This is to certify that the
dissertation entitled

DEVELOPMENT OF SPECIAL MOLECULAR DYNAMICS BASED METHODS TO ACCELERATE SAMPLING OF IMPORTANT PROTEIN MOTIONS

presented by

Li Su

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Chemistry

Robert Cukier

Major Professor's Signature

August 18, 2009

Date

DEVELOPMENT OF SPECIAL MOLECULAR DYNAMICS BASED METHODS TO ACCELERATE SAMPLING OF IMPORTANT PROTEIN MOTIONS

By

Li Su

A DISSERTATION

Submitted to

Michigan State University
In partial fulfillment of the requirements
for degree of

DOCTOR OF PHILOSOHPY

Chemistry

2009

ABSTRACT

DEVELOPMENT OF SPECIAL MOLECULAR DYNAMICS BASED METHODS TO ACCELERATE SAMPLING OF IMPORTANT PROTEIN MOTIONS

By

Li Su

Molecular Dynamics (MD) applied to complex systems such as proteins that have energetic barriers separating different configurations may suffer from slow sampling of the configuration space. Therefore, new methods are needed to improve the ability of exploring conformational space. One way to accelerate the exploration of configuration space is with various Hamiltonian replica exchange methods (HREM) whereby multiple systems differing in their Hamiltonians are run by MD in parallel and, periodically, attempts at exchanging them is made according to a Monte Carlo rule that maintains the Boltzmann distribution.

The HREM (implemented in the CUKMODY MD code designed for the efficient simulations of solvated proteins) is first used to study the conformational states of the zwitterionic form of the pentapeptide Met-enkephalin. There is a competition between open forms of the peptide driven by polar solvation of the terminal ammonium and carboxylate groups and closed forms driven by their salt-bridge formation. Normal MD started from an open state does not sample closed conformations. A small number of HREM systems were found to be sufficient to sample closed and open states a sufficient number of times to obtain potentials of mean force along various reaction coordinates. A Principal Component Analysis (PCA) shows that the first two principal modes capture more than one-half of the HREM generated fluctuations. The first mode corresponds to the end-to-end distance fluctuations and shows that the closed zwitterionic state is the

predominant species. The second mode describes the presence of two conformations of similar end-to-end distance that differ in the values of neighboring psi and phi dihedral angles, because such psi/phi compensation can still produce the same end-to-end distance.

HPPK (6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase) catalyzes the transfer of pyrophosphate from ATP to HP (6-hydroxymethyl-7,8-dihydropterin). This first reaction in the folate biosynthetic pathway is an important target for potential antimicrobial agents. The conformations of the HP binding pockets of E. coli HPPK and Y. pestis HPPK are studied by HREM methods. Root mean square fluctuation (RMSF) calculated based on backbone atoms shows that loop 2 (which is close to the HP binding pocket) of YpHPPK has a much larger flexibility than that of loop 2 of EcHPPK. By using clustering methods, we find that EcHPPK and YpHPPK have residue conformations around the HP binding pocket that are are close to but distinct from those found in the crystal structures. There are near-closed conformations that have different HP binding pocket shapes in EcHPPK and YpHPPK that have potential for discriminating among ligands.

The conformational space of ATP binding to HPPK and the unbinding process of ATP from HPPK is studied using a restraint MD method and a targeted reweighting scheme (implemented in CUKMODY). ATP remains remarkably stable in its binding pocket when HPPK is driven, by using the restraint method, towards its open form. When the ATP is induced to leave its binding pocket by using a targeted reweighting method, it uses a special path that preserves the hydrogen bonds and salt bridges existing in previous stages along the path.

Dedicated to

my family in China and my life at MSU

Table of Contents

List of Tables

List of Figures

Images in this dissertation are presented in color.

Chapter 1. Introduction

1. Molecular Dynamics Simulation

General description

Molecular Dynamics[1-3] is a form of computer simulation for studying molecular motion by allowing atoms to interact under force fields, which are developed to approximate certain known physics. Normally, MD simulation can provide dynamic information of a system at atomic detail. (The method only deals with the classical Newtonian forces and does not explicitly take into account any quantum level details.) Thus, in principle, MD simulation can be used to answer any questions about a model system. In other words, in principle, not only any equilibrium property of the model system can be evaluated by calculating a trajectory time average based on the ergodic hypothesis (in practice, the MD trajectory usually only provides a sample of the ensemble) , but also the real dynamics of the system can be obtained. By examining the trajectory, not only can one calculate average system properties such as free energy but also one can learn the main functional motions of a protein. Of course, in practice, a study based on MD simulation will be limited by the time-scale affordable under current MD technique and the correctness of the force field parameters in use.

