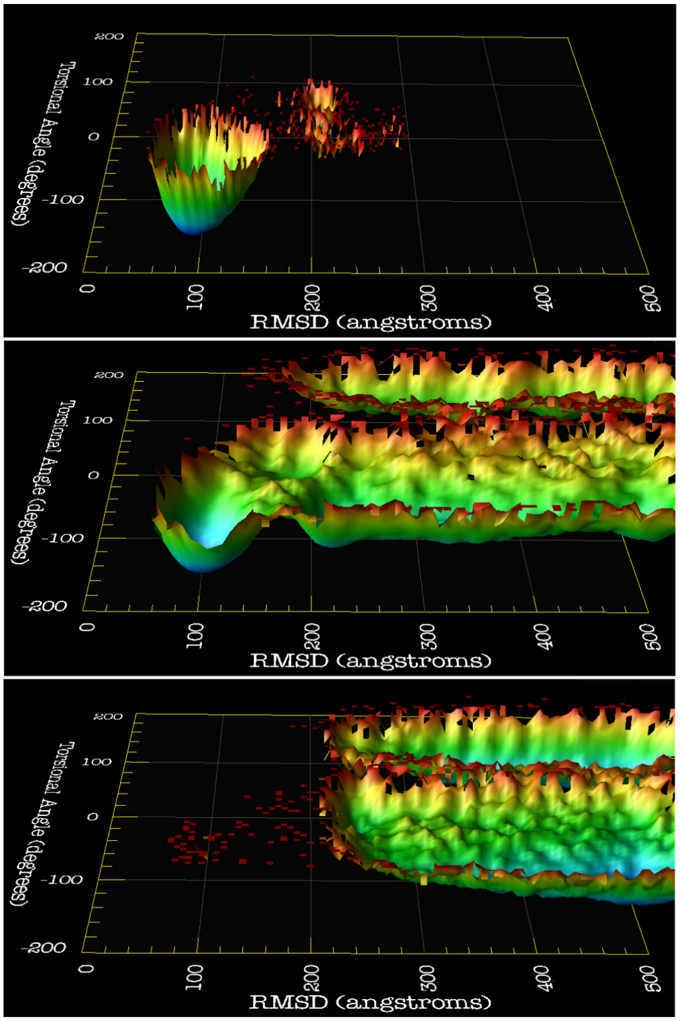
Figure 3.11 The loop RMSD compared to the correct conformation as a function of time for LREM simulation; each copy (replica) is shown in different color.

The sensitivity to the temperature in GB+LES simulations may get the remedy through the use of the LREM approach. As described above, one of the advantages of the LREM is not to be hampered by the local minima problem since the low temperature copies can escape the trap by exchanging with high temperature copies; on the other hand, the high temperature copies have chances to relax the structure in a way similar to the temperature annealing method. In addition, a range of temperatures used in the LREM can guarantee to generate a population of structures where not only the most stable (correct) structure are sampled but also other meta-stable structures are sampled as well just with a relatively lower probability. We therefore investigated whether our approach would be successful in this aspect. In Figure 3.11 we show the RMSD value for each of the copies as a function of time during the 4 ns LREM simulation at temperatures ranging from 80K to 120K. It is apparent that each copy undergoes a much quicker I→C transition. In contrast to the simulation with GB+LES at 130K using 3 copies, not all the copies finally convert to single correct conformation, instead there are multiple transitions between the correct and other structures during the entire simulation, thus possibly giving a temperature-dependent equilibrium distribution of conformational probability.

It has been shown above the sampling efficiency has been greatly enhanced in our modified approaches. At this point, it is necessary to ask to what extent the results from these modified approaches resemble those from standard REM approaches. Since the canonical ensemble properties are maintained in REM, the free energy landscape can be constructed through all saved structures according to the appropriate reaction coordinates. Approximately we could do the same analysis on the data obtained from the modified approaches; nevertheless, one should keep it in mind although the correct thermodynamic properties can be reproduced from PREM approach under the weak coupling assumption, the resulting free energy landscape constructed from the LREM method is absolutely not the same as that of the original system due to the modification of the potential function. But if the original landscape has very simple topology, for instance with only very few deep energy minima, it might be possible for LREM to generate a free energy landscape, which reflects the general feature of the real one.

To compare our simulation results with standard REM, we now calculate the free energy landscape using the loop region RMSD (as compared to the correct structure) and U5 glycosidic angle as two reaction coordinates. During the simulations we saved about 60,000 structures for each replica and removed the first 2000 to minimize the influence of the starting structures. The RMSD of each structure compared to correct structure and U5 glycosidic angle were calculated and then histogramed, and the free energy was calculated as $G = -k_B T ln\{\mathrm{P}(RMSD,\chi_{U5})\}$, where $\mathrm{P}(RMSD,\chi_{U5})$ is the probability distribution function for each pair of RMSD and torsion angle values.




Torsional Angle (degrees)
200
100
0
-100
-200
0
100
200
300
400
500
RMSD (angstroms)
Torsional Angle (degrees)
200
100
0
-100
-200
0
100
200
300
400
500
RMSD (angstroms)
Torsional Angle (degrees)
200
100
0
-100
-200
0
100
200
300
400
500
RMSD (angstroms)

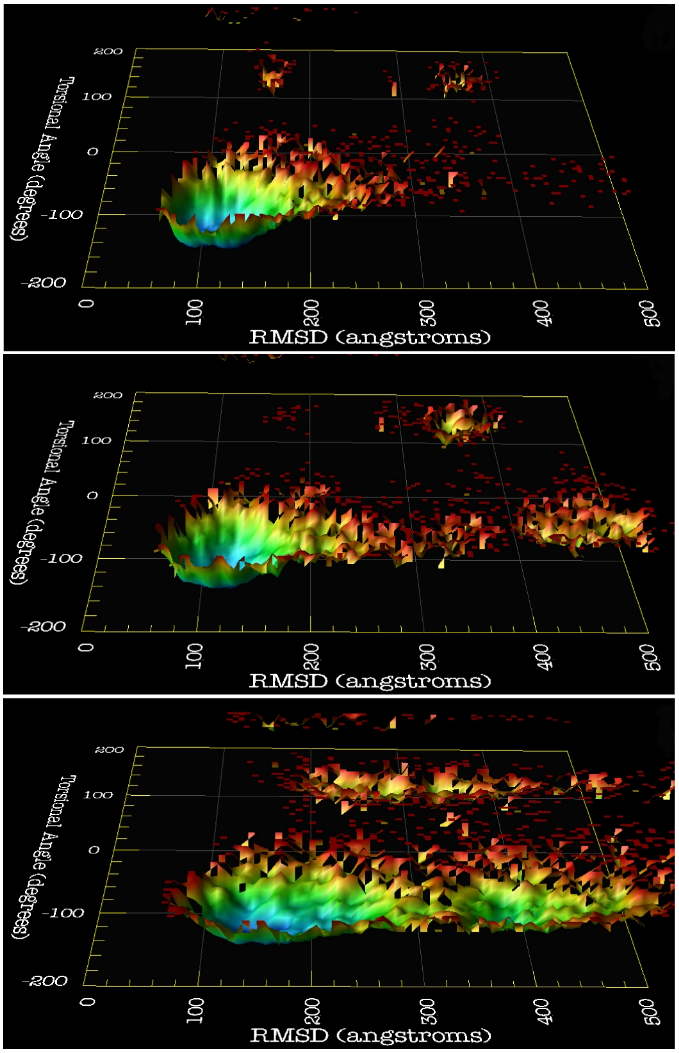


Figure 3.12 The free energy landscape of RNA tetraloop constructed from various REM simulations using the loop region RMSD (compared to the correct structure) and U5 glycosidic angles ($\chi_{U5}$) as two reaction coordinates. (Top): standard REM; (Middle): PREM; (Bottom): LREM. (*x* axis is from 0.00-5.00Å).

The free energy maps calculated from PREM and LREM (middle and bottom columns) are compared with those from REM (on the top) in Figure 3.12. The convergence of the data has been tested using methods we mentioned in the following Practical Consideration section. Different rows are from different temperatures. For REM, three selected temperatures are 300K, 338K and 405K; For PREM the temperatures are 295K, 327K and 403K; and for LREM, the temperatures are 80K, 99K and 120K. All the free energy surfaces show very similar pattern. At the lowest temperature only one deep minimum is observed in all REM methods, which corresponds to the NMR determined correct structure (C). At the middle temperatures, PREM map appears more similar to REM map. There are two major minima at the PREM energy map as in the REM map. Although the correct structure family is highly dominant, the incorrect structure is only a meta-stable structure corresponding to a much shallow minimum in the energy map. However, the second minimum disappears from the LREM map. And moreover the landscape is much more flattened as compared to REM and PREM maps. These results are not totally surprising given mean field approximation employed by LREM. The goal of LREM is not to faithfully reproduce the thermodynamic properties but to efficiently simulate the mean behavior of the system. It has been shown that every local energy minimum on the potential surface of the real system is also a local energy minimum on the LES potential surface[52]. And the copies with $n_{copy}/n$ (n is the total number of LES copies and $n_{copy}$ is the number of LES copies occupying one particular energy minimum) near unity will be the closest to the real minimum in the original system whereas the copies with low values of $n_{copy}/n$ may deviate from the real minimum[53]. For LREM, the correct structure, which is actually the global minimum in the energy surface, has much greater value of $n_{copy}/n$ than that of the incorrect structure. Therefore, LREM does predict the correct structure but the second minimum is blurred from the LREM surface.

We now turn to the seemly inconsistent occurrence of the incorrect and correct structures in the energy surfaces. Previous theoretical studies have shown that the incorrect structure (I) is stable during the explicit solvent simulation and can last several ns in the GB simulations. However, all of REM approaches can barely explore the incorrect structure basin at the low temperature as shown in the free energy maps, although at higher temperature sampling the incorrect structure becomes feasible for REM and PREM simulations. Perhaps two facts could lead to the above observations: firstly, the observed incorrect conformation is simply kinetically trapped and it has to cross a high energetic barrier in order to reach the correct structure especially in water simulations as reported previously; secondly, the free energy difference is relatively high between the C and I structures. The large free energy difference of about 7.5 kcal/mol as estimated by MMGB method explains why the incorrect conformation is not highly populated at low temperature REM and PREM simulations. Moreover, once the C structure is reached, it will be rare chance of going back to that I structure again. The reason for the high barrier of conformational change between correct and incorrect structures in water simulation is due to the hydrogen bond formed by 2' hydroxyl groups of the ribose. The hydrogen bond has been reported before to be primarily responsible for the unusual stability of the loop conformation. In all single-copy MD simulations 2' hydroxyl groups of U5 forms hydrogen bond to backbone 5' oxygen of U6[74]. The hydrogen bond together with a reverse wobble base pair pattern between U5 and G8 have to be broken simultaneously to form the correct structure. The GB and LES simulations somehow weakened the hydrogen bonding interactions. That also explains why the I→C transition is achieved in GB or LES simulations while not achieved in the explicit solvent single-copy simulation.

We further demonstrate that the resulting free energy landscape from PREM is consistent with our previous normal MD simulations. In our previous standard GB simulations at 300K, three different I→C pathways were observed[98]. In Figure 3.13 we show the free energy surface calculated from the PREM data, with white spheres showing the sampling of this landscape during simulations representing the three transition pathways.  The first two pathways appear to be similar and just sample structures in the vicinity of the incorrect (I) and correct (C) structures. The transitions are through the direct crossing of the barrier between two minima. However, upon careful examination, the two pathways are found to cross the barrier at different locations. The first pathway involves slightly flipping-out of U5 base and rotation of its glycosidic bond, and meanwhile converts to the correct structure while the second pathways only involves minor reorganization in the loop region. On the basis of PREM energy landscape, we estimate the free energy barrier for the first path involving significant rotation of glycosidic bond to be about 3 kcal/mol, while that for the second path to be 4 kcal/mol or so. The third pathway is apparently different to the first two and samples much broader region on the free energy landscape. At an early point it starts to show great deviation from both I and C structures and then after a while it comes back to the correct structure basin which makes much sense by looking at the energy map. Actually the third pathway has been exploring other highly populated regions as sampled by PREM method. It reveals the interconvertion somehow involves first partially unfolding of the incorrect RNA structure, and then refolds to the correct one.

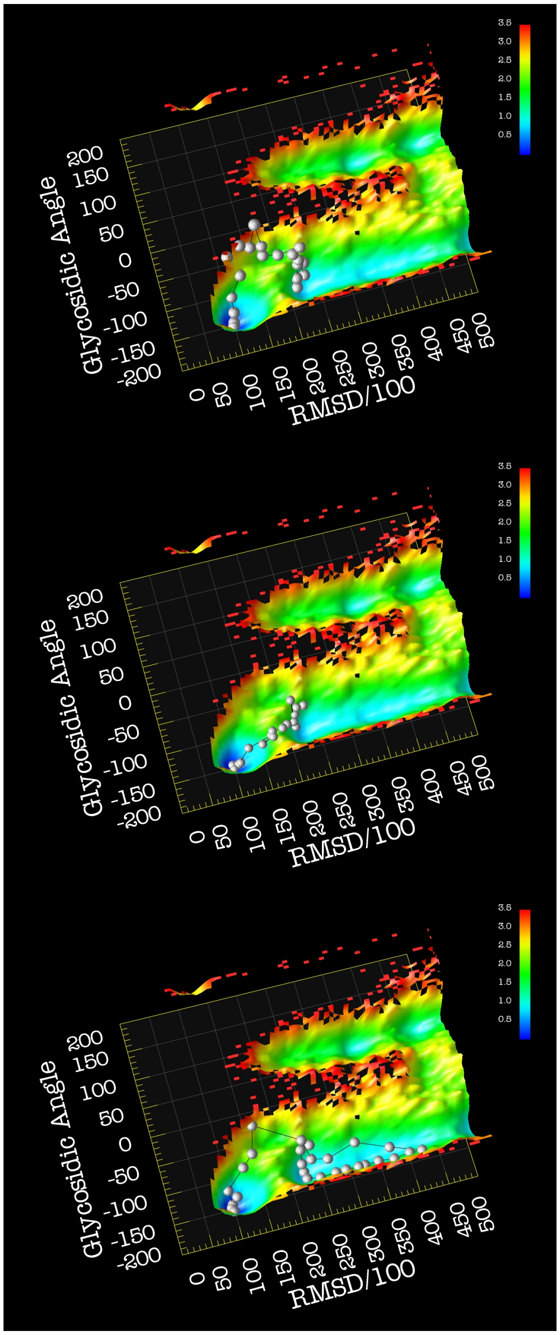


Figure 3.13 Projections of three MD *I*→*C* transition pathways at 300K onto the free energy surface constructed from PREM simulations using the loop region RMSD (compared to the correct structure) and U5 glycosidic angles ($\chi_{U5}$) as two reaction

coordinates. (Top): base flipping-out pathway; (Middle): minor reorganization pathway; (Bottom): partially unfold then refold pathway.

## 3.5 Conclusions

Replica exchange method has been demonstrated as a powerful tool to sample thermodynamic equilibrium quantities of many non-trivial systems. With the recent implementation of REM in AMBER[79], the use of this method has gained great popularity in peptide and small protein folding simulations. However, one of the problems that prevent the method from extending to even bigger system is due to the fact that when the system gets bigger, the number of replicas grows proportionally as $f^{1/2}$. To address this difficulty, we introduced herein two variant replica exchange methods called PREM and LREM, which employ a similar strategy to reduce the number of replicas needed for the simulation. When using these two methods, one source of primary concerns is to what extent two parts of the system can be decoupled to separate temperature baths. An important assumption used in the both modified REM methods is that the bath or non-replicated region is not significantly affected by its interactions with the other part of the system. In other words, the bath or non-replicated region should be relatively rigid and conformational variations of replicas are not strongly dependent on the remainder of the system. This is generally a reasonable approximation since we always choose the flexible part of the system to be replaced by the multiple copies. However, the coupling does exist between the replicated region and non-replicated region. Therefore whether this method holds true or not strongly depends on how strong the coupling is. The second concern is the mean field approximation employed by LREM. In spite of these potential concerns, the variant REM methods are expected to be useful tools for many applications. One of tasks for which it might be particularly well suited is to use these methods for loop structure refinement simulations, which is actually the primary driving force so that we developed these methods. Another interesting application of these methods could be ligand docking, where ligand can be simulated over a range of temperatures while the receptor will be maintained at a single physiological temperature.

## 3.6 Some Practical Considerations When Running a REM Simulation

### 3.6.1 How to determine the optimal temperature distribution in REM

Clearly, the proper choice of temperatures for REM simulation is important. As has been shown before, the exchange probability in each REM step is given by[58],

$$\frac{\rho(X \to X')}{\rho(X' \to X)} = \exp\{-(\frac{1}{k_B T_n} - \frac{1}{k_B T_{n+1}})(E(q^{[i]}) - E(q^{[i+1]}))\} \qquad (3.19)$$

where $T_n$ and $T_{n+1}$ are the target temperatures for two neighboring replicas attempting to exchange, $E(q^{[i]})$ and $E(q^{[i+1]})$ are the corresponding instantaneous potential energies respectively.