The algorithms used to implement MD simulation are all based on answering a fundamental question[3]: Given the current configuration of the system, what is the next one? MD propagates the model system in time by answering the above question repeatedly. As the method only deals with classical dynamics, MD propagates the system following the classical equations of motion (equations 1.1 and 1.2):

$$\dot{p}_i = -\frac{\partial H}{\partial r_i} = -\frac{\partial V}{\partial r_i} = f_i \tag{1.1}$$

$$\dot{r}_i = -\frac{\partial H}{\partial p_i} = \frac{p_i}{m_i} \tag{1.2}$$

where, r_i , p_i . f_i and m_i are the coordinate, momentum, force and mass of particle *i*, respectively, while H and V are the relevant Hamiltonian and potential. In practice for MD, the above first-order differential equations are usually solved for each particle to obtain better accuracy solutions of the equations of motion although the second-order Newton's Equation (equation 1.3):

$$f_i = m_i \cdot \ddot{r}_i \tag{1.3}$$

are sometimes used for implementation due to efficiency concerns. Therefore, in general, after the composition of the system (including the number of atoms and their types, masses and interaction potentials) are specified, there are three steps involved in the process used to propagate the system. First, one needs to assign initial values of positions and velocities to the particles within the model system. In terms of biological systems, simulation is usually started out with the crystal structures of the molecules of interest and, when conducting explicit solvent simulation, solvents can be added by either putting the system in a lattice box or an equilibrated box of solvent molecules. Then, the initial velocities are usually either generated randomly from a Maxwell distribution for a specified temperature[3] or assigned to a special value (typically zero). Second, forces are evaluated according to the atoms' interaction potentials used to define the system. In fact, potentials are defined through choosing developed force fields. Third, an integrator is

used to predict the next configuration given the chosen time step. The second and third steps are repeated until reaching the pre-set number of integration steps.

Force Field

In MD, a molecule is generally described as a series of charged points (atoms) linked by springs (bonds). Therefore, the relevant potential interactions in a MD simulation can be decomposed into the following interaction terms: bonding terms for bond length vibration, bond angle vibration, and proper torsion and improper torsion angles, and non-bonded Lennard-Jones and electrostatic terms. These terms must be parameterized to obtain a force field. The parameterization is obtained by reproducing molecular geometry and selected properties of tested systems and the parameter values are chosen to fit experimental results or ab initio quantum data. By now, a substantial number of force fields have been developed, including some popular ones for biological systems such as the AMBER[4], CHARMM[5] and GROMOS[6] force fields.

Integrator

Many integrators[3] have been developed for MD simulation. One of the most popularly used integrators is the leapfrog algorithm which is a modified version of the standard Verlet algorithm mainly used for integrating the second order Newtonian equations. The Verlet algorithm uses the positions and accelerations at the time *t* and the positions at the time t - Δt to predict the positions at the time t + Δt, where Δt is the integration step (time step). By using a Taylor expansion, it can easily be derived that the errors in the atomic positions and velocities are of the order of Δt^4 and Δt^2, respectively. To obtain more accurate velocities, the leapfrog algorithm is used, using velocities at half

time step. The following two equations (equations 1.4 and 1.5), in which r represents atom positions, t represents time and Δt represents one time step, show the leap-frog algorithm[7].

$$r\left(t+\Delta t\right)=r\left(t\right)+\dot{r}\left(t-\frac{\Delta t}{2}\right)\Delta t \tag{1.4}$$

$$\dot{r}\left(t+\frac{\Delta t}{2}\right)=\dot{r}\left(t-\frac{\Delta t}{2}\right)+\ddot{r}\left(t\right)\Delta t \tag{1.5}$$

One thing worth noting is that the integration step size should be chosen as large as possible to make the simulation efficient but small enough to catch the fastest motion (As a rule of thumb, roughly the step size needs to be set as about 1/10 of the time scale of the fastest motion.) Normally bonds involving hydrogen atoms vibrate in the scale of 10 fs. Therefore, in order to make the MD simulation more efficient, the SHAKE[8] algorithm is usually used to constrain at least hydrogen related bonds so that the proper time step size can be increased to 2 fs. Since, for the normal simulation time scale, those quick bond vibration motions will be averaged out, using the above constraints should be a good approximation.

Temperature and pressure control

When only solving the Newton's equation with no temperature modification, then the total energy of the system is conserved. (This condition is often used as a measure to check the accuracy of the integrators used.) Therefore, without temperature modification, the MD simulation is in fact performed in the microcanonical (fixed number (N), volume (V) and energy (E)) ensemble. Although MD simulations within the microcanonical ensemble are the easiest to perform, in most situations, the canonical (fixed number (N),

volume (V) and temperature (T)) or the isothermal-isobaric (fixed number (N), temperature (T) and pressure (P)) ensemble is more appropriate for normal experimental conditions. Therefore, we need methods to control temperature and pressure.

A number of methods have been developed to control temperature or pressure [9]. One popularly used method is the Berendsen method[10]; the essential idea behind the method is that, in fact, the system in simulation is not isolated but coupled to an external bath. So, Berendsen at al proposed the following equation of motion with a modification on velocities of atoms, $\ddot{r}_i = \frac{f_i}{m_i} + \frac{1}{2\tau}(\frac{T_{ref}}{T} - 1)\dot{r}_i$, where, τ is the coupling time, and T_{ref} is the reference temperature, which is the temperature pre-chosen for the external bath. T is the instantaneous temperature, which is equal to $\frac{2}{k_B(Nd)}K$, where k_B is the Boltzmann constant, N_d is the number of degrees of freedom in the system and the kinetic energy is $K = \sum_{i=1}^{N} \frac{1}{2} m_i \dot{r}_i^2$. The extra term added in the above equation, in fact, serves as a frictional force and, hence, the actual implementation of the method simply employs a feedback mechanism to control the temperature. The velocity is scaled by a factor $\sigma = \sqrt{1 - \frac{\Delta t}{\tau}\frac{T - T_{ref}}{T}}$. Usually, the coupling time, τ, is set to around 0.2 ps. The principle behind controlling pressure is similar to that behind controlling temperature. The instantaneous pressure P is calculated as $P = \frac{2}{3V}(K - Virial)$ with V the volume, K the kinetic energy and *Virial* the virial function[3]. Thus, for an isotropic system, the pressure is controlled by scaling the box side and all atoms' positions by a

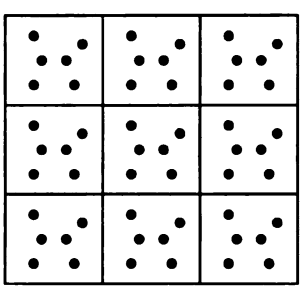
$\text{factor}\,\gamma = \left(1 - \frac{\Delta t}{\beta}\frac{P - P_{ref}}{P}\right)^{-\frac{1}{3}}$, where β is the coupling time for pressure control. The implementation of pressure control is much more complex for anisotropic box, in which case a pressure tensor has to be used[10]. As shown above, the implementation of Berendsen method is quite simple, but the compromise for this simplicity in implementation is that the Berendsen thermostats, in theory, do not lead to any known ensemble. To guarantee an NVT or NPT ensemble theoretically, more complicated methods such as Langevin dynamics method[11] or Nose-Hoover method[12] need to be introduced.

Periodic Boundary Conditions (PBC)

In mathematical models and computer simulations, periodic boundary conditions (PBC) are a set of boundary conditions that are often used to simulate an infinite system by using only a finite size box. Since usually only a small box of atoms (usually referenced as primary box) can be afforded in practical MD simulation, the boundary particles' interactions are a serious concern. The PBC is introduced to help address this difficulty.[3] In the PBC approximation, an infinite system is represented as a periodically repeated array of the finite small system being simulated. A schematic diagram for a two-dimensional system is shown in the following Figure 1.1. Clearly by using PBC, certain artificial periodicity (as shown in Figure 1.1, the simulated system is constructed by putting periodic images around the primary box, so there clearly are certain periodicities in the constructed system which might not exist in the real system) is imposed on the system under study. But, it can be shown that this system can still represent a real system

provided that an appropriate box size is used.[3] In conjunction with using PBC, the long-range interactions can be calculated approximately by using cut-off methods or, more accurately, by using an Ewald Lattice Sum.[3] Details of calculation of long range interactions are given in the next paragraph.

Figure 1.1. Illustration of Periodic Boundary Conditions (The shadowed box is the primary box and the boxes around it are its periodic images.)



Calculation of long-range (non-bonding) interactions

For most current MD simulation force fields (including the one described in more detail before), the non-bonded energy V_{nb} for the molecular system can be written as a sum of pair-wise electrostatic and Lennard-Jones contributions as shown in equation 1.6:

$$V_{nb} = \sum_{i,j} \left(\frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \tag{1.6}$$