Therefore, the overall acceptance ratio $P_{acc}$ can be calculated as the following,

$$P_{acc} = <\exp\{-(\frac{1}{k_B T_n} - \frac{1}{k_B T_{n+1}})(E(q^{[i]}) - E(q^{[i+1]}))\}> \qquad (3.20)$$

In the NVT ensemble, the potential energy $E$ is usually very sharply peaked around the mean value $\overline{E}(T)$. The fluctuation of $E$ can be calculated easily, for instance in the case of a system of $N$ atoms[116],

$$\left\langle \delta E^2 \right\rangle_{NVT} = k_B T^2 (C_V - \frac{3}{2} N k_B) \qquad (3.21)$$

For convenience, we define $F(E') = -(\frac{1}{k_B T_n} - \frac{1}{k_B T_{n+1}})(E(q^{[i]}) - E(q^{[i+1]}))$ (3.23)

where $E' = E(q^{[i]}) - E(q^{[i+1]})$, and equation 3.20 can be expressed as,

$$P_{acc} = <\exp(F(E'))> \qquad (3.24)$$

In the thermodynamic limit of infinite system size, we obtain simply, $E = \overline{E}(T)$, which will lead to

$$<\exp(F(E'))> = \exp(F(\overline{E'})) = \exp(F(\overline{E}_n - \overline{E}_{n+1})) \qquad (3.25)$$

The situation for a finite system, equation 3.24 can be expanded about the mean value of $\overline{E}(T)$.

If we write $E = \overline{E}(T) + \delta E$, then we obtain,

$$<\exp(F(E'))>=\exp(F(\overline{E'}))+\frac{1}{2}(\frac{\partial^2}{\partial E^2}\exp(F(E')))_{E=\overline{E}}\,\delta E'^2+..... \qquad (3.26)$$

The first term $\exp(F(\overline{E'}))$ corresponds to $\exp\{-(\frac{1}{k_B T_n}-\frac{1}{k_B T_{n+1}})(\overline{E}_n-\overline{E}_{n+1})\}$, and the second correction term, which is proportional to the mean square fluctuation of potential energy *E*. For any sizable system, $\delta E^2$ is expected to be relatively small. Therefore, as a first-order approximation, equation 3.26 can be conveniently written in terms of the mean average potential energy $\overline{E}$ as the following,

$$P_{acc} = \exp\{-(\frac{1}{k_B T_n}-\frac{1}{k_B T_{n+1}})(\overline{E}_n-\overline{E}_{n+1})\} \qquad (3.27)$$

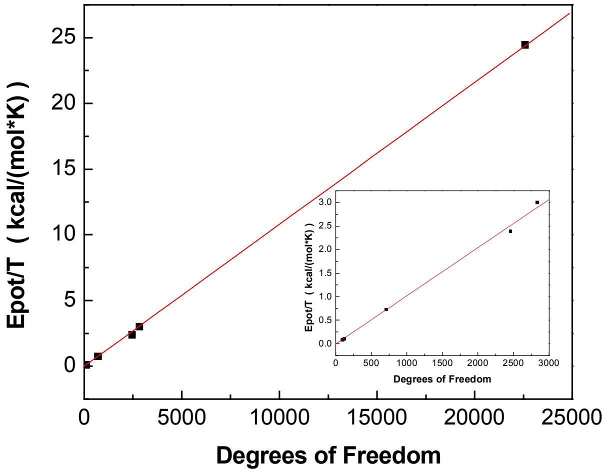


Figure 3.14 The ratio of potential energy vs. temperature as a function of degrees of freedom. The ratio is calculated from best-fitting of mean potential energies vs. target temperatures in MD simulations for systems of different size. Shown as an inset is the amplified region with number of degrees of freedom below 3000.

In the NVT ensemble, under the assumption of no phase transition, the mean potential energy $\overline{E}$ has the following relationship with temperature $T$[117]

$$\overline{E} = (rept)\ T = 1/2\, f k_B T \qquad (3.28)$$

where *rept* stands for the slop of mean potential energy versus target temperature; *f* is the number of degrees of freedom of the system and $k_B$ is Boltzmann factor.

The validity of the relationship given in equation 3.28 is demonstrated in Figure 3.14. Indeed, the temperature dependence of average potential energy gives very good straight line for many systems over a range of temperatures from 200K to 400K. And in most cases, the correlation coefficients are as good as 0.99. Further, we plot the ratio obtained from the above fittings vs. the number of degrees of freedoms in Figure 3.14. Apparently, it also gives a perfect straight line with a correlation coefficient of 0.99 for systems ranging from several tens of atoms to ten thousands of atoms; this confirms the relationship presented in equation 3.28 is qualitatively accurate enough for our following derivation.

If insert the relationship of equation 3.28 to equation 3.27, gives

$$P_{acc} = C(f) \cdot \exp\{-(\frac{1}{T_n} - \frac{1}{T_{n+1}})(T_n - T_{n+1})\} \qquad (3.29)$$

where *C(f)* is only the function of degrees of freedom *f*. For the exchange ratio to be evenly spaced between any neighboring temperature pairs, it requires $(\frac{1}{T_n} - \frac{1}{T_{n+1}})(T_n - T_{n+1})$ MUST not be related to any particular $T_n$

Assuming $T_{n+1} = xT_n$ (this happens to give an exponential temperature distribution) would be a straightforward choice. Plugging $T_{n+1} = xT_n$ into equation 3.29 would give us,

$$P_{acc} = \exp\{(-\frac{rept}{k_B})\frac{(1-x)^2}{x}\} \qquad (3.30)$$

Solving the equation 3.30, will lead to

$$x = 1.0 + 0.5 * \{k_B \ln(P_{acc})/rept + \sqrt{[2.0 - k_B \ln(P_{acc})/rept]^2 - 4.0}\} \qquad (3.31)$$

Let's consider, since $k_B \ln(P_{acc})/rept = 2\ln P_{acc}/(fT)$ is usually less than $10^{-5}$, a very useful result can be derived as shown in the following, which states the relationship between the number of replicas *m* and the number of degrees of freedom *f* of the system.

$$\Delta T \sim (x-1) \sim 1/\sqrt{\text{rept}} \sim 1/\sqrt{f} \qquad (3.32)$$

Hence, in order to cover a temperature range of $T_{max}$ -$T_{min}$ ~ $(m-1)\Delta T$ ~ $(m-1)/\sqrt{\text{f}}$ . This suggests that the number of replicas is proportional to the square root of the number of degrees of freedom for the system.

Based on the above discussion, a typical procedure for establishing a series of replica temperatures before running a REM simulation would be as follows. First, initiate multiple MD simulations (4 simulations are usually good to start with) at different temperatures; then calculate the mean potential energy for each MD simulation and obtain the best-fit slope of potential energy vs. target temperature; finally given the predefined acceptance ratio $P_{acc}$, plug the slope into equation 3.31, one will be able to obtain a series of temperatures.

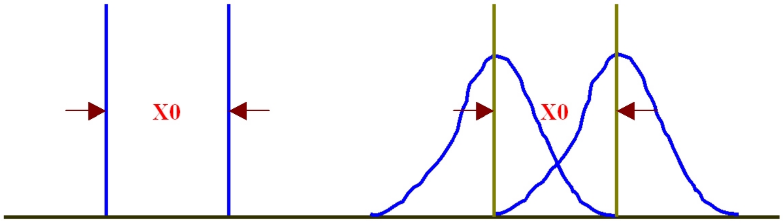


Figure 3.15 Instantaneous potential energy *E* distributions. (Left) no fluctuations; (right) *E* fluctuates

In reality, one always observes much greater exchange ratio than what has been designed to. This is especially true for small system with only hundreds of or fewer atoms. Now for explanation we turn to equation 3.20 and 3.27. The approximation used in equation 3.27 is actually quite a big step. It will only hold true under limiting conditions. Let's look at the acceptance exchange ratio according to equation 3.20 and 3.27 respectively; these two scenarios can be well illustrated as in the Figure 3.15.

The designed exchange ratio $P_{acc}$ is $\exp(x_0)$. If the fluctuations are introduced and assuming they obey a Gaussian distribution, the actual exchange ratio $P_{acc}$ would be $\iint dx_1 dx_2 \exp\{(x_1 - x_2) e^{-ax_1^2} e^{-a(x_2-x_0)^2}\}$, which is usually much greater than $\exp(x_0)$.

Additionally, we give some numerical results to justify the above arguments. Table 3.5 shows some quantities calculated from four independent MD simulations at four different temperatures respectively. The numbers given in the first two columns is by post-processing potential energy snapshots taken from each pair of trajectories.

The simulated exchange ratio is calculated as follows: take two independent trajectories, and collect the instantaneous potential energy and calculate

$$\Delta = (\frac{1}{k_B T_n} - \frac{1}{k_B T_{n+1}})(E(q^{[i]}) - E(q^{[i+1]}))$$

$\Delta \leq 0$, accept the exchange

$\Delta > 0$, accept the exchange if random(0,1) $\leq$ exp($-\Delta$)

reject the exchange if random(0,1) $>$ exp($-\Delta$)

As matter of the fact, no real exchange is performed during the calculation. The exchange ratio for REM as shown in the last column is drawn from real REM simulation using the same set of 4 target temperatures. It is evident that the first column data is close to target exchange ratio 0.15 while the corresponding values are much larger in the second column, which corresponds to the exchange ratio of an unperturbed REM system. In principle the second and the third columns should give close numbers. However, the small deviation here is probably due to the poor convergence of the sampling.

Table 3.5 Quantities calculated from four independent MD simulations

| | expr1[a] | expr2[b] | Exchange ratio | |
|---|---|---|---|---|
| | | | Simulated[c] | REM |
| 267→283 | 0.1626 | 0.4079 | 27.5 | 33.2 |
| 283→300 | 0.1775 | 0.4750 | 30.8 | 32.4 |
| 300→318 | 0.1920 | 0.5688 | 32.4 | 31.1 |

[a] expr1=$\exp[-(\frac{1}{k_B T_n}-\frac{1}{k_B T_{n+1}})(\overline{E}_n-\overline{E}_{n+1})]$; [b] expr2 =$<\exp[-(\frac{1}{k_B T_n}-\frac{1}{k_B T_{n+1}})(E_n-E_{n+1})]>$; [c] see text

In conclusion, the exchange ratio calculated using equation 3.27 is always higher than what is designed to (the smaller the system, the greater deviation will be seen). As has been discussed above, it is almost impossible to give an exact solution to this problem. Many factors are involved in determining the overall exchange ratio, such as the size of system, temperature-coupling method and among many others. Fortunately, the exact exchange ratio is not very important. The only thing one shall make sure is that the exchange ratio lies within his desired range. Otherwise a simplest solution to this would be to scale down the expected exchange ratio accordingly. As an example, test was run for a system of about 400 atoms, if you want an exchange ratio of 10% then you just plug 5% instead of 10% into the right side of equation 3.31. In the end, you will obtain an overall exchange ratio of about 9.3%.

### 3.6.2 How to determine the optimal equilibration time between exchanges

In a number of recent REM simulations, how often to attempt the exchange appears to be a somewhat arbitrary choice. Several published values of the exchange frequencies can be found, such as 0.375 ps for a 46-residue three-helix bundle[118], 5 ps for a 20-residue Trp-cage miniprotein[119], 0.01ps for 5-residue peptide[58].

If the exchange is attempted more frequently (fewer steps between exchanges), the replicas will be more mobile as they are more likely to change temperature. More frequent exchanges therefore mean replicas can be spaced further apart, for the same level of mobility. This saves on computational expense. On the other hand, if the exchange occurs too frequently, then the temperature of the system will not be equilibrated, thus leading to errors when one wants to compute equilibrium thermodynamics on the structures sampled under this condition.

The equilibration time depends on the temperature distribution, system size, and thermostat method etc. Importantly, even if the velocities are rescaled after every exchange, thus keeping the temperature perfect, the potential energies still take time to match that of the new temperature. The equilibration of the system can be monitored by watching how quickly the thermostat method equilibrates to the new desired temperature.

In the following, we present two methods to quantitatively address the question of how fast each replica takes to reach the equilibrium. In one of the methods, we use force metrics to explore the thermostat method and watch the equal partition of the kinetic energies after a temperature jump; in the second method, the effect on the overall exchange ratio is evaluated by varying the equilibration time.

### 1) Force metric criterion

Firstly, we estimated the kinetic energy fluctuation metric introduced by Mountain and Thirumalai[89-91]. The kinetic energy of each atom is

$$F_i(t) = \frac{1}{2} m_i v_i^2 \qquad (3.33)$$

and one can define

$$f_i(t) = \frac{1}{t} \int_0^t F_i(s) ds \qquad (3.34)$$

$$f(t) = \frac{1}{N}\sum_{i=1}^{N} f_i(t) \tag{3.35}$$

and then the kinetic energy fluctuation metric is defined as

$$\Omega(t) = \frac{1}{N}\sum_{i=1}^{N} [f_i(t) - f(t)]^2 \tag{3.36}$$

If the system is ergodic, $\Omega(t)$ approaches 0 in the limit of long time. It has been found that the time dependence of $\Omega(t)$ is a diffusion-like equation at the long time limit,

$$\Omega(0)/\Omega\,(t) \approx D_{KE}\,t \tag{3.37}$$

where $D_{KE}$ is referred as the generalized diffusion constant, and $D_{KE}^{-1}$ indicates the approximate time needed for adequate sampling to reach equilibrium.

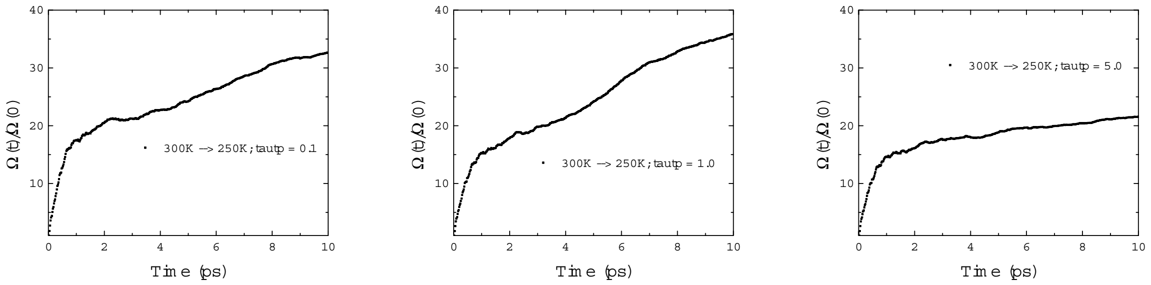


Figure 3.16. Plots of the normalized kinetic energy fluctuation metric, $\Omega_{FN}(0)/\ \Omega_{FN}(t)$, as a function of time for Trp-cage simulation after a temperature jump from 300K to 250K.

Table 3.6 Kinetic energy metrics relaxation time for different coupling constants in a Trp-cage temperature jump simulation.

| | Coupling Constant | Relaxation Time $D_{KE}$ |
|---|---|---|
| | ($ps^{-1}$) | ($ps^{-1}$) |
| 1 | 0.1 | 1.8 |
| 2 | 1.0 | 2.5 |
| 3 | 5.0 | 0.6 |

We compared relaxation processes of the kinetic energy fluctuation metrics $\Omega_{FN}(0)/\ \Omega_{FN}(t)$ after a 300K→ 250K temperature jump with the Brendesen temperature coupling method[115] using three different coupling constants: 0.1ps, 1.0ps and 5.0ps. As shown in Figure 3.16, the relaxation time is very sensitive to the coupling constant used. In general, based on the $D_{KE}$ values given in Table 3.6, 1.0 ps seems to be a reasonable length of time for the Trp-cage system to relax to the new temperature as monitored by thermodynamic equal partition data.

### 2) Exchange ratio criterion

As mentioned previously, if the exchange is attempted more frequently, the replicas will be more mobile as they are more likely to change temperature. That implies that the fewer steps between exchanges, the larger the overall exchange ratio will be expected.

In Figure 3.17, a plot of the normalized overall exchange ratio as a function of equilibration time is shown for Trp-cage REM simulations. Totally, 5 sets of REM simulations with different equilibration time have been carried out for 10,000 exchanges each. Each simulation started with the same set of temperatures, which were designed to give exchange acceptance ratio of $r_{target}$. The normalized exchange ratio drops quickly with time. After 0.4ps or so, it reaches a plateau; this indicates that a length of 0.4 ps seems necessary and sufficient for the equilibration of Trp-cage system between every exchange trials.

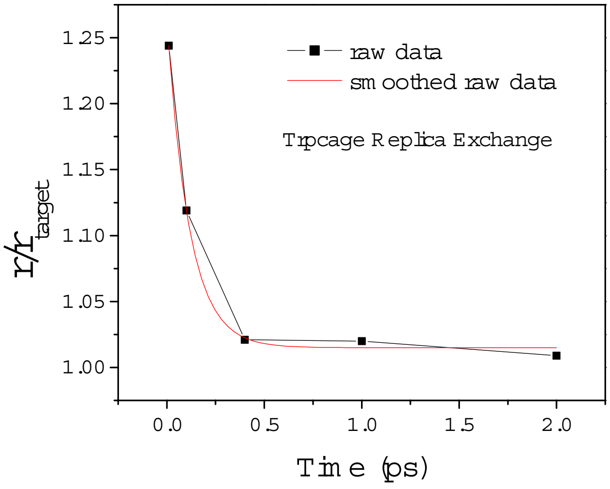


Figure 3.17 Plots of the normalized overall acceptance ratio as a function of equilibration time during 10,000 REM exchanges for Trp-cage simulations. Data points were fit to an exponential decay.