When conducting simulation using PBC concept, as far as the system is determined, all its images are determined. Therefore, in theory, all the long-range interactions can be calculated by the above expression. But, in practice, it can be seen that, in the first term of the above equation representing the electrostatic contribution, the

order of r_{ij}, the distance between two atoms, is -1. It can be shown that a summation of the long-range interactions whose energy has the form of $\frac{Kq_iq_j}{r^p}$ with $p \leq 3$, converges very slowly.[13](Of course, if $p \leq 1$, the summation does not converge at all.) Probably the easiest method to solve this problem in calculation of those non-bonding interactions is to use a cut-off (truncation) method, which simply ignores all interactions beyond some chosen threshold distance (typically 10 Å). But later it was found that for proper simulation of many biological systems more efficient methods, such as Ewald summation[3]are needed.[14] Using an Ewald summation method would allow the long range interactions to be calculated fully for a periodic system. The Ewald method is implemented by dividing the summation calculations into two parts, within the so-called reciprocal space and within the so-called real space. Then, assuming a charge neutral system, the two summations both converge fairly rapidly given the special formula developed by Ewald in 1921.[3] A naive implementation of the Ewald summation method would be implemented in $O(N^2)$, which is too expensive for many biological systems such as proteins simulated in explicit solvents. There are specialized implementations of the Ewald method, which can run as $O(N \log N)$. A popular implementation is called the Particle Mesh Ewald (PME) method[15], which makes interpolation of the reciprocal space structure factor in the lattice points using Cardinal BSpline interpolation method and utilizes the popular Fast Discrete Fourier Transform (FDFT) method.

2. Methods for enhanced sampling

General Introduction

Although, in theory, MD simulation can provide all structural, equilibrium and dynamic information of a system under study at atomic detail, the quality of the information obtained is limited by the quality of the force fields in use and the extent of the coverage in the sample space. In our work we have focused on the sampling problem (In other words, the force field issue, which is another important problem cumbering effective application of MD simulation, is not the focus for the work summarized in this dissertation.). The current affordable time scale for conducting MD simulation through integrating Newton's equation of motion with a femtosecond time step is usually only tens of nanoseconds for a normal-sized protein simulated under explicit solvents unless large parallel machines are used so that hundreds of nanoseconds or a microsecond is achievable. ,Thus, using MD simulations to study conformational changes that may occur on micro second to seconds time scales is problematic.[16] (Some typical motions in proteins are listed in the following Table 1.1.) Proteins usually have complex energy landscapes with energy barriers many times the thermal energy, RT (for room temperature at 300 K, RT is around 0.6 kcal/mol), while going over barriers is exponentially hard according to the Arrhenius equation, $k = Ae^{-E_a / RT}$, where k is rate constant, A is a pre-factor, and E_a is the activation energy. Therefore, structures are frequently trapped in local energy minima where the thermal energy of the system is insufficient to traverse the energy barriers in normal MD simulation times. Special techniques, such as the replica exchange method (REM) and the reweighting method, have been developed to, respectively, help surmount or reduce those barriers among

minima so that, by using those methods appropriately, a better sampling within the given affordable amount of time might be obtained.

Table 1.1. Typical motions in proteins

Protein Motion	Log_{10} of Characteristic times (s)
Local denaturation and opening of secondary structure	3 to -1
Aromatic side-chain rotation	1 to -5
Conformational changes and allosteric trasitions	0 to -5
Diffusion	-6 to -9
Segmental motion	-7 to -9
Aliphatic side-chain rotation	-8 to -9
Vibrational and torsional modes	-11 to -13

Replica Exchange Methods

The Replica exchange method (REM)[17-24] is a popularly used method designed to address the sampling issue. The method was first developed for the simulations of spin systems [22] and now is widely used to study peptides and proteins. The first version of the REM introduced into the MD simulation field was temperature replica exchange method (TREM), and recently a Hamiltonian REM (HREM) was introduced into the field by Fukunishi et al.[25] The essential algorithm for all the REM approaches contains two parts. In part one, each REM system (a REM system is characterized by a certain property, e.g., temperature for TREM or Hamiltonian for HREM) containing a certain configuration is simulated simultaneously and independently for a certain number of MD steps. In part

two, exchange attempts are made among systems according to a Metropolis-Hastings algorithm.[9] (To improve the effectiveness, usually exchange attempts are only conducted between adjacent systems). In the TREM, the low temperature systems can borrow fast equilibration properties from high temperature systems, providing doorways for those low temperature systems to overcome energy barriers and thus improve the sampling.[26] Of course, the sampling of the low temperature system will not be guaranteed to improve through exchanging with high temperature systems because, for example, the high temperature system may simulate the same sample space as the low temperature system or the high temperature system wastes most time on sampling unproductive regions in the sample space. The TREM suffers from the deficiency that the number of system copies needed is approximately proportional to the square root of the number of degrees of freedom of the system of interest.[25] The HREM was introduced to address the above-mentioned difficulty with TREM.[25], In HREM, the potential energy function in different Hamiltonians for the systems differ only in a limited set of the total number of degrees of freedom required to characterize a system, thereby, in principal, reducing the number of systems needed. Of course, the basic issue of implementing HREM becomes choosing which degrees of freedom to focus on.