In conclusion, the proper choice of how often exchange should be attempted is largely determined by the specific details of the simulated system, such as the size and even energy surface features etc. There really is no easy solution to this question. However, as illustrated in the previous analysis of the Trp-cage system, our proposed force metric and exchange ratio criteria can be helpful in evaluating the equilibration time between each exchange trial.

### 3.6.3 How to determine the convergence of the REM simulations

REM has been demonstrated to be an efficient method for improving sampling[58,118,119], which means that REM results usually converge much faster than standard MD. However, the convergence of REM still can’t be assumed. Moreover, different thermodynamic quantities of one particular molecular system may have significantly different time scale of convergence rate[116]. Therefore, even though the accurate evaluation of convergence rate is of great importance in both MD and REM simulations, a thorough and practical measure of convergence is not easily available.

**1) Population distribution**

One of the most straightforward means of evaluating convergence is to monitor the time-dependent data (or step-dependent data in case of REM) of some structural or thermodynamic quantities. Those quantities such as RMSD, clusters identified and conformational sub-states hopping rate have been previously used for this purpose in a number of studies[15]. The convergence of simulations can then be determined by using pre-defined tolerance value or standard statistical test.

Since the earlier work of Flory[120], it has been appreciated that significant conformational change in peptides or proteins are tied to rotation of torsion angles. We have studied the backbone dihedral angle changes in $(ala)_3$ tri-peptide using 0.5 μs MD and 60ns REM simulations. The probability density map of backbone torsion angles phi and psi from normal MD simulation is shown in Figure 3.18, and it can been seen that β is the most dominant and contributes about 78% while α contributes about 22%. The density map obtained from 60ns REM simulation is very similar to that from MD simulation, with the same locations of two maximum and the same probability for each one. The basin-hopping rate of α→β for the normal MD is about 5.3 $\text{ns}^{-1}$ while the rate of β→α is about 17.9 $\text{ns}^{-1}$. The rates in REM simulations are about 8 times faster respectively. During the overall simulation, the total hopping events are more than 6000 times in both MD and REM simulations, which provides an enormous data set to study the convergence of sampling for this system.

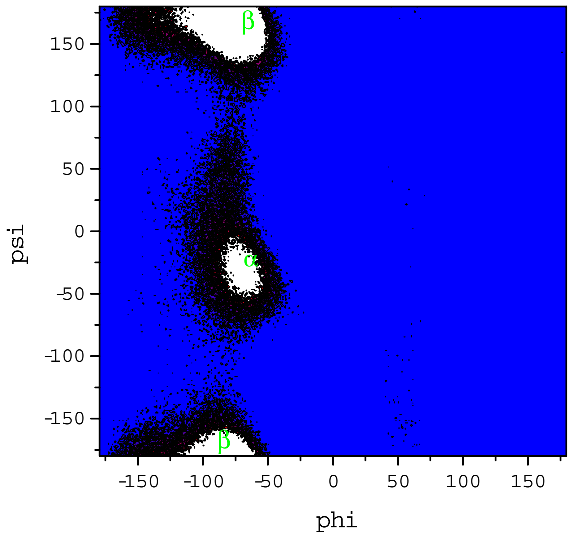


Figure 3.18 Backbone torsion angles phi and psi of $(\text{ala})_3$ shown as probability density maps for the normal MD simulation. White regions indicate potential energy wells and are heavily sampled.

Figure 3.19 shows the fraction of α configuration as a function of the time for the normal MD (left) and REM (right) simulations. It clearly shows that REM is much faster to reach the convergence in terms of dihedral angle sampling as compared to normal MD. However, it is worthy of noting, even in such a trivial system, the convergence of torsion angle sampling takes about 10 ns in REM; and much longer time will be needed for the normal MD since after 60 ns the number of α population is still fluctuating more than 2% around the final value.

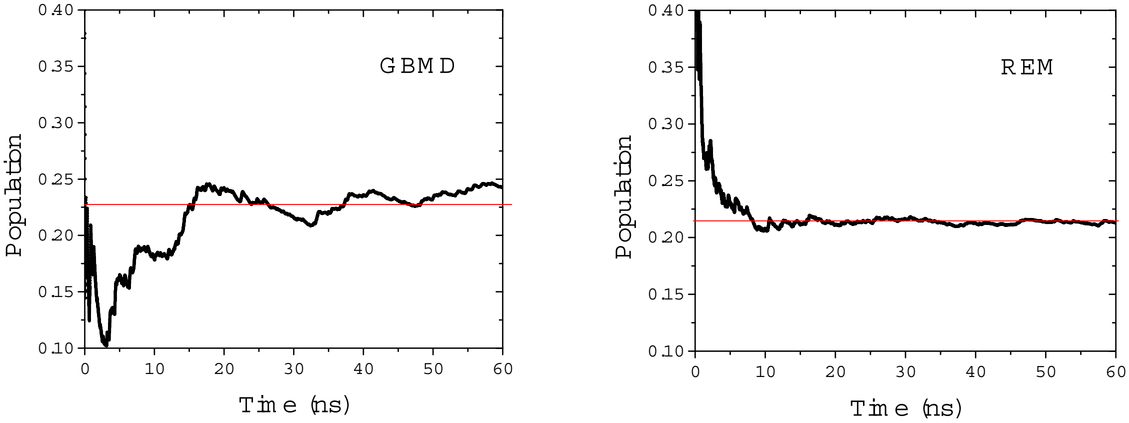


Figure 3.19 Fraction of $\alpha$ configuration shown as a function of simulation time in normal MD (Left) and REM (Right) simulations.

**2) Lowest frequency motion mode convergence**

Recently, quasi-harmonic analysis or principle component analysis (PCA) has been widely used to compute the modes or the collective variables responsible for most of the variation in the data[121-123]. The lowest frequency mode is of particular interest because it corresponds to the largest and slowest collective motion occurred in the simulation. Consequently, in order to have a thorough measure of the overall converging behavior in a system with multi-time scale motions, time series of the lowest frequency mode $r(x_1, x_2, ...., x_n, t)$ is one of the most illustrative quantities can be used for this purpose.

The QUASIH module in AMBER has been used to compute the frequencies and modes from MD trajectories as previously described[79]. The quasi-harmonic calculation is based on the assumption that energy surfaces are quadratic in the vicinity of energy minima. The lowest frequency mode $r(x_1, x_2, ...., x_{3N})$ is a vector with 3N components, indicating the contribution to the overall motion from the component motions along 3N Cartesian coordinates of the molecule.

In the case of a single long trajectory, a possible procedure would start with the full simulation data, and then divide it into sequential, non-overlapping blocks of the same size. If each block has the length of $\tau$ and total number of data points of $l$, then the total number of blocks will be $n = l/\tau$. We do quasi-harmonic analysis for each block of data, and then calculate the average mode as follows (Figure 20 top),

$$\vec{r}_{mean}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\vec{r}_i(\tau) \qquad (3.38)$$

The mean value of deviations from the average can then be determined by,

$$R_{lowest}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\| \vec{r}_i(\tau) - \vec{r}_{mean}(\tau) \| \qquad (3.39)$$

This process will be repeated with different length of $\tau$ ($\tau < 0.5l$). For any given value of convergence tolerance, the convergence time $\tau$ can be found very straightforwardly.

Another interesting and informative way of looking at the convergence is by comparing two independent trajectories. For instance, after the same period of time $t$, we can calculate lowest frequency mode using full simulation data sampled before $t$ for each trajectory, and then the deviation between two vectors can be defined as (Figure 20 bottom),

$$D_{lowest}(t) = \| \vec{r}_a(t) - \vec{r}_b(t) \| \qquad (3.40)$$

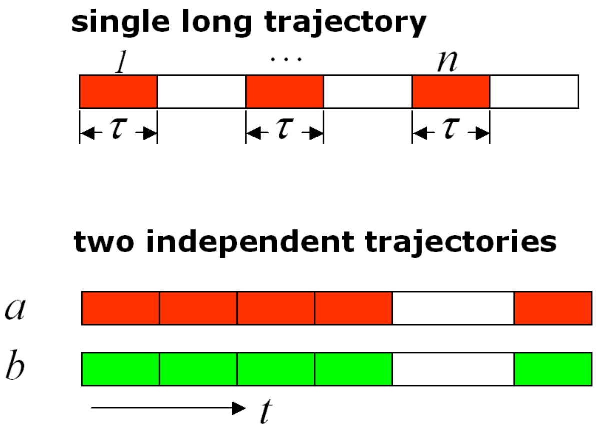


Figure 3.20 For the single trajectory, the mean deviation of the lowest frequency modes $R_{lowest}$ is calculated within different block length of $\tau$. And the use of multiple blocks improves the calculation accuracy. For two independent trajectories, the deviation of the lowest frequency modes between two trajectories $D_{lowest}$ is calculated with the accumulative time t.

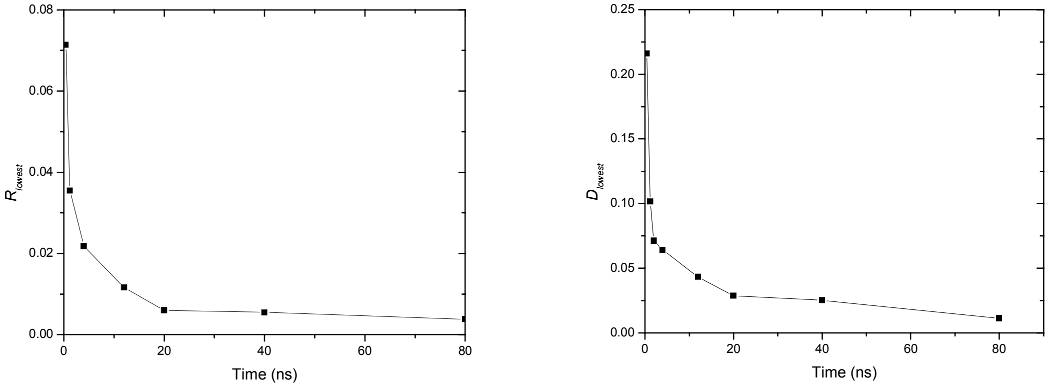


Figure 3.21 The convergence of the lowest frequency mode shown as a function of time. The left shows the data computed from a single 0.5 μs MD simulation, the right shows the data computed from two independent 80 ns MD simulations; all simulations were run on $(ala)_3$ test system.

According to the results shown in Figure 3.21, 20 ns or so is the time needed to reach convergence, which is consistent with our previous results obtained from backbone dihedral population analysis. The slowest motion in $(ala)_3$ dynamics is mainly the backbone torsion transition. It is not surprising, however, the results calculated from two independent 80ns MD trajectories show a little slower convergence rate as compared to single trajectory one. The variation between two methods indicates that the use of multiple simulations is probably better than a single very long trajectory to monitor the convergence of sampling.

# Chapter 4

# Insight to Structures and Dynamics of Damaged DNA from MD and REM Simulations

## 4.1 Introduction

Reactive oxygen species cause oxidative DNA damage that must be repaired in order for a cell to maintain genomic stability. One of the most abundant forms of DNA oxidative damage is 8-oxo-7, 8-dyhydroguanine (8oxoG, Figure 4.1)[66,124]. In *Escherichia coli*, repair of DNA containing 8oxoG can occur by three key enzymes, MutT, MutM (Fpg) and MutY, thus protecting cells from deleterious effects of guanine oxidation.

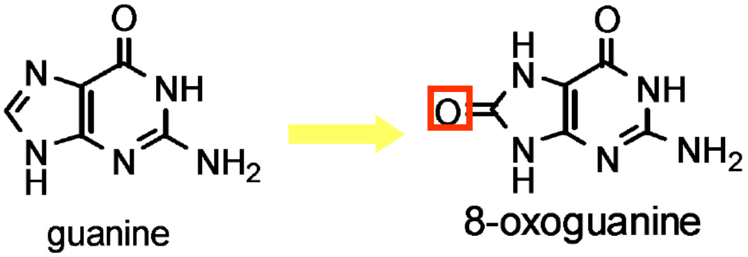


Figure 4.1 Structures of guanine and oxidized 8-oxoguanine (8oxoG).

The role of DNA repair enzyme MutY and Fpg in preventing mutagenesis by oxidized guanines in DNA is demonstrated in Figure 4.2. Oxidation of guanine in DNA generates 8oxoG opposite cytosine (8oxoG:C). Fpg recognizes the lesion and excises the 8oxoG base. If the 8oxoG strand is not corrected rather replicated at this point, the lesion preferentially mispairs with adenine to give an 8oxoG:A intermediate, which can be repaired by MutY as the second line of defense. Otherwise, replication of the adenine strand in the 8oxoG:A intermediate completes the mutational process, resulting in a permanent G:C to T:A transversion mutation.

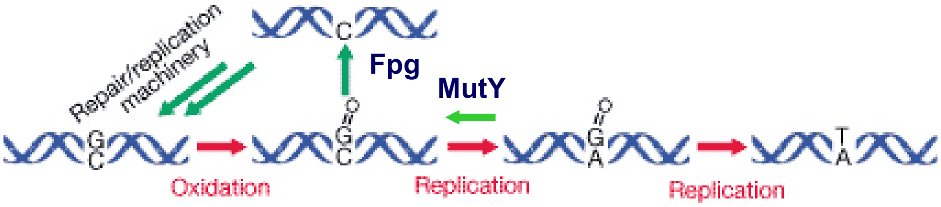


Figure 4.2 A schematic illustration of the mechanism of DNA repair enzyme MutY and Fpg in preventing mutagenesis by 8oxoG in DNA. (Adapted from reference [125])

Long before the enzyme/DNA complex structure was determined, the structures of DNA containing 8oxoG:A and 8oxoG:C pairs had been extensively studied by NMR and X-ray crystallography[66,126]. These studies have suggested that 8oxoG:C has the normal *anti:anti* form of base pair, and 8oxoG:A has the Hoogsteen *syn:anti* form of base pair. But as far as we know computational studies supporting these observations have never been reported.

About two years ago, the crystal structure of Fpg complexed with DNA containing 8oxoG:C damage (PDB code: *1K82*) was determined by treating the Schiff base intermediate with $NaBH_4$ to form a stable covalent complex between the first residue Proline and DNA[127]. Recently, the structure of 8oxoG:A DNA bound to MutY (PDB code: *1RRQ*) was also crystallized through the use of disulphide crosslinking[128]. Even though the crystal structures of both complexes reveal significant aspects of the damage recognition and repair mechanism, many crucial details of dynamic recognition remain unclear. For example, the 8oxoG nucleotide has an *anti* glycosidic bond conformation when bound to MutY and disengaged from base pairing, whereas its conformation is *syn* when paired with adenine and not bound to MutY[129,130].

And in the case of Fpg complex, using the model of 8oxoG in a complex with eukaryotic 8oxoG DNA glycosylase hOgg1[125], the deoxyribose linked to the oxidized base was overlapped upon the backbone ($P^0$-O-C5'-C4'-C3'-O-$P^{-1}$) of Pr site, it was found that both *anti* and *syn* conformations of the 8oxoG can be fit into the binding pocket. However, it is somewhat surprising that the size and the nature of this pocket are even better and require fewer rearrangements of the surrounding residue side chains for a *syn* conformation. So now the questions are what does the enzyme see in the 8oxoG DNA to know if it needs repair; if the free form of 8oxoG:A assumes a *syn* configuration, how is the *syn* → *anti* transition achieved upon MutY binding; and how is this glycosidic bond rotation related to the adenine partner's extrusion into the active site?

Furthermore, recent structural and biochemical studies have suggested that the specific recognition of the damaged 8oxoG base or 8oxoG:A mismatch arises from the distinct dynamic behavior of the system. Osman et. al.[131] studied the coupling between base opening and local bending and proposed that the changes in dynamics were primarily responsible for the lesion recognition; David and Helquist et. al.[132] applied restrained MD simulations to G:T mismatch and suggested the dynamic differences in helical parameters between normal pair and G:T mismatch was important to mismatch recognition; however, the studies conducted by Miller et. al[133] concluded that the primary effect for specific damage recognition is due to preferred bending direction.

These apparent controversies make it necessary to use more converged data in describing the structure and dynamics of the damaged DNA.

In the present work, we use MD simulations to study the structures and dynamics of the native and several damaged DNA 13-mers. Our simulation results confirmed the predominance of the normal *anti*:*anti* form of the 8oxoG:C base pair and the Hoogsteen *syn:anti* form of the 8oxoG:A pair[129,130]. We also complemented the multiple MD simulations with our modified REM approaches to the study of the equilibrium properties of structural quantities such as local bending and base pair opening in the DNA. We find that the *anti* → *syn* transition and base opening of 8oxoG:A mispair are highly coupled on the basis of free energy landscape. Similar to those have been proposed by Osman in thymine dimmers[131], the bending enhanced by damage base greatly lowers the barrier of base flipping, and thus having significant impacts on the enzyme recognition.