Reweighting

The essential idea underlying a reweighting method[27-31], which is a form of umbrella sampling[9], is that we can use another potential to run a trajectory, and then by reweighting back we can still get the original ensemble average of any configurational property. The ensemble average of any property $A\left(r^N\right)$, which is a function of the system configuration r^N, can be calculated as shown in equation 1.7[30]

$$\left\langle A\left(r^N\right)\right\rangle = \frac{\int dr^N A\left(r^N\right) e^{-\beta V\left(r^N\right)}}{\int dr^N e^{-\beta V\left(r^N\right)}} = \frac{\int dr^N \{A\left(r^N\right) e^{+\beta\Delta V\left(r^N\right)}\} e^{-\beta V^*\left(r^N\right)}}{\int dr^N \{e^{+\beta\Delta V\left(r^N\right)}\} e^{-\beta V^*\left(r^N\right)}}$$

$$= \left\langle A\left(r^N\right) e^{+\beta\Delta V\left(r^N\right)} \right\rangle^* \Big/ \left\langle e^{+\beta\Delta V\left(r^N\right)} \right\rangle^*$$

$$= \frac{\lim_{T\to\infty} \int_{t_0}^{t_0+T} A\left(r_*^N(s)\right) e^{+\beta\Delta V\left(r_*^N(s)\right)} ds}{\lim_{T\to\infty} \int_{t_0}^{t_0+T} e^{+\beta\Delta V\left(r_*^N(s)\right)} ds} \tag{1.7}$$

where, the potential energy $V = (V + \Delta V) - \Delta V \equiv V^* - \Delta V$ defines V^*, a modified potential energy surface upon which the dynamics is carried out, <..>* denotes an average with modified (V^*) weighting and $r_*^N(s)$ $(t_0 \le s \le t_0 + T)$ denotes the V^* modified trajectory over the simulation time interval, *T*. The last identity indicates that, with the assumption of ergodicity, the ensemble averages are to be evaluated as time averages over the trajectories generated from the modified surface. At each step of the dynamics on the modified surface, the data can be "re-weighted" with the factor $\exp(\beta\Delta V)$ to guarantee that the average is Boltzmann weighted. If barriers are reduced by use of the modified potential surface, the modified trajectory will explore the configuration space more rapidly.

Some successful examples of application of the method including protein folding[32], exploration of dihedral conformational space in the alanine dipeptide [33], and study of HIV-1 protease [34].

The most technical and difficult part for reweighting method is to choose V^*. Examples of V^* choices from some developed reweighting methods[32-34] are shown in the following equations 1.8 to 1.10.

$$V^* = \exp[(-V(r^N) - V_{thre})/n_f] \tag{1.8}$$

$$\text{if } V_{thre} > V(r^N), \text{ then } V^* = \Delta V(r^N) + V(r^N)$$
$$\text{with}$$
$$\Delta V(r^N) = \frac{(V_{thre} - V(r^N))^2}{\alpha + (V_{thre} - V(r^N))} \tag{1.9}$$

$$V^* = 1 - \exp[(-\gamma V(r^N) - V_{thre})] \tag{1.10}$$

where, V is the true potential energy; V_{thre} is called a threshold energy and is a constant chosen before conducting simulation; n_f in eq. 1.8 is the number of degrees of freedom of the system; α and γ are constants used in equations 1,9 and 1.10, respectively, to further modify the energy landscape. Because a good choice of $V_{thre.}$ and α and γ, needs to be made based on some knowledge of the energy landscape, usually a substantial amount of trials need to be made before the final choice can be made on those values.

3. Methods to compute the potential of mean force (PMF)

To find a low energy path and barriers between different states, a potential of mean force (pmf)[35] is defined as: $W(\xi) = -kT \ln \rho(\xi) + W_0$, where, W_0 is an arbitrarily

chosen reference value to set a zero free energy; and $\rho(\xi)$ is the distribution function which is defined in following equation 1.11 as:

$$\rho(\xi) = \frac{\int dr^{3N} \delta(\xi'[r^{3N}] - \xi) \exp[-V(r^{3N})/kT]}{\int dr^{3N} \exp[-V(r^{3N})/kT]} \quad (1.11)$$

in which, V, the potential energy, depends on the $3N$ coordinates of the system, r^{3N}; ξ is a reaction coordinate, which is a function depending on a certain number of degrees of freedom in the system (e.g. an angle or a distance of interest). The $\delta(\xi'[r^{3N}] - \xi)$ is a delta function used to pick out the reaction coordinate value ξ of the snapshot being checked, $\xi'[r^{3N}]$. Of course, in practice, a certain range will be used to consider $\xi'[r^{3N}]$ being equal to ξ.