## 4.2 Methods

### 4.2.1 Force Field Parameters

**8oxoG parameters** Partial atomic charges for 8-oxo-7, 8-dihydroguanine in the 6,8-diketo form were adopted from Miller[133], where charges were evaluated in the RESP model of Bayly et al.[27] based on an electrostatic potential energy calculated with the 6-31G* basis set for consistency with other atomic charges in the AMBER force field. The resulting atomic charges are shown in Table 4.1a. Bond parameters were also taken from the same reference[133], where all of them use standard AMBER parameters except those for carbon and nitrogen on the five-member ring of 8oxoG. These parameters are approximated by the closest substitutes in the existing AMBER parameters. The correspondence is shown in Table 4.1b.

Table 4.1a Partial Charges for the 8oxoG Base

| ATOMS | CHARGE (e) |
|---|---|
| N1 | -0.4025 |
| H1 | 0.3266 |
| C2 | 0.7208 |
| N2 | -0.9625 |
| 1H2 | 0.4371 |
| 2H2 | 0.4371 |
| N3 | -0.6118 |
| C4 | 0.2108 |
| C5 | -0.0211 |
| C6 | 0.4299 |
| O6 | -0.5500 |
| N7 | -0.5129 |
| H7 | 0.4077 |
| C8 | 0.4468 |
| O8 | -0.5558 |
| N9 | 0.1110 |

Table 4.1b Existing AMBER force field parameters substituted for parameters needed to simulate DNA containing 8oxoG[133]

| Missing parameter | AMBER parameter |
|---|---|
| CK-O | C-O |
| CB-NA | CB-NB |
| CK-NA | CK-NB |
| CB-CB-NA | CB-C-NA |
| C-CB-NA | C-CB-NB |
| CB-NA-H | CC/CR/CW-NA-H |
| CB-NA-CK | CB-NB-CK |
| NA-CK-O | NA-C-O |
| N*-CK-O | N*-C-O |
| N*-CK-NA | N*-C-NA |
| CK-NA-H | CC/CR/CW-NA-H |
| X-CB-NA-X | X-CB-N*-X |
| X-CK-NA-X | X-CC-NA-X |
| X-X-CK-O | X-X-C-O |

**[4Fe-4S] cluster parameters** The catalytic domain of MutY/DNA complex contains the structural [4Fe-4S] cluster. Since this highly charged cluster is within 5 Å range of the damaged base, its strong influence on the structure of DNA is expected from their electrostatic interactions. Therefore, we include this motif in the MutY/DNA complex simulation. But during the simulation, the [4Fe-4S] cluster is kept fixed to its starting X-ray structure using the positional restraint with a force constant of 3 kcal·mol$^{-1}$·Å$^{-1}$, thus the bond and angle parameters for the cluster are not very important. Therefore, the bond and angle equilibrium values are average values from MutY crystal structure; the stretching and bending constants were estimated from the combination of experimental values and AMBER existing ones. All the dihedral parameters are set to 0. Lennard-Jones parameters for iron was taken from Karplus's parameter used in myoglobin[134] while for inorganic sulfur, cysteine sulfur value from AMBER force field was used.

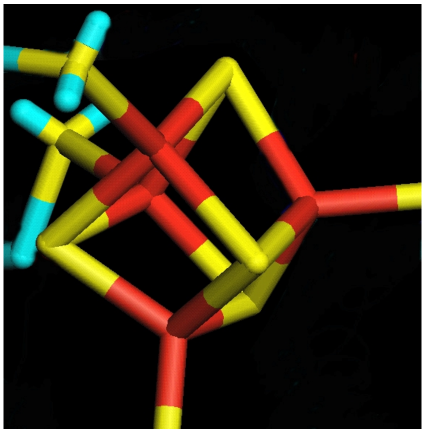


Figure 4.3 Structures of [4Fe-4S] cluster with sulfur in yellow and iron in red.

As far as charge parameters are concerned, computation of atomic charges can be accomplished using quantum mechanic calculation with RESP procedure. However, as a starting calculation, we adopted charges for the [4Fe-4S] cluster from Bertini[135] and modified accordingly the charges for cysteine. Since metal ion charge is usually spread out partially on ligand, two extra points have been added around sulfur atoms. These two points have zero mass and vdw parameters, only bearing -0.4972e charge for each point. The geometry of each point is as specified in Table 4.2. Overall, our parameters are working properly with AMBER ff94 force field, and can reproduce satisfactory geometry around the cluster center. All the parameters for [4Fe-4S] and cysteine in the simulation are shown in Table 4.2.

Table 4.2 Force field parameters used for [4Fe-4S] cluster

| **BOND** | $\mathbf{K_{bond}}$(kcal·mol$^{-1}$·Å$^{-1}$) | $\mathbf{R_{bond}}$ (Å) |
|---|---|---|
| EP-S | 600.0 | 0.700 |
| FE-S | 150.0 | 2.420 |
| FE-SE | 150.0 | 2.210 |

| **ANGLE** | $\mathbf{K_\theta}$(kcal·mol$^{-1}$·degree$^{-1}$) | $\mathbf{R_\theta}$ (degree) |
|---|---|---|
| CT-S -EP | 150.0 | 109.60 |
| FE-S -EP | 150.0 | 109.60 |
| EP-S -EP | 150.0 | 120.00 |
| FE-S -CT | 150.0 | 91.65 |
| SE-FE- S | 55.0 | 114.61 |
| SE-FE-SE | 55.0 | 105.35 |
| FE-SE-FE | 55.0 | 74.88 |

| **DIHE** | **IDIVF** | **PK**(kcal·mol$^{-1}$·degree$^{-1}$) | **PHASE**(degree) | **PN** |
|---|---|---|---|---|
| X-SE-FE-X | 1 | 0.0 | 180.0 | 2. |
| X-S -FE-X | 1 | 0.0 | 180.0 | 2. |
| EP-S-CT-X | 1 | 0.0 | 180.0 | 2. |
| FE-S-CT-X | 1 | 0.0 | 180.0 | 2. |

| **NONBOND** | **R***(Å) | **ε**(kcal/mol) |
|---|---|---|
| EP | 0.00 | 0.000 |
| FE | 0.90 | 0.300 |
| SE | 2.00 | 0.250 |

| **CHARGE** | **e** |
|---|---|
| FE | 1.3400 |
| S | -1.0400 |
| N | -0.4157 |
| HN | 0.2719 |
| CA | -0.0351 |
| HA | 0.0508 |
| CB | -0.2413 |
| HB2 | 0.1122 |
| HB3 | 0.1122 |
| SG | 0.3100 |
| EP1 | -0.4972 |
| EP2 | -0.4972 |
| C | 0.5973 |
| O | -0.5679 |

### 4.2.2 Base Sequence and Starting Structures

The 13-mers CCAGGA(8oxoG)GAAGCC and GGCTTCCTCCTGG has the form shown by Grollman to be cleaved efficiently by Formamidopyrimidine DNA glycosylase

(Fpg)[127]. Three others used as controls for comparison. The first four sequences used for our MD simulation are as follows:

G:C (dCCAGGAGGAAGCC)·(dGGCTTCCTCCTGG)

G:A (dCCAGGAGGAAGCC)·(dGGCTTCATCCTGG)

8oxoG:C (dCCAGGA(8-oxoG)GAAGCC)·(dGGCTTCCTCCTGG)

8oxoG:A (dCCAGGA(8-oxoG)GAAGCC)·(dGGCTTCATCCTGG)

In order to study the flanking sequence impact on the DNA flexibility, we added three additional sequences as follows,

5'AT-AT3' (dCCAGGA(8-oxoG)AAAGCC)·(dGGCTTTATCCTGG)

5'GC-AT3' (dCCAGGG(8-oxoG)AAAGCC)·(dGGCTTTACCCTGG)

5'GC-GC3' (dCCAGGG(8-oxoG)GAAGCC)·(dGGCTTCACCCTGG)

Canonical B-DNA structures were constructed by AMBER NUCGEN module, and the starting structure for DNA containing 8oxoG was obtained from canonical B-DNA, by replacing the hydrogen at C8 position on G7 with an oxygen atom and adding hydrogen to N7 position. The system was relaxed by 2000 steps of energy minimization to remove bad contacts. And it was subsequently equilibrated to 320K over a period of 2.4 ns using cycling heating protocol. Note that with GB solvation, the slow heating protocol seems extremely important to obtain stable trajectory at the start of simulation.

### 4.2.3 MD Simulation Details

**MD simulation in continuum solvent model** All the simulations were carried out with AMBER8 package[79] using AMBER ff94 force field[23] except the parameters for 8oxoG that have been described above. The time step was 2 fs, and all bonds involving hydrogen were constrained with SHAKE algorithm with a tolerance of $10^{-4}$Å[86]. No non-bonded cutoff was used. The Born radii were adopted from Bondi with modification of hydrogen by Case etc[46], and an offset of 0.13Å was used as recommended for nucleic acid simulations. The scaling factors for Born radii were taken from the Tinker modeling package[87]. The data was collected after the equilibration of about 2ns for each sequence.

**MD simulation in explicit water model** For the explicit solvent simulation, the canonical DNA structure was immersed into TIP3P water truncated octahedron box with the closest box clearance of 10Å, resulting in a total of 6400 water molecules for the simulation. 24 counter ions $Na^+$ were added to neutralize the net charges on the 13-basepair DNA. The solvated system was then equilibrated in 3 following steps: (i) 5000 steps of MD simulations of water (DNA fixed), (ii) 5 cycles (with decreasing restraints on DNA) of 500-step minimizations of the system, and (iii) 4 cycles (holding DNA with decreasing restraints and the last cycle with no restraint) of 5000-step MD simulations of the system. The time step was 2 fs, and all bonds involving hydrogen was constrained by SHAKE with a tolerance of $10^{-4}$Å. Simulations were carried out in NPT ensemble at 320K. A short-range cutoff of 9Å was used for nonbonded interactions and long-range interactions were treated with particle-mesh Ewald PME method[68].

**MD simulation of MutY complex** The force field parameters used for [4Fe-4S] cluster and 8oxoG have been described above. The crystal structure[128] (2.22Å resolution, protein data bank code 1RRQ) of the MutY complex with DNA containing an 8oxoG:A mispair was used as the starting model except that Asn (in 1RRQ mutant) was modified to Asp (in native MutY) and two missing loops(Res229-234 and Res287-292) in the crystal structure were built in using homologous structure of 1RRS[128]. The structure was further modified to replace the *anti* DNA with our simulated *syn* DNA with adenine partner also buried inside the duplex. Hydrogen atoms were added with *tleap* program in AMBER. Then the complex was solvated using TIP3P water model in a truncated octahedral box with about 18,000 water molecules. 28 counter ions $Na^+$ were added to neutralize the system. After initial minimization, the system was equilibrated using the same protocol as described above. After equilibration, simulations started in NPT ensemble at 300K.

### 4.2.4 Umbrella Sampling

The method used to establish the free energy profiles for *anti* → *syn* transition and base flipping process is through umbrella sampling method[60]. In order to allow sufficient overlapping between different windows, a soft harmonic bias potential of the kind $V_{\phi} = K_{bias}(\phi - \phi_n)^2$ was introduced, where $K_{bias}$ and $\phi_n$ are the force constant and reference angle respectively. And finally the simulation results are analyzed using WHAM method[136]. The PMFs have been calculated along two coordinates; the glycosidic torsion angle χ and the base flipping angle.

***anti* → *syn* transition** Taking the dihedral angle χ (defined as O4'-C1'-N9-C4 in Figure 4.4) as the reaction coordinate, potential mean force calculations were performed to determine the relative free energy values of conformations at 72 increments between -180° and 180°. At each dihedral angle, a harmonic potential with a force constant of 0.2 kcal/mol degree$^{-2}$ was applied, and a simulation consisting of 200ps of equilibration and 300ps of production was used to sample the fluctuations of dihedral angles. To test the sensitivity to the quadratic restraints, three different constraints ($K_{bias}$ = 0.05, 0.1, and 0.2 kcal/mol degree$^{-2}$) have been used to obtain consistent results.

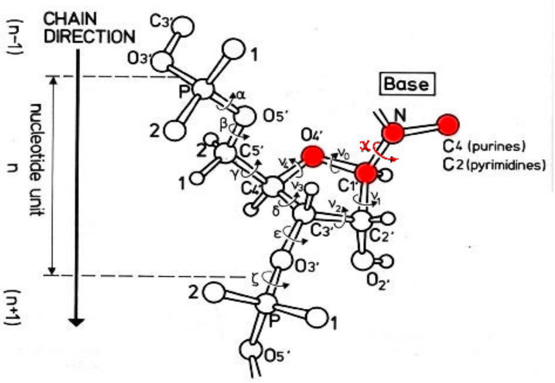


Figure 4.4 Four atoms forming 8oxoG glycosidic angle χ are shown as filled spheres, also shown are backbone torsion angles and sugar pucker angles.

**Base flipping process** The base-flipping angle used by MacKerell was taken as the driving coordinates for the base opening process[137]. The pseudo dihedral angle was defined by four mass centers of A, B C and D rings as shown in Figure 4.5. In a similar manner, a total of 72 windows were used to cover the entire range from -180° to 180° of base flipping angle. For each window, we run 200ps of equilibration and 300ps of sampling.

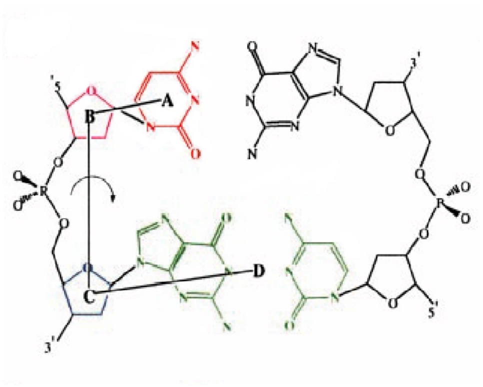


Figure 4.5 The base flipping angle shown on the right is defined by the center of masses of (*i*) the target C (red, A), (*ii*) its sugar moiety (purple, B), (*iii*) the adjacent 3' sugar moiety (blue, C), and (*iv*) the 3' GC base atoms (green, D)[137].

### 4.2.5 Modified REM Simulations

**PREM simulation** The mismatched base pair and its two flanking base pairs were defined as the focused region and the other part of the molecule as the bath region. The neighboring base pairs were included because the motion of damaged base pair was seen to be coupled with the breaking of neighboring base pairs in our previous multiple GB simulations. Eight replicas were used with target temperatures as 253K, 275K, 300K, 327K, 356K, 388K, 422K and 461K for each replica. The bath region temperature was maintained at 300K. The coupling constant of 0.5ps was applied to both bath and focused regions. The exchange was attempted every 1ps, and the anticipated overall acceptance ratio was 15%. We ran 12,000 exchanges in total for each of four sequences. The data was collected after every exchange for later analysis.

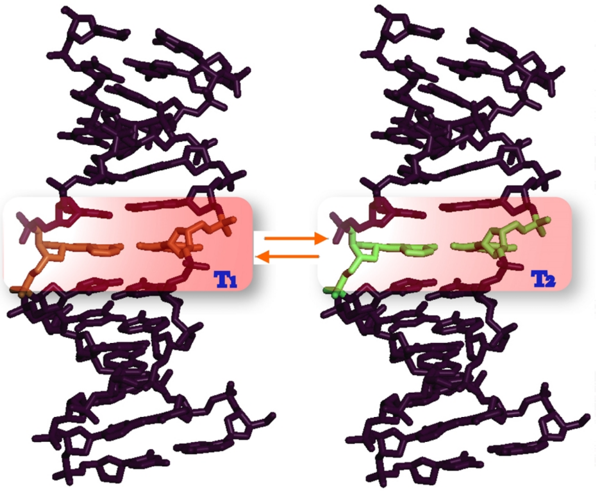


Figure 4.6 A schematic diagram shows the mismatched base pair along with its two neighboring base pairs are defined as the focused region in PREM simulation while the same region were replaced by five LES copies in LREM simulation.

**LREM simulation** Only two bases at the mismatched site were replaced by five LES copies using AMBER ADDLES module[79]. Accordingly, all the force field parameters of LES copies were scaled by a factor of five. Each LES copy was assigned one of the following target temperatures 80K, 93K, 108K, 125K, and 145K while the non-LES region was maintained at 100K. We ran 8,000 exchanges in total for each of four sequences.

### 4.2.6 Calculation of Bending Angle

DNA Bending was calculated with the program CURVES[138] as the angle between the local helical axis segments of A6/G8 (Figure 4.7). This angle was selected because it was anticipated to exhibit large changes for damaged DNAs from the native ones.

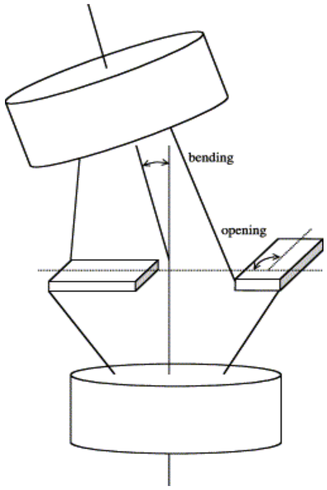


Figure 4.7 A schematic representation of the bending angle as defined in CURVES[138].

## 4.3 Results and Discussion

### 4.3.1 GB Simulations of Four DNA 13-mers

**G:C and 8oxoG:C in GB simulation** We first simulated four different DNA 13-mer sequences containing G:C, G:A, 8oxoG:C and 8oxoG:A base pairs. Canonical B-DNA was used as the starting structure for fully unrestrained MD simulations at 320K. Figure 4.8 shows the heavy-atom RMSD (compared to the representative structure from cluster analysis) as a function of time for the four simulations. As controls, both G:C and 8oxoG:C are stable, showing no major conformational changes in up to 40ns simulations. The exception is that after 40ns the terminal base pairs begin to fray a lot and cause dramatic conformation change for both molecules. However, no evidence for opening of the G:C or 8oxoG:C base pair is found. Neither do we find any significant change for G and 8oxoG glycosidic angles.

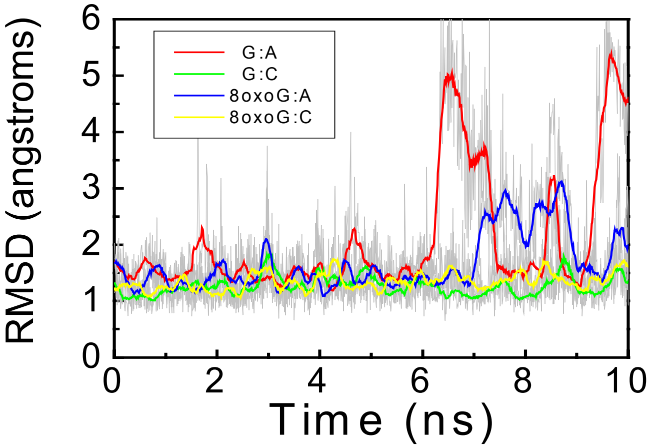


Figure 4.8 The heavy-atom RMSD (compared to representative structure) as a function of time for four DNA sequences. RMSD is calculated without including two end base pairs.

**G:A in GB simulation** G:A sequence shows greater variability than G:C and 8oxoG:C and has some periods of large change where the DNA gets very bent at the mismatch, and even shifts the register of base pairs at the mismatch to make alternate purine-pyrimidine pairs, then comes back. Replacement of cytosine by adenine base causes a remarkable destabilization of the duplex is in agreement with the proposed base-pairing scheme. G:A mismatch base pairs disrupt the normal Watson-Crick base pairs. The glycosidic angles of G:A mismatch were previously shown to adopt different conformations in different sequence context and solution conditions[126], such as G(*anti*):A(*anti*), G(*syn*):A(*anti*) and G(*anti*):A(*syn*). Particularly, the G(*syn*):A(*anti*) conformation dominates at low pH solution. Our simulation starts with G(*anti*):A(*anti*) and converts to G(*anti*):A(*syn*) in 10ns or so, then stays there for the rest of 40ns.

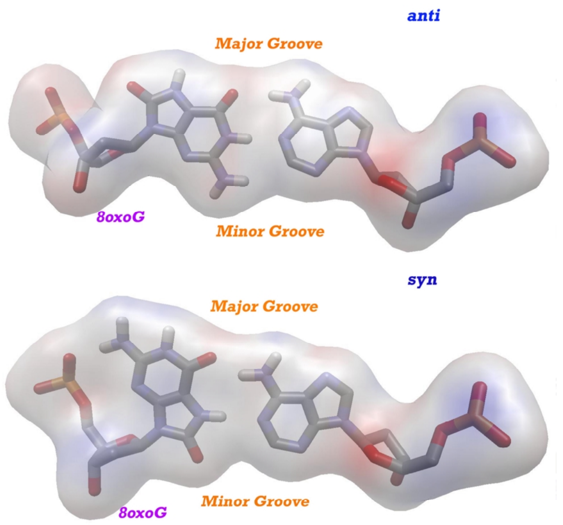


Figure 4.9 In 8oxoG:A simulation, a spontaneous *anti*→*syn* transition of 8oxoG glycosidic angle was observed, following the flipping of the base out of duplex. The above figure shows *anti*- and *syn*- configurations of 8-oxoG before and after the transition.

***anti* → *syn* transition of 8oxoG:A** In the case of 8oxoG:A sequence, the simulation starts with the 8oxoG(*anti*):A(*anti*) shown as *anti* in Figure 4.9. During the first several ns of simulation the 8oxoG seems to form a fine Watson-Crick base pair with the mismatch adenine, which is in agreement with the NMR experiment that Watson-Crick base pairs adopt *anti* glycosidic torsion angle. In about 4.2 ns, we observe a spontaneous opening of the 8oxoG:A base pair, flipping of the 8oxoG base, rotation around the glycosidic angle and subsequent re-formation of the base pair, resulting in an *anti* → *syn* transition for 8oxoG. The *syn* conformation is shown as *syn* in Figure 4.9. Initially, it is quite puzzling that G:A mismatch structure gets very distorted provided there isn't any interaction with the extra O in the 8oxoG. The simple reason is that G:A can not have a stable structure of either *anti* or *syn*. On the one hand, G:A has significant backbone strain penalty to stay as the *anti* conformation because G:A pair has much larger size than the normal G:C or A:T pairs. On the other hand, the formation of a *syn* structure at 8oxoG established a Hoogsteen base pair pattern for 8oxoG:A, which is however impossible for G:A mismatch since no hydrogen in N7 position is available for hydrogen bonding.

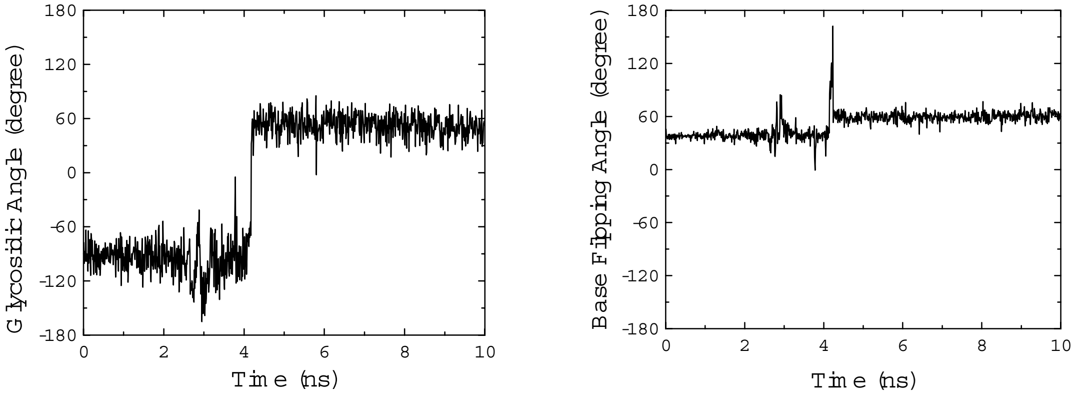


Figure 4.10 8oxoG glycosidic angle (on the left) and base flipping angle (on the right) as a function of time for 8oxoG:A GB simulation at 320K.

As shown in Figure 4.10, during the transition, the motions of 8oxoG glycosidic angle and base flipping angle are clearly coupled. Base opening occurs just before the glycosidic bond rotation. The importance of this coupling is that the rate of glycosidic angle rotation becomes of the same magnitude as that of base flipping, which is usually greatly increased in the mismatch DNA. Otherwise the barrier for this rotation would be much higher because inside the duplex leaves no room for the rotation to be accomplished. This manner of transition is similar in the essence to many cases, where local structural rearrangement requires partially unfolding of the protein and RNA.

On the left of Figure 4.11, we show three pairs of hydrogen bond distances as function of time during the transition. In 3ns, there is transient base pair breaking, but it fluctuates back immediately. In 4.2ns when the base 8oxoG flips out, the base pair is lost again. However, this time it comes back with a new pair of hydrogen bonds. During the transition, backbone torsion angles also participate in the transition, at least two backbone torsions $\varepsilon$ and $\xi$, which are shown on the right plot of Figure 4.11, are involved when going from *anti* to *syn*. Note that the backbone torsion motion is not just a consequence of base flipping; rather it may cause compression of the helix at the *anti* orientation so that facilitate the base flipping. When comparing damaged 8oxoG:A DNA and native DNA simulations, significant differences in the backbone torsions are found. The unusual conformation at the mismatch site suggests that the damaged DNA is more flexible and more easily disrupted, and therefore can have great impact on the enzyme recognition.

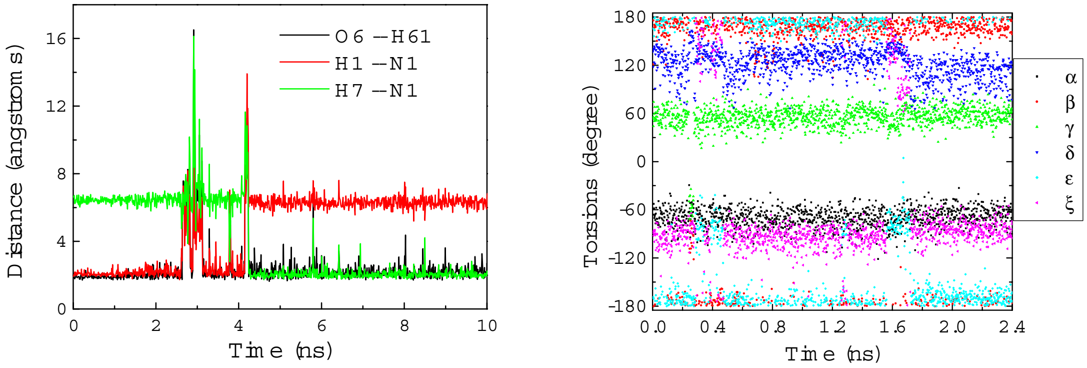


Figure 4.11 Three pairs of hydrogen bond distances (on the left) and backbone torsion angles (on the right) of 8oxoG as a function of time for 8oxoG:A GB simulation at 320K. It clearly shows *anti* → *syn* transition is coupled with hydrogen bond and torsion angle fluctuations.

To further understand why the *anti* → *syn* transition occurs, we did some energy component analysis on the system at the transition. Particularly, we looked at two energy components during the simulation. The definition of these two components is shown on the top of Figure 4.12. One is the base pair energy, which includes the non-bonded interactions between 8oxoG and its adenine partner. The other is called stacking energy, consisting of non-bonded interactions between 8oxoG and its two flanking bases, adenine and cytosine. Two energy components for each snapshot structure sampled at the transition are given on the left column of Figure 4.12. The big spike is the transition point where the 8oxoG is hanging out of the duplex. A detailed histogram analysis of these energy terms is shown on the right column of Figure 4.12.

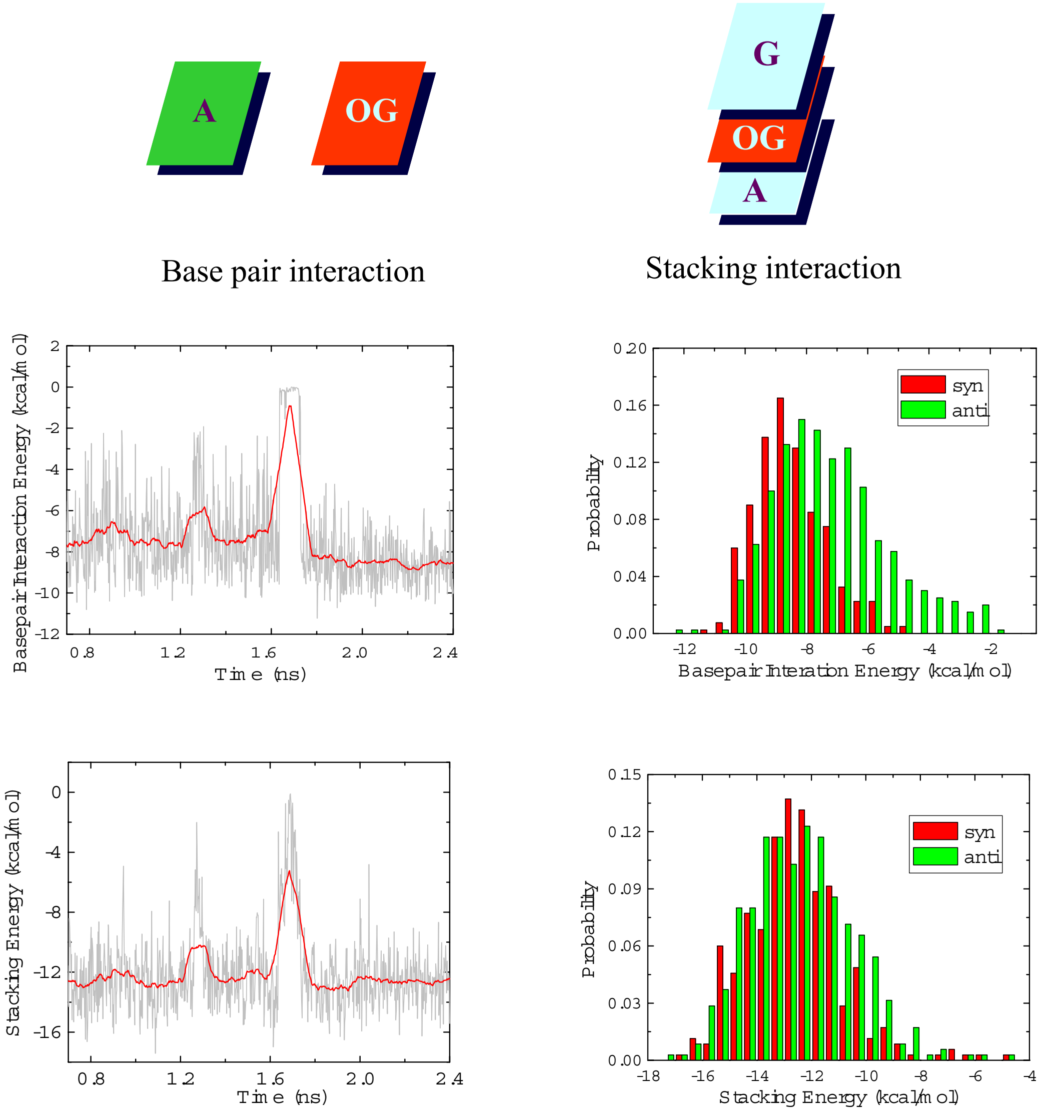


Figure 4.12 Energy decomposition analysis during 8oxoG *anti* → *syn* transition simulation. On the top shows the base pair interaction energy of 8oxoG with the opposite base A20; on the bottom shows the stacking energy of 8oxoG with the neighboring bases A6 and G8; The left panel shows energies as the function of time during the transition; the right panel shows the probability distribution of energy componets before and after the transition.

We summarize the energy component calculations in Table 4.3. The results reveal both *anti* and *syn* structures have similar stacking energies, the preference for the *syn* conformation is due to its more favorable base pair interactions. In structural terms, although both *anti* and *syn* structures can form two hydrogen bonds, the *anti* orientation of the purine-purine bases at mismatch site increases the cross-strand C1'-C1' distance from its value in purine-pyrimidine pairs. The perturbation resulting from the increased C1'-C1' separation seems to diminish the effective hydrogen bonding of the *anti* structure and partly lead to the *anti* → *syn* transition.

Table 4.3 Energy decomposition analysis during 8oxoG *anti* → *syn* transition simulation.

| | Stacking<br>(kcal/mol) | Base pair<br>(kcal/mol) | Total<br>(kcal/mol) |
|---|---|---|---|
| *anti* | -11.9±2.5 | -7.2±1.8 | -19.1 |
| *syn* | -12.4±1.7 | -8.5±1.2 | -20.9 |
| *diff* | -0.5 | -1.3 | -1.8 |

### 4.3.2 Multiple GB Simulations of 8oxoG:A

While the result of our MD simulation is consistent with available structure data for the sequence containing the 8oxoG:A pair, the transition process itself has never before been observed in atomic detail. We validated this observation by performing an additional 45 MD simulations starting from the *anti:anti* conformation. We observed multiple *anti* → *syn* transition events that tend to follow two pathways.

We plotted the distances of 8oxoG:A pair and its two neighboring base pairs as a function of time for two transition pathways in Figure 4.13. The left plot shows a base flipping-out pathway without breaking of the neighboring AT base pair while the right shows a pathway involving breaking of the neighboring AT base pair. However, in neither case, the 3' neighboring GC base pair breaks (green line in Figure 4.13). Moreover, as shown by blue lines in Figure 4.13, the major groove width decreases about 3Å after the transition. This suggests the resulting *syn* conformation stabilizes the structure quite a bit by releasing backbone stress due to two bulky purine bases 8oxoG and adenine in their unfavorable *anti:anti* orientation.