Usually, there will be some barriers along the chosen reaction coordinate that are difficult to overcome by normal MD simulation. Then, some artificial potential terms need to be added to help the sampling. One widely used scheme is the Umbrella Sampling scheme.[36] For applying this method, a restraining potential $U(\xi)$ (sometimes called umbrella potential or window potential) is used to help improving the sampling of specific regions around ξ. $U(\xi)$ can be whatever which is convenient and proper, though the most commonly used potential is the harmonic potential. Then, the chosen path along reaction coordinate will be divided into properly spacing windows to provide enough overlap in the reaction coordinate distributions for neighboring windows, and the expression of the restraint potential biased distribution for some window i can be represented as shown in equation 1.12:

$$\rho_i^b(\xi) = \exp(-\beta U_i(\xi)) \rho_i^u(\xi) < \exp(-\beta U_i(\xi)) >_i^{-1} \quad (1.12)$$

where, $\rho_i^u(\xi)$ is the unbiased (without the restraint potential) distribution for the i^{th} window. Substituting it into the $W(\xi)$ expression , the corresponding W_i can be obtained for each window, which is shown in following equation 1.13,

$$W_i(\xi) = -kT \ln \rho_i^b(\xi) + W_0 - U_i(\xi) + f_i \tag{1.13}$$

where, f_i is defined as

$$f_i = -kT \ln < \exp(-\beta U_i(\xi)) >_i \tag{1.14}$$

There have been numerous efforts spent to address problems of unbiasing and recombining the information gotten from all windows. WHAM[37,38] is currently the most popular method since all traditional methods, which calculate f_i by overlapping $\rho_i^u(\xi)$ as well as possible, only use data within overlapped area between neighboring $\rho_i^u(\xi)$ while wasting a substantial amount of data outside the overlapped area. WHAM instead linearly combines the $\rho_i^u(\xi)$ to one unbiased distribution function:

$$\rho^u(\xi) = \sum_{i=1}^{N} p_i(\xi) \rho_i^u(\zeta), \quad \sum_{i=1}^{N} p_i(\xi) = 1 \tag{1.15}$$

where

$$p_i(\xi) = \frac{n_i \exp(-\beta[U_i(\xi) - f_i])}{\sum_{j=1}^{N} n_j \exp(-\beta[U_i(\xi) - f_i])} \tag{1.16}$$

WHAM scheme is developed to make the statistical error minimal.[39] The f_i's can be obtained from:

$$\exp(-\beta f_i) = \int d\xi \rho^u(\xi)\exp(-\beta U_i(\xi))$$
$$= \int d\xi \frac{n_i \exp(-\beta U_i(\xi))}{\sum_{j=1}^{N} n_j \exp(-\beta[U_i(\xi) - f_i])} \rho_i^b(\xi) \quad (1.17)$$

By plugging eq.1.14 into eq 1,12 and making some rearranging, we can get the following equation:

$$\rho_i^u(\xi) = \exp(-\beta[U_i(\xi) - f_i])\rho_i^b(\xi) \quad (1.18)$$

Using the above four equations (equations 1.15 to 1.18), the f_i can be computed self-consistently (note $U_i(\xi)$ is the added window restraint potential which is known), and thus compute the unbiased distribution function $\rho^u(\xi)$ shown in eq. 1.15. And, it is clear that this method can be easily extended to higher dimensional distributions naturally. For example, for two dimensions, the only difference is that the iteration will be carried over f_{ij} instead of f_i. Though for practical issue, the number of data points needed for pmf construction exponentially depends upon the dimensionality.

4. Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is a statistical multivariate analysis method used in many fields[40]. The analysis method essentially transforms a number of possibly correlated variables into a smaller number of uncorrelated variables that are usually called principal components. The first principal component is constructed to account for as much of the variability in the data as possible, and then each succeeding component is

constructed to account for as much of the remaining variability as possible. In field of simulations of biological systems, PCA is essentially used to extract those modes with biggest fluctuations for further study.[41-45] One basic method to implement PCA is to diagonalize the covariance matrix for coordinates of atoms, $Cov_{ij} = < \delta r_i \delta r_j >$, where δr_i is defined as $r_i - < r_i >$, with $r_i = \{x_i, y_i, z_i\}$ denoting the Cartesian components of the position of the i^{th} atom.and <...> denoting the ensemble average. By diagonalizing, the orthonormal eigenvectors and corresponding eigenvalues of the covariance matrix can be obtained. Then the configuration point $r(t) = (x_1(t), y_1(t), .., z_N(t))$ can be decomposed as:

$$\mathbf{r}(t) = \sum_{i=1}^{3N} \left[\mathbf{r}(t) \circ \mathbf{m}_i \right] \mathbf{m}_i = \sum_{i=1}^{3N} p_i(t) \mathbf{m}_i \tag{1.19}$$

where the $\mathbf{m}_i$ are the orthonormal eigenvectors of the covariance matrix Cov_{ij}, and the $p_i(t)$ are the corresponding PCA mode displacements. Typically the first few "important" modes have much larger eigenvalues than the remaining modes. (Normally the number of remaining modes is much larger than that of the first few important modes). The first few modes are considered to be important since they often describe large scale motions such as conformational changes while the remaining modes with small eigenvalues usually correspond mostly to small Gaussian fluctuations.

5. Cluster Analysis

Cluster analysis or clustering is a common technique for statistical data analysis.[46] In essence, clustering is the assignment of objects into groups (referenced as clusters) so that objects from the same cluster are more similar to each other than objects from different clusters. Often this similarity is assessed according to a distance measure. One widely used clustering method in analysis of simulated trajectories for bio-systems is the "g_cluster" method[47], which uses the root mean square distance (RMSD) as its similarity measure. For each input structure, the method counts the number of neighbors within an RMSD cut-off, takes the structure with the largest number of neighbors, along with all its neighbors, as a cluster, and eliminates all the structures within this cluster from the pool of structures. The above steps are repeated for the remaining structures in the pool until no structures are left. There are other clustering methods such as k-means[48] or hierarchical clustering.[49]

6. Studied peptides and proteins

Met-Enkephalin

Met-enkephalin is an endogenous opioid peptide neurotransmitter found in the central nervous system and gastrointestinal tract where they bind to opiate receptors.[50,51] The pentapeptide (tyr-gly-gly-phe-met) is one of the two forms of known enkephalins and the other is leu-enkephalin (tyr-gly-gly-phe-leu). The opioid peptide neurotransmitters are shown to play important roles in pain mediation, opiate dependence, and euphoria.[52,53] After being identified in 1975, a substantial amount of studies have been performed to investigate the enkephalins. These studies include experimental

studies such as nuclear magnetic resonance (NMR)[53-55], infrared (IR)[56], ultraviolet (UV)[57,58], and circular dichroism (CD)[59,60], and numerous computations.[19,32,61-69] Almost all studies show that the enkephalins exhibit great conformational plasticity, and possibly because of the flexibility of the peptides, the exact dominant conformations of the peptides are still unknown (conflicting evidences for structures are obtained.) However, as will be discussed in more detail in Chapter 2, it seems to be hard to get good samples of met-enkephalin or leu-enkephalin with normal MD simulation. So some part of the work summarized in this dissertation is focused on using a special HREM approach to help improve sampling of met-enkephalin. The simulation work of met-enkephalin is being conducted in explicit water, because although the receptors of enkephalins are usually membrane proteins, the regions of those membrane bound receptors to which enkephalins bind are considered to probably be in the aqueous parts[70].

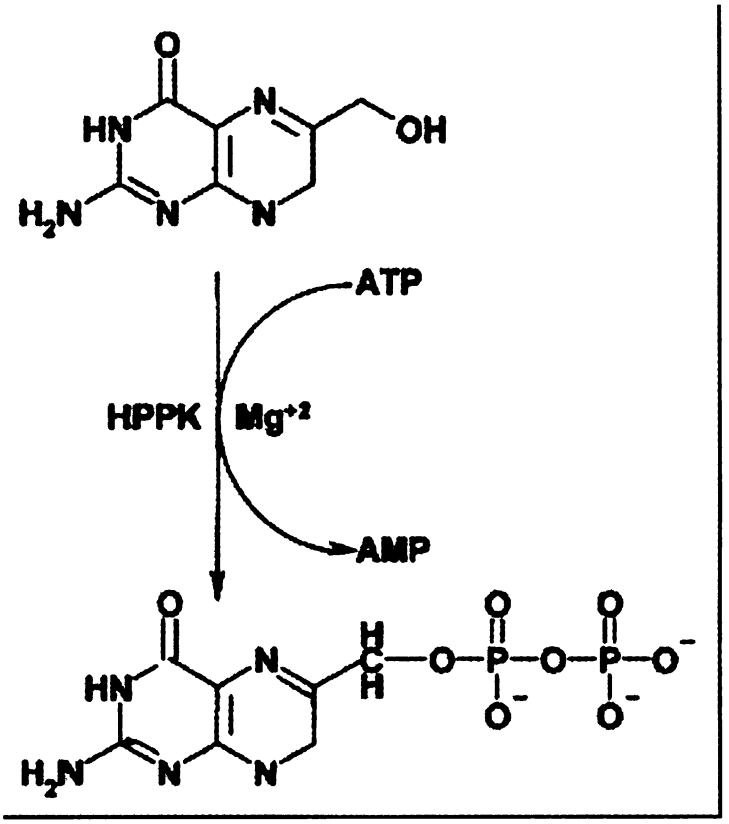
6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK)