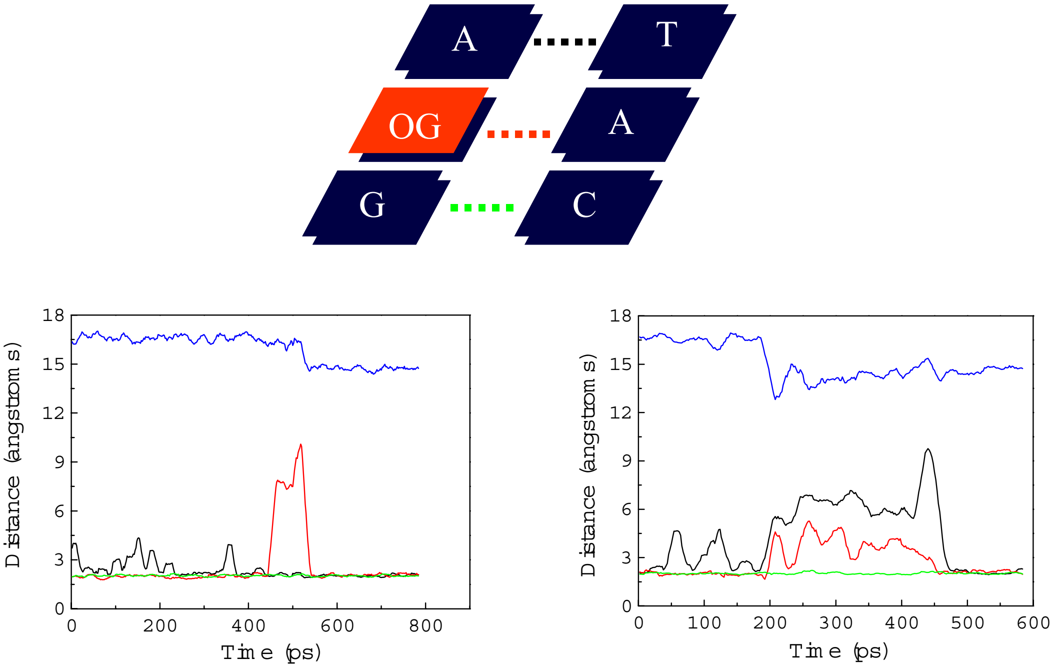


Figure 4.13 The distances of 8oxoG:A pair and two neighboring base pairs of 8oxoG:A as a function of time for two transition pathways. Major groove width in blue, A:T base pair distance in black, 8oxoG:A base pair distance in red, G:C base pair distance in green

In Table 4.4, we summarize *anti* → *syn* transition events occurred in 45 simulations. Overall, in about 20% of 45 trajectories, the *anti*→ *syn* transition is observed. Between two pathways, the transition involving the base flipping is more than two times preferred as compared to the transition by breaking neighboring AT pair. This suggests the latter pathway will have a higher kinetic barrier than the former one. The higher barrier of breaking AT pair pathway will further be confirmed in the next section by REM energy landscape data.

Table 4.4 Summary of *anti* → *syn* transition events occurred in 45 independent MD simulations for 8oxoG:A starting with *anti:anti* orientation.

| | Path | Events | Probability (%) |
|---|---|---|---|
| *anti* → *syn* | Flipping | 9 | 20 |
| | No flipping | 4 | 9 |
| Total *anti* → *syn* | | 13 | 29 |

### 4.3.3 Modified REM Simulations

In an attempt to further understand the coupling between the glycosidic angle and base flipping angle we applied our LREM method to the lesion site of the above four DNA sequences, and obtained equilibrium probability distributions for alternate base pair conformations for each sequence. In Figure 4.14 we show the free energy surface of two torsion angles for 8oxoG:A, 8oxoG:C and native G:C DNA. In general, we define the conformation as a *syn* if the glycosidic angle lies between 0° and 90° and as an *anti* if from –90° to -180°**.** The distributions clearly show that *syn* is dominant for 8oxoG:A(50°) in contrast to the dominant *anti* in G:C(-150°) and 8oxoG:C(-150°). Additionally, 8oxoG:C spreads out a little wider than the native G:C, but the difference is not significant.

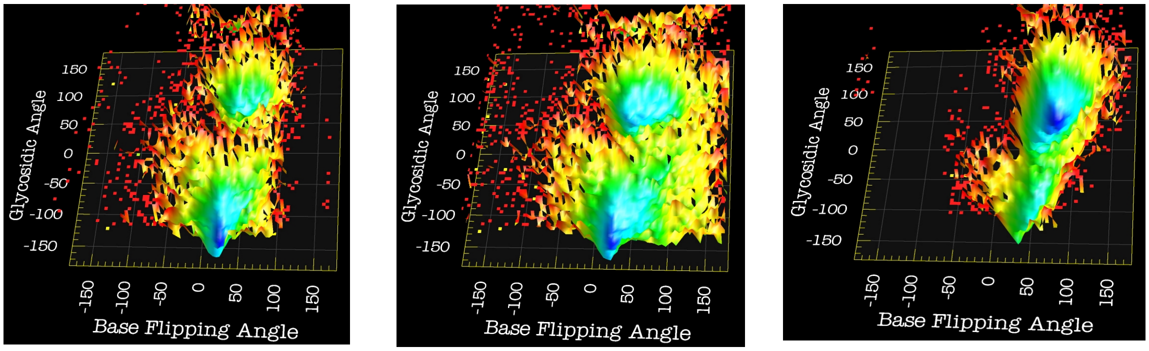


Figure 4.14 Free energy surface constructed from modified REM simulations using the base flipping and glycosidic angles as two reaction coordinates. (on the left): standard DNA; (in the middle): 8oxoG:C; (on the right): 8oxoG:A. The preferred conformation of the 8oxoG glycosidic torsion varies in the three sequences.

We further demonstrate that the resulting free energy landscape of 8oxoG:A is consistent with our multiple transition path simulations described above. In Figure 4.15 we show the free energy surface calculated from the LREM data, with white spheres showing the sampling of this landscape during simulations representing the two transition pathways. We estimate the free energy barrier for the path involving base flipping to be about 9 kcal/mol, while that for a direct *anti*→*syn* transition without flipping to be about 12 kcal/mol. The estimation is made based on the fact that the energetic barrier in LREM surface is approximately reduced by a factor of 1/N (N is the number of LREM replicas). These results are in good agreement with the probability of two transition pathways in the previous 45 MD simulations. The lower probability path has a barrier of about 3 kcal/mol higher than that of the dominant path. The consistence achieved among various simulations provides us with confidence that our data is well converged and our model is quantitatively accurate enough in describing the structure and dynamics of different DNA sequences.

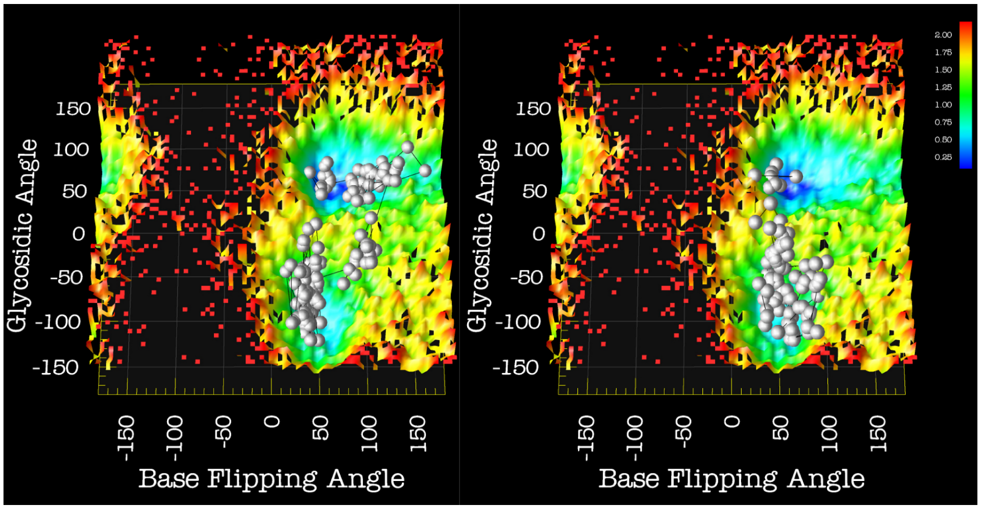


Figure 4.15 Projections of two *anti* → *syn* transition pathways onto the free energy surface constructed from variant REM simulations using the base flipping and glycosidic angles as two reaction coordinates. (left): base flipping-out pathway; (right): pathway sampled without flipping of the 8oxoG.

**DNA bending** It has been suggested that different bending behavior resulting from DNA damage might contribute to specific damage recognition. Thus, local DNA bending angles were calculated for 3 DNA sequences using the REM data. G:A is excluded from calculation because of its too large distortion in simulation. As shown in the Figure 4.16, oxidative damaged DNA, especially 8oxoG:A, adopt significant bent structures than the native DNA. The most probable bending angle of 12.5° for native DNA is well compared to previous reported values. In fact, the pattern of bending is dramatically dependent on the sequence context. A gentle degree of helix bending 7° for a DNA dodecamer, and more substantial bending for a longer sequence (14° per dedecamer) have been observed[139].

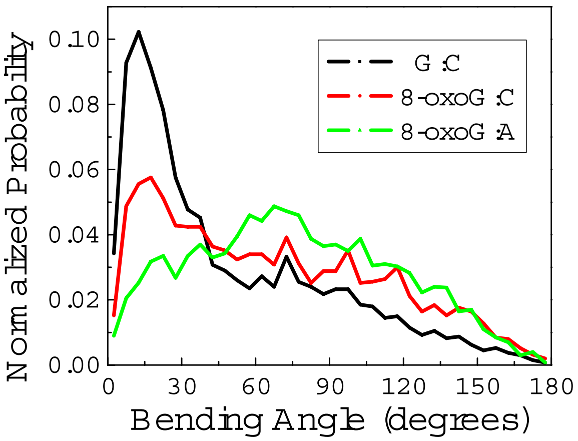


Figure 4.16 The probability distribution of bending angle during G:C, 8oxoG:C and 8oxoG:A simulations.

The most probable and average bending angle values sampled from simulations for G:C, 8oxoG:C and 8oxoG:A are summarized in Table 4.5. In crystal structures[128], MutY bends the DNA substrate 55°, which is roughly the same as the bend induced by EndoIII[140], but less than that induced by AlkA (66°)[141] or hOGG1 (70°)[125]. Our simulated most probable value of 67.8° is well compared to those numbers. MutY/DNA complex structure also reveals that the bend is localized to the lesion, with normal unbent B-form DNA helices projecting from either side. Comparison of the MutY-bound (from crystal structure 1RRQ) and unbound (from our simulation) 8oxoG:A DNA structures indicates that although some local structural variations exist in the lesion site, there is no significant global conformational changes after DNA-MutY binding. This observation strongly indicates that 8oxoG:A DNA bending is not induced by enzyme after binding, but it somehow bends first, thus being pre-organized for the enzyme recognition.

Table 4.5 The most probable and average bending angle values sampled from MD simulations for G:C, 8oxoG:C and 8oxoG:A

| | Probable | Mean |
|---|---|---|
| **G:C** | 12.5° | 56.3° |
| **OG:C** | 17.4° | 70.1° |
| **OG:A** | 67.8° | 78.7° |

The correlation between the bending angle and the rate of base opening is shown in Figure 4.17 for 3 sequences from PREM simulations, which suggests that it becomes easier to flip the base out in a more bent environment. Therefore, the presence of 8oxoG seems to lower the barrier of base opening by increasing the magnitude of DNA bending. This observation is however in contrast to Miller's results[133]. They concluded based on their simulations that 8oxoG had very little effect on the magnitude of bending but rather the bending direction changed significantly. On the basis of our well-converged REM data, we would suggest that Miller's 2 ns simulation in explicit solvent probably is still too short to reach the equilibriums. But we can't exclude the possibility of insufficient treatment of solvation with GB model in our simulations.

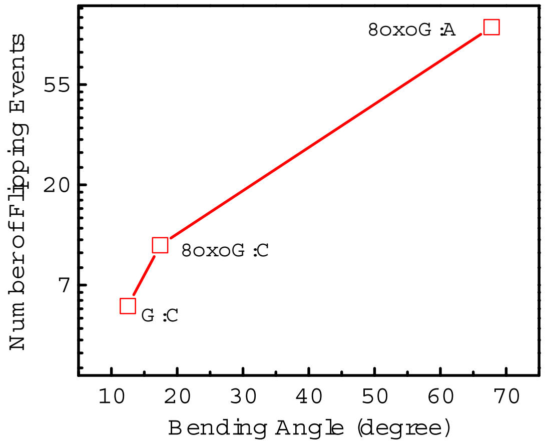


Figure 4.17 A plot shows the relationship between DNA bending and nucleotide flipping events observed in PREM simulations.

### 4.3.4 Umbrella Sampling

Although the time scale is usually limited in MD simulations, it has recently become feasible to carry out PMF calculation using umbrella sampling to clarify the detail of long time-scale events. We established the free energy profile for the rotation of 8oxoG glycosidic bond by this method. In Figure 4.18, we show the PMF as the function of glycosidic angle for 8oxoG:A (left column) and G:A (right column). It is evident that 8oxoG:A and G:A have their global minima at different locations. For 8oxoG:A, *syn* structure with glycosidic angle of 50° is almost 3 kcal/mol more stable than *anti* while this trend is switched in G:A where *syn* is 1.5 kcal/mol less stable than *anti*. These numbers match the REM results pretty well. For both sequences, there are considerable differences on the direction of rotation. In 8oxoG:A, towards the decrease of torsion, the rotation of glycosidic bond breaks AT pair and leads to a cost of about 8 kcal/mol for the *anti* → *syn* transition. This is about 3 kcal/mol less than the rotation in the other direction, which involves breaking of the GC pair instead. The same trend is observed in the G:A. But for either direction, the barrier is almost 3 kcal/mol lower than that of 8oxoG:A. This is believed to be related to increased flexibility of the G:A sequense.

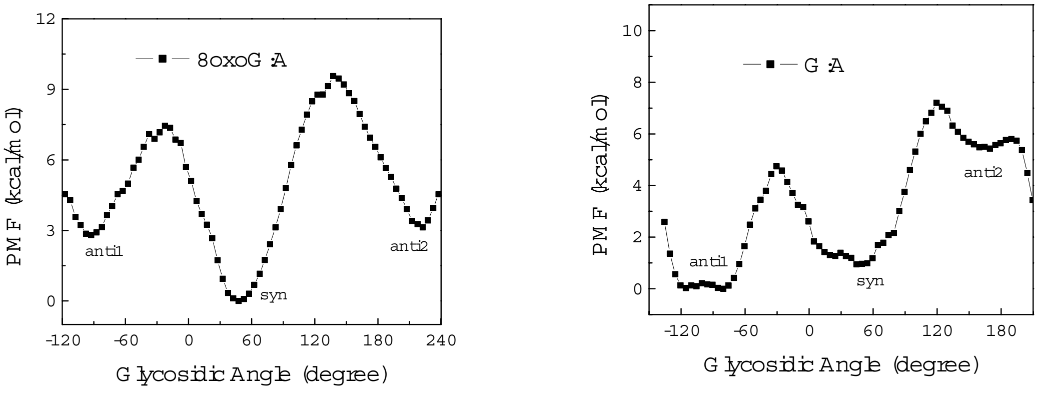


Figure 4.18 The potential of mean force as the function of glycosidic angle for 8oxoG:A and G:A sequences.

Base flipping has been established as a ubiquitous mechanistic feature of DNA glycosylase as the enzyme must gain access to the damaged base that is normally buried in the duplex structure of DNA[66]. Although the crystallography has provided a solid description of the initial and final states in the base flipping process, the reaction pathway and energetic are still unknown. Moreover, since the base opening has been shown to be coupled with glycosidic rotation, we would expect comparable energy barrier for the base flipping as that of glycosidic bond rotation.

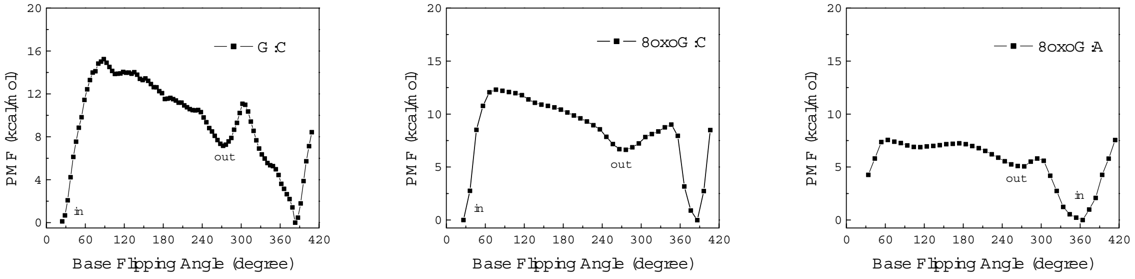


Figure 4.19 The potential of mean force as the function of base flipping angle for G:C, 8oxoG:C and 8oxoG:A three sequences.

In the literature, two types of reaction coordinates have been used to characterize the base flipping process. One is the pseudodihedral used by Mackerell in the study of base flipping in the DNA/cytosine-5-methyltransferase complex[137]. As shown in Figure 4.5, the pseudodihedral is defined by the mass centers of the cytosine ring, the sugar moiety, the adjacent 3' sugar moiety, and the 3' GC base atoms. The other type of base flipping angle has been described by using a base pair opening angle as define in CURVES[138]. Nevertheless, our PMF calculations using these two reaction coordinates turned out to be rather noisy and trajectories initiated from the PMF barrier failed to convert to either “in” or “out” structure. The above study calls attention to the possible danger of adopting a satisfactory order parameter as a reaction coordinate in the case of conformational changes involving many degrees of freedom. In fact, more and more experimental and theoretical studies on the 8oxoG DNA have revealed that the base pair flipping and the local structure fluctuations (such as local bending and backbone compression) are highly coupled[131,133]. Therefore we tried a new reaction coordinate to calculate the PMF, which is defined as a linear combination of the base flipping and the backbone torsion angle.

In Figure 4.19, we show the PMF as a function of base flipping angle for G:C, 8oxoG:C and 8oxoG:A. Similar to the results reported previously in the literature, in all three sequences, base opening toward minor groove is found to be more difficult than major groove opening, possibly due to steric clashes of the exocylic groups and the proximity of the sugar. 8oxoG:A has the lowest base opening barrier among the three sequence, with the barrier of 7.5 kcal/mol. This value agrees well with our results from both REM simulation and PMF of glycosidic bond rotation. A summary of all calculated PMF barriers is given in Table 4.6.

Table 4.6 Summary of PMF results for the glycosidic bond rotation and the base opening in four DNA 13-mers

| Sequence | Barrier Heights (kcal/mol) | | | |
|---|---|---|---|---|
| | Glycosidic rotation | | Base opening | |
| | Break 3’ GC | Break 5’ AT | Minor groove | Major groove |
| G:C | - | - | 15.8 | 11.7 |
| G:A | 7.2 | 4.9 | - | - |
| 8oxoG:C | - | - | 12.4 | 9.7 |
| 8oxoG:A | 10.7 | 7.5 | 7.5 | 5.9 |

### 4.3.5 Explicit Solvent Simulations of 8oxoG:A

Even though the GB model provides reasonably good description of solvent effect at probing the structure and dynamics of DNA, there are concerns about the hydrophobicity, mobile ions etc. For example, it has been shown that attraction of the counter ions to the damaged site didn’t preferentially neutralize atoms on the major groove, thus having substantial effects on the bending direction[133]. Moreover, the time scale of motion in GB simulation is lost due to its lack of friction. To ensure what we observed in GB simulations is not an artifact of the simplified solvent model, we carried out two additional simulations of 8oxoG:A in explicit water. One of the simulations starts from the same canonical B-DNA as that in the GB simulation while the other one starts from a *syn* conformation, which is a snapshot taken from the previous GB simulation after the *anti* → *syn* transition.

Figure 4.20 shows RMSD from the starting structure for *anti* and *syn* simulations. Not surprisingly, the *anti* simulation shows greater variability than the *syn* simulation. However, as shown in Figure 4.21, no evidence for the glycosidic bond rotation or base opening is seen in either of the simulations. Rather, the kink seen in the plot of *anti* simulation is caused by the change of sugar puckering of 8oxoG nucleotide instead of the rotation of the glycosidic bond. This is consistent with our previous observations that GB accelerates the sampling due to its lack of friction. Therefore, in order to observe the *anti* → *syn* transition in explicit water, much longer simulation or REM will be required. (This part of work is still in progress.)

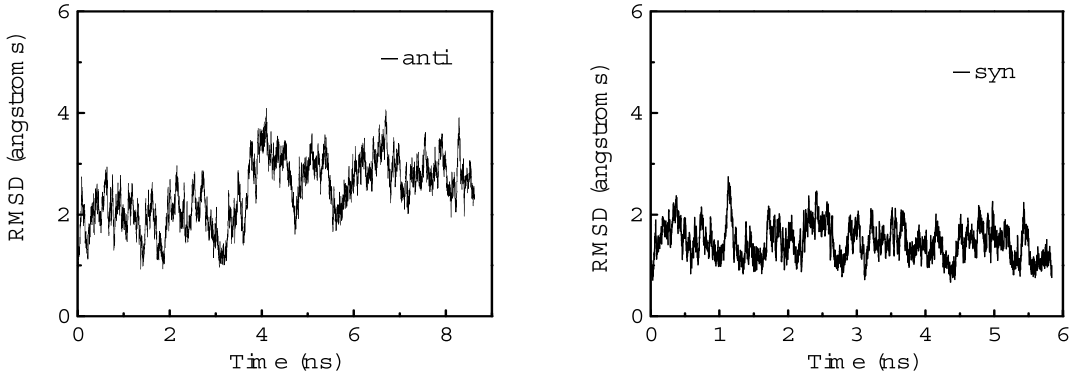


Figure 4.20 The heavy-atom RMSD (compared to the first structure) as a function of time for explicit solvent simulation 8oxoG:A starting from *anti* (on the left) and *syn* (on the right) configuration respectively.

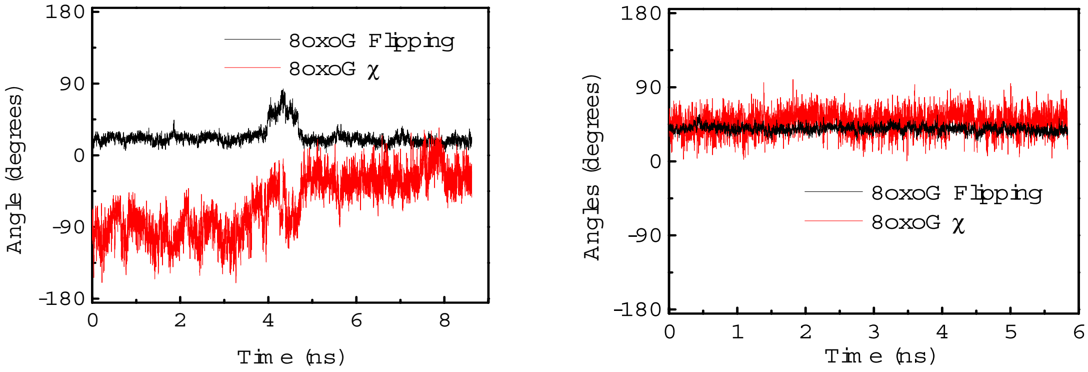


Figure 4.21 The glycosidic angle and base flipping angle of 8oxoG as a function of time for 8oxoG:A explicit solvent simulations at 320K. The left panel for *anti* 8oxoG:A and the right panel for *syn* 8oxoG:A.

### 4.3.6 Different Sequences Effects

8oxoG:A mismatches in different sequence contexts have distinct thermodynamic and structural outcomes[66]. We have observed several times of breaking of the flanking AT base pair in our 45 simulations of original 5’AT-GC3’ sequence. However, the other flanking base pair GC was hardly perturbed structurally and dynamically as compared to AT pair. The above observation leads us to hypothesize the greater stability of GC pair over AT pair might be the reason. Therefore, additional simulations for three alternate sequences all with a 8oxoG:A base pair were initiated, such as two neighboring GC pair (5’G:C-8oxoG:A-G:C3’), two neighboring AT pairs (5’A:T-8oxoG:A-A:T3’), and the switched AT-GC pair (5’G:C-8oxoG:A-A:T3’). In the following, we present comparisons of the data from different sequences that will enhance our understanding of the complex sequence effects on the glycosidic bond rotation.

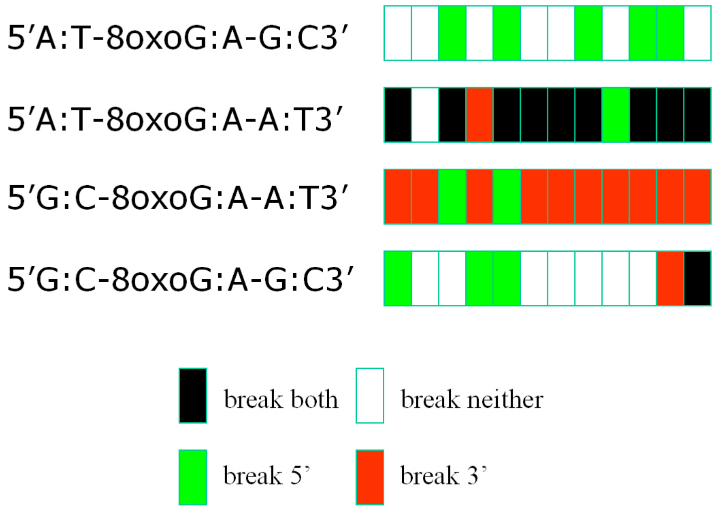


Figure 4.22 The summary of 5’and 3’ base pair breaking events occurred in three alternate sequences as compared to 8oxoG:A parent sequence, for each 12 independent MD simulations have been carried out.

As shown in Figure 4.22, 5’AT-AT3’ sequence is more structurally perturbed and dynamic compared to its parent 5’AT-GC3’ sequence. And the switched 5’GC-AT3’ sequence also shows increased flexibility. However, the increase in the breaking of flanking base pair doesn't result in an increase in the *anti* to *syn* transition as compared to their parent sequence. In contrast, the 5'GC-GC3' shows comparable *anti* → *syn* transition rate of 5'AT-GC3' whereas the breaking of flanking base pair is less frequent. This result is not totally surprising since it is clear from the free energy landscape (Figure 4.15) that the breaking of flanking base pair is only a minor pathway, thus to which the overall *anti* → *syn* transition rate is not directly correlated.

Table 4.7 Summary of *anti* → *syn* transition events occurred in 12 independent MD simulations for three alternate sequences as compared to 8oxoG:A

| | Events | Probability (%) |
|---|---|---|
| 5'A:T-8oxoG:A-A:T3' | 1 | 8 |
| 5'G:C-8oxoG:A-A:T3' | 3 | 25 |
| 5'G:C-8oxoG:A-G:C3' | 5 | 42 |
| 5'A:T-8oxoG:A-G:C3' | - | 29 |

### 4.3.7 Implication for Enzyme Recognition

In *Escherichia coli,* DNA repair enzyme MutY initiates repair of 8oxoG:A by removing the mispaired adenine from the DNA backbone. One of the remarkable differences between MutY and other glycosylases is the recognition of a normal base mispaired with 8oxoG. However, the interesting question is how this enzyme recognizes adenine only in the context of A/8oxoG or A/G mismatch (about $10^6$ fold greater than AT pair[66]). The puzzle was partially resolved when Noll and Clarke found the C-terminal domain of MutY shows significant homology to the MutT enzyme to which 8oxoG is a major substrate[142]. Recently, when the crystal structure of MutY/DNA complex was determined[128], it is clear the enzyme has two separate domains tethered by a polypeptide linker as shown in Figure 4.23.

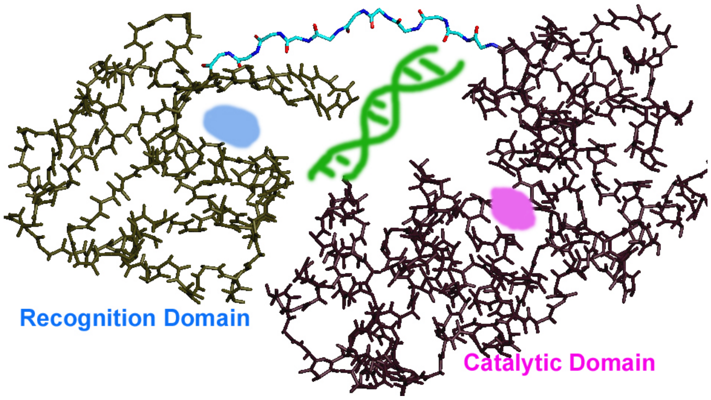


Figure 4.23 A schematic illustration of MutY/DNA complex. The recognition domain is shown in blue and the catalytic domain in pink.

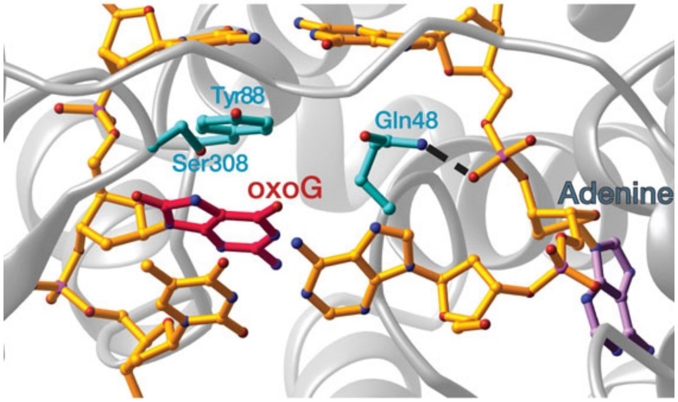


Figure 4.24 The binding site of MutY/DNA complex. 8oxoG is shown in red, the substrate adenine in purple, and the remainder of the DNA in gold. The protein backbone is presented as a grey ribbon trace, with side chains shown explicitly in cyan (adapted from reference [128]).

As anticipated, the crystal structure provides important insight to the enzyme activity and specificity. The interaction model of MutY catalytic domain and adenine is similar to that of AlkA, hOGG1 and EndoIII. The C-terminal domain shows an exquisite chemical complementary to the surface of 8oxoG. The binding mode is shown in Figure 4.24 with residues Gln48, Tyr88 and Ser308 indicated, among which the highly conserved Ser308 is proposed to be responsible for the 8oxoG and G discrimination[128].

Although the crystal structure of MutY/DNA provides critical information concerning MutY function, many dynamic details of the initial enzyme/DNA interactions are still missing. As we know, upon binding, a series of conformational changes occurred to both DNA and MutY, such as 8oxoG flipping out of duplex and subsequently coming back with an *anti* orientation, adenine partner also flipping out and inserting to the active site. Therefore, it is useful to clarify how these events are correlated; and particularly, how the swivel of 8oxoG from *syn* to *anti* facilitates the adenine flipping? In an attempt to answer these questions, we carried out one MD simulation of the MutY/DNA complex in water. The starting structure was obtained by replacing the crystal DNA structure with our simulated structure with a *syn* 8oxoG:A pair. The *syn* structure was then docked into the enzyme binding cleft by overlapping to the backbone heavy atoms of the DNA structure in MutY/DNA complex. After 0.4 ns of simulation, the development of RMSD and energy reach plateau (Figure 4.25). A snapshot during the simulation is shown in Figure 4.26. Compared to final binding complex as shown in Figure 4.23, different array of residues have been identified around the 8oxoG:A surface, such as direct hydrogen bonds between 8oxoG and Gln40, Arg83 instead of Gln48, Tyr88 and Ser308. It is evident that, after binding, significant local rearrangements have occurred to the MutY active site.

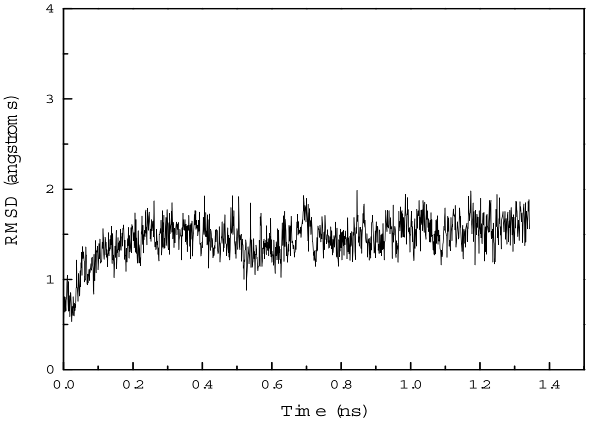


Figure 4.25 The heavy atom RMSD as a function of time for MutY/8oxoG:A complex explicit solvent simulation.

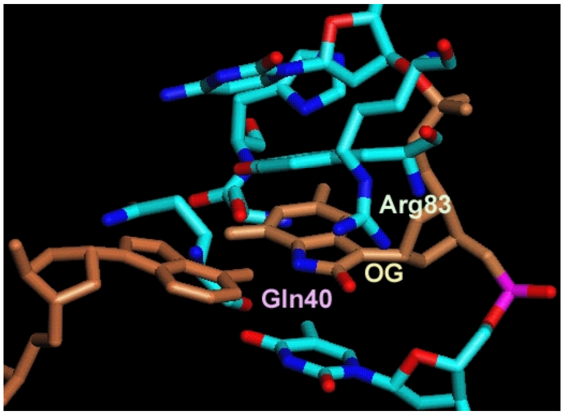


Figure 4.26 A snapshot sampled from MD simulation, showing the interaction surface of 8oxoG:A and several key residues of MutY.

Based on our simulations, a preliminary binding model was established to illustrate how the enzyme and DNA interact with each other before reaching their stable binding mode.  Our model (Figure 4.27) shows that 8oxoG:A specificity comes primarily from the shape complementary, especially the stereochemistry of the 8oxoG. The 8oxo group of 8-oxoG reaches much deeper in the minor groove compared to the standard GC pair. Specifically, interactions of 8oxoG with Gln40 and Arg83 in the minor groove, His301 and Ser300 in the major groove are the main factors of determining the initial orientation of DNA upon MutY surface. Interaction of Arg83 with 8oxoG will possibly facilitate the insertion of Tyr80 into the duplex and further trigger the eversion of 8oxoG base.

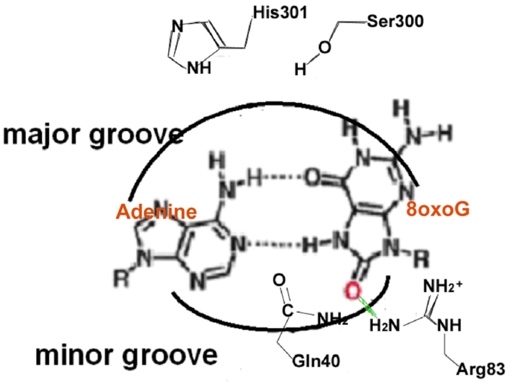


Figure 4.27 Our proposed binding model for the initial recognition of 8oxoG:A DNA by MutY based on the preliminary MD simulation.

It is intriguing that GA mispair is also a substrate of MutY, but the enzyme shows a less than six fold kinetic preference for G:A than 8oxoG:A despite the fact that helix disruption with 8oxoG:A is disfavored G:A by 3.4 kcal/mol[143]. As revealed by NMR experiments[126], GA mispair has a dominant G(*syn*):A(*anti*) alignment at low pH. Therefore, we hypothesize that under normal pH condition, equilibrium between G(*syn*):A(*anti*) and G(*anti*):A(*syn*) still exists, but G(*syn*):A(*anti*) has much lower probability. Since enzyme only recognizes G(*syn*):A(*anti*), the lower probability of G(*syn*):A(*anti*) means lower enzymatic activity. This further underscores the importance of the shape at the mismatch site for the initial recognition. Even so, it is unlikely that shape complementary alone directs the binding of mismatch to MutY. As has been proposed previously[131,133], more dynamic features resulting from the damaged 8oxoG base, such as increased bending and flexibility, are also believed to play important roles in the initial binding.

## 4.4 Conclusion

In the present work, we studied four different DNA 13-mer sequences with G:C, G:A, 8oxoG:C and 8oxoG:A pairs. Our simulation results confirmed the predominance of the normal *anti*:*anti* form of the 8oxoG:C base pair and the Hoogsteen *syn:anti* form of the 8oxoG:A pair. Particularly, we observed multiple *anti* → *syn* transition events that tend to follow two pathways. In order to gain further insight into the details of this structure transition and local structural fluctuations, we complemented the non-equilibrium simulations with a detailed study of the thermodynamic properties of this system. Our modified REM was used to construct the free energy landscape with the 8oxoG glycosidic and base flipping angles as two reaction coordinates. The resulting free energy landscape is shown to be consistent with our MD results. The combination of free dynamics and the thermodynamic data from REM provides new insights into the dynamic behavior of this system and how this behavior is affected by the chemical modifications involved in the oxidative damage.

We also carried out PMF calculations for several sequences and demonstrated that 8oxoG:A greatly decreases the barrier of 8oxoG flipping, whereas 8oxoG:C has comparable barrier of the standard G:C sequence.

In addition, as shown by our simulations, the flanking base pairs have profound effects on the rotation of glycosidic bond. Replacement of AT pair with more stable GC pair on 3’ side of 8oxoG:A will direct most of *anti* → *syn* transitions through the base flipping pathway.

Finally, we present a preliminary recognition mode based on our MD simulation, which reveals that the special *syn* orientation of 8oxoG in 8oxoG:A is primarily responsible for the initial damage recognition although significant bending and local structural fluctuations at the DNA damage site also contribute to the enzyme activity.

# Bibliography

1)van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem. Int. Ed. in English* **1990**, *29*, 992-1023.

2)Karplus, M.; McCammon, A. *Nature Struct. Biol.* **2002**, *9*, 646.

3)McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585-590.

4)Verlet, L. *Phys. Rev.* **1967**, *159*, 98.

5)Brooks, C. L. *Acc. Chem. Res.* **2002**, *35*, 447-454.

6)Shea, J. E.; Brooks, C. L. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499-535.

7)Bockmann, R.; Grubmuller, H. *Nature Struct. Biol.* **2002**, 198-202.

8)Groot, B. L. d.; Grubmuller, H. *Science* **2001**, *294*, 2353-2357.

9)Roux, B.; Allen, T. W.; Berneche, S.; Im, W. *Quat. Rev. Biophys.* **2004**, *37*, 15-103.

10)Viloca, M. G.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186-195.

11)MacKerell, A. D.; Feig, M.; III, C. L. B. *J. Am. Chem. Soc.* **2004**, *126*, 698-9.

12)Ponder, J.; Case, D. *Advances in Protein Chemistry* **2003**, *66*, 27-85.

13)Simmerling, C.; Strockbine, B.; Roitberg, A. *J. Am. Chem. Soc.* **2002**, *124*, 11258.

14)Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C. *J. Comput. Chem.* **2003**, *24*, 21-31.

15)Hansson, T.; Oostenbrink, C.; Gunsteren, W. F. v. *Curr. Opin. Struct. Biol.* **2002**, 190-196.

16)Tuckerman, M. E.; Martyna, G. J. *J. Phys. Chem. B* **2000**, *104*, 159-178.

17)Tai, K.; Bond, S. D.; MacMillan, H. R.; Baker, N. A.; Holst, M. J.; McCammon, J. A. *Biophys. J* **2003**, *84*, 2234-2241.

18)Tama, F.; III, C. L. B. *J. Mol. Biol.* **2002**, *321*, 297-305.

19)Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471-2474.

20)Sprik, M.; Hutter, J.; Parrinello, M. *J. Chem. Phys* **1996**, *103*, 1142.

21)sagnella, D. E.; Laasonen, K.; Klein, M. *Biophys. J.* **1996**, *71*.

22)Shen, T.; Tai, K.; Henchman, R. H.; AcCammon, J. A. *Acc. Chem. Res.* **2002**, *35*, 332-340.

23)Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

24)MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

25)Daura, X.; Mark, A. E.; vanGunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535-547.

26)Jorgensen, W. L.; Maxwell, D. S.; Tiradorives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

27)Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269-10280.

28)Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620-9631.

29)Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. *J. Phys. Chem. B* **2002**, *106*, 11673-11680.

30)Halgren, T. A.; Damm, W. *Curr. Opin. Struct. Biol* **2001**, 236-242.

31)Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926-935.

32)Berendsen, H. J. C.; Vangunsteren, W. F.; Postma, J.; Hermans, J. *Interaction models for water in relation to protein hydration.*; (editor), P. B., Ed.: Dordrecht, Reidel, 1981, pp 331-342.

33)Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335.

34)Cheatham, T. E.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193-4194.

35)Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1-20.

36)Simonson, T.; Brunger, A. T. *J. Phys. Chem.* **1994**, *98*, 4683-4694.

37)Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*.

38)Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144-1149.

39)Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435-445.

40)Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem.B* **2001**, *105*, 6507-6514.

41)Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993**, *97*, 3591-3600.

42)Roux, B.; Yu, H. A.; Karplus, M. *J. Phys. Chem.* **1990**, *94*, 4683-4688.

43)Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127-6129.

44)Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275-291.

45)Cornell, W.; Abseher, R.; Nilges, M.; Case, D. A. *J. Mol. Graphics & Modelling* **2001**, *19*, 136-145.

46)Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489-2498.

47)Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238-250.

48)Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578-1599.

49)Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045-1066.

50)Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122-129.

51)Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161-9175.

52)Roitberg, A.; Elber, R. *J. Chem. Phys.* **1991**, *95*, 9277-9287.

53)Stultz, C. M.; Karplus, M. *J. Chem. Phys.* **1998**, *109*, 8809-8815.

54)Neuhaus, B. A. B. a. T. *Phys Lett. B* **1991**, *267*, 249-253.

55)Hannsmann UHE, O. Y., Eisenmenger F *Chem. Phys. Lett.* **1996**, 321-330.

56)Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776-1783.

57)Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96-123.

58)Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141-151.

59)Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *329*, 261-270.

60)Torrie, G. M.; Valleau, J. P. *J. Compt. Phys.* **1977**, *23*, 187-199.

61)Schlitter, J.; Engels, M.; Kruger, P. *J. Mol. Graphics* **1994**, *12*, 84-89.

62)Molnar, F.; Norris, L. S.; Schulten, K. *J. Mol. Graphics & Modelling* **1998**, *16*, 294-295.

63)Wu, X. W.; Wang, S. M. *J. Phys. Chem.B* **1998**, *102*, 7238-7250.

64)Chandler, D. *Finding transition pathways: throwing ropes over rough mountain passes, in the dark*; Berne, B. J., Ciccotti, G. and Coker, D. F., Ed.; World Sci: Singapore, 1998, pp 51-66.

65)Fukunishi H, W. O., Takada S *J. Chem. Phys.* **2002**, *116*, 9058-9067.

66)Stivers, J. T.; Jiang, Y. L. *Chem. Rev.* **2003**, *103*, 2729 -2760.

67)Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Interation models for water in relation to protein hydration*; B, P., Ed.: Dordrecht, Reidel, 1981, pp 331-342.

68)Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

69)Zhou, R. H.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U S A.* **2001**, *98*, 14931-14936.

70)Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 345-354.

71)Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161-2200.

72)Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243-252.

73)Williams, D. J.; Hall, K. B. *Biophys. J.* **1999**, *76*, 3192-3205.

74)Miller, J. L.; Kollman, P. A. *J. Mol. Biol.* **1997**, *270*, 436-450.

75)Czerminski, R.; Elber, R. *Proteins: Struct., Funct., Genet.* **1991**, *10*, 70-80.

76)Simmerling, C.; Elber, R. *J. Am. Chem. Soc.* **1994**, *116*, 2534-2547.

77)Simmerling, C. L.; Elber, R. *Proc. Natl. Acad. Sci. U S A.* **1995**, *92*, 3190-3193.

78)Hornak, V.; Simmerling, C. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 577.

79)Case, D. A.; Pearlman, D. A.; Caldwell, J. A.; Cheatham, T. E.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 6*; University of California: SanFrancisco, 1999.

80)Cui, G.; Simmerling, C. *J. Am. Chem. Soc.* **2002**, *124*, 12154.

81)Varani, G.; Cheong, G.; Tinoco, I. J. *Biochemistry* **1991**, *30*, 3280-3289.

82)Allain, F. H. T.; Varani, G. *J. Mol. Biol.* **1995**, *250*, 333-353.

83)Simmerling, C.; Miller, J. L.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 7149-7155.

84)Guvench, O.; Shenkin, P.; Kolossvary, I.; Still, W. C. *J. Comput. Chem.* **2002**, *23*, 214-221.

85)Onufriev, A.; Case, D.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297-1304.

86)Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327-341.

87)Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8*, 1016-1024.

88)Lolis, E.; Alber, T.; Davenport, R. C.; Rose, D.; Hartman, F. C.; Petsko, G. A. *Biochemistry* **1990**, *29*, 6609.

89)Thirumalai, D.; Mountain, R. D.; Kirkpatrick, T. R. *Phys. Rev. A* **1989**, *39*, 3563-3574.

90)Thirumalai, D.; Mountain, R. D. *Phys. Rev. A* **1990**, *42*, 4574-4587.

91)Strab, J. E.; Thirumalai, D. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 360-373.

92)Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401-9409.

93)Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889-897.

94)Straub, J. E.; Karplus, M. *J. Chem. Phys.* **1991**, *94*, 6737-6739.

95)Straub, J. E.; Lim, C.; Karplus, M. *J. Am. Chem. Soc.* **1994**, *116*, 2591-2599.

96)Weston, R. E.; Schwarz, H. A. *Chemical Kinetics*; Prentice Hall: New York, 1972.

97)Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 11-27.

98)Cheng, X.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2004**, *108*.

99)Straub, J. E.; Rashkin, A. B.; Thirumalai, D. *J. Am. Chem. Soc.* **1994**, *116*, 2049-2063.

100)Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181-189.

101)Wolynes PG; JN, O.; D, T. *Science* **1995**, *267*, 1619.

102)Straub, J. E. *Protein Folding and Optimization Algorithms*; P. v. R. Schleyer, N. L. A., T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner, Ed.; John Wiley & Sons: Chichester, 1998; Vol. vol. 3, pp 2184-2191.

103)Ulitsky, A.; Elber, R. *J. Chem. Phys.* **1993**, *98*, 3380-3388.

104)Zheng, W. M.; Zheng, Q. *J. Chem. Phys.* **1997**, *106*, 1191-1194.

105)Keasar, C.; Elber, R. *J. Phys. Chem.* **1995**, *99*, 11550-11556.

106)Zheng, Q.; Rosenfeld, R.; Delisi, C.; Kyle, D. J. *Protein Science* **1994**, *3*, 493-506.

107)Verkhivker, G.; Elber, R.; Nowak, W. *J. Chem. Phys.* **1992**, *97*, 7838-7841.

108)Joseph-McCarthy, D.; Tsang, S. K.; Filman, D. J.; Hogle, J. M.; Karplus, M. *J. Am. Chem. Soc.* **2001**, *123*, 12758-12769.

109)Sugita, Y.; Okamoto, Y. *Progress of Theoretical Physics Supplement* **2000**, 402-403.

110)Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042-6051.

111)Rhee YM, P. V. *Biophysical J.* **2003**, *84*, 775-786.

112)Mitsutake A, S. Y., Okamoto Y *J Chem Phys* **2003**, *118*, 6664-6688.

113)Jang, S.; Shin, S.; Pak, Y. *Phys. Rev. Lett.* **2003**, *91*, 58305.

114)Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

115)Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684-3690.

116)Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: London, 1987.

117)McQuarrie, D. A. *Statistical Mechanics*; Harper & Row: New York, 1976.

118)Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 13898-13903.

119)Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7587-7592.

120)Flory, P. J. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, 1953.

121)Ichiye, T.; Karplus, M. *Biophys. J.* **1984**, *45*, A377-A377.

122)Teeter, M. M.; Case, D. A. *J. Phys. Chem.* **1990**, *94*, 8091-8097.

123)Hess, B. *Phys. Rev. E* **2000**, *62*, 8438-8448.

124)Burrows, C. J.; Muller, J. G. *Chem. Rev.* **1998**, *98*, 1109-1152.

125)Bruner, S. D.; Norman, D. P. G.; Verdine, G. L. *Nature* **2000**, *403*, 859-866.

126)Gao, X.; Patel, D. J. *J. Am. Chem. Soc.* **1988**, *110*, 5178-5182.

127)Gilboa, R.; Zharkov, D. O.; Golan, G.; Fernandes, A. S.; Gerchman, S. E.; E. Matz, J. H. K.; Grollman, A. P.; Shoham, G. *J. Biol. Chem.* **2002**, 19811-19816.

128)Fromme, J. C.; Banerjee, A.; Huang, S. J.; Verdine, G. L. *Nature* **2004**, *427*, 652-656.

129)Michaels, M. L.; Tchou, J.; Grollman, A. P.; Miller, J. H. *Biochemistry* **1992**, *31*, 10964-8.

130)Grollman, A. P. *Struct. Funct., Proc. Conversation Discip. Biomol. Stereodyn.* **1992**, 165.

131)Fuxreiter, M.; Luo, N.; Jedlovszky, P.; Simon, I.; Osman, R. *J. Mol. Biol.* **2002**, *323*, 823-834.

132)Francis, A. W.; Helquist, S. A.; Kool, E. T.; David, S. S. *J. Am. Chem. Soc.* **2003**, *125*, 16235-16242.

133)Miller, J. H.; Chiang, C. P.; Straatsma, T. P.; Kennedy, M. A. *J. Am. Chem. Soc.* **2003**, *125*, 6331-6336.

134)Case, D. A.; Karplus, M. *J. Mol. Biol.* **1978**, *123*, 697-701.

135)Banci, L.; Bertini, I.; Carloni, P.; Luchinat, C.; Orioli, P. L. *J. Am. Chem. Soc.* **1992**, *114*, 10683-10689.

136)Kumar, S., Bouzida, D. , Swendsen, R. H. , Kollman, P. A. & Rosenberg, J. M. *J. Comput. Chem.* **1992**, 1011-1021.

137)Huang N; Banavali NK; Jr., M. A. *Proc. Natl. Acad. Sci. U S A.* **2003**, *100*, 68-73.

138)Lavery, R.; Sklenar, H. *J. Biomol. Struct & Dyn.* **1989**, *6*, 655-667.

139)Strahs, D.; Schlick, T. *J. Mol. Biol.* **2000**, *301*, 643-666.

140)Fromme, J. C.; Verdine, G. L. *EMBO J.* **2003**, *22*, 3461-3471.

141)Hollis, T.; Ichikawa, Y.; Ellenberger, T. *EMBO J.* **2000**, *19*, 758-766.

142)Noll, D. M.; Gogos, A.; Granek, J. A.; Clarke, N. D. *Biochemistry* **1999**, *38*, 6374.

143)Porello, S. L.; Leyes, A. E.; David, S. S. *Biochemistry* **1998**, *37*, 14756-14764.