Folate derivatives are essential cofactors for life; mammals, including humans, obtain folates from their environment, while most microorganisms must synthesize folates *de novo* via the folate biosynthetic pathway. Therefore, biosynthetic pathway of folates is an important target for the development of antibiotics to treat infections from microorganisms. Due to rapidly increasing antibiotic resistance in recent years which has rendered the current antibiotics ineffective for treating many microbial infections, resulting in a worldwide health care crisis, new antimicrobial agents are urgently needed to control the infections that are resistant to the current antibiotic drugs.[71]

6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK) catalyzes the transfer of pyrophosphate from ATP to 6-hydroxymethyl- 7,8-dihydropterin (HP)[72,73] and

is a new target in the folate biosynthesis pathway for developing antimicrobial agents. Its function mechanism is shown in the following figure[73]:

Figure 1.2. The function of HPPK in the folate biosynthesis pathway[73]



A number of studies have been conducted on HPPK[72-82], most of the studies focus on E. coli. HPPK.[72,74-79,81,82] The crystal structure of apo E.coli HPPK [74] reveals a three-layered α-β-α fold formed by six β-strands and four α-helices. The fold of the HPPK molecule creates a valley that is approximately 26 Å long, 10 Å wide, and 10 Å deep. Three flexible loops, β2-β3 (loop2), β1-α1 (loop1), and α2-β4 (loop3), form one wall of the valley. The other wall of the valley is relatively rigid and is constructed by the part of the protein's hydrophobic core. The crystal structure of the ternary complex E.coli HPPK

with AMPCPP (a non-reactive analog of ATP) and two associated Mg^{2+} ions and with HP shows significant conformational changes relative to the apo structure concentrated in the flexible loops[75] that serve to sequester the ligands in a catalytically competent form. In the ternary form[75], HP is sandwiched between two aromatic rings of Phe123 and Tyr53 and forms six hydrogen bonds with residues Thr42, Pro43, Leu45, and Asn55. Twelve residues are involved in binding AMPCPP, Gln74, Glu77, Arg84, Arg88, Try89, Arg92, Ile98, Arg110, Thr112, His115, Tyr116, and Arg121, of which Glu77, Arg92, His115, and Arg121 are conserved[77]. The crystal structure data on the binary complex of E. coli HPPK with AMPCPP show that this binary complex is trapped in a "super open" conformation where one of the loops (loop 3) is in an extended conformation.[76] NMR studies of HPPK with a related non-interactive ATP analog, AMPPCP present on average a different but also more open (than apo) conformation.[79] Actually, examination of the ensemble of structures reveals that some are in conformations that approach the ternary closed conformation. The NMR results suggest that the binary complex of HPPK with ATP most likely is labile, sampling both open and closed-like conformations.

In part of the work summarized in this dissertation, a special HREM approach is applied to study the HP binding pockets of E. coli HPPK and Y. pestis HPPK in the hope of discovering different important HP binding pocket conformations of the proteins, which might be used to develop new narrow band antibiotics.

Chapter 2. A specific Hamiltonian Replica Exchange Method (HREM) and its validation by application to Met-Enkephalin

1. Introduction

In the work covered in this chapter, the conformational space of Met-enkephalin is explored with the use of the Hamiltonian Replica Exchange Method (HREM) applied in explicit solvent MD simulations. As mentioned in Chapter 1, Met-enkephalin has been shown to exhibit great conformational plasticity by experiments[53-56,58,60] and computation.[19,32,61-69] Due to this plasticity and its practical importance as an important opioid peptide, Met-enkephalin is popularly used to test methods that can enhance the rate of conformational sampling. Of particular interest is the zwitterionic form (protonated N-terminus and ionized C-terminus), which should predominate in polar media such as water. The competition between "closed" forms, where the N and C termini are salt-bridged, and charge solvated "open" forms in which the terminal peptide charges interact with solvent dipoles may lead to reasonable co-existence of closed and open forms. A number of MD-based simulations of Met-enkephalin and the closely related Leu-enkephalin (both of them are enkephalins mentioned in Chapter 1) have been carried out. Berendsen et al.[63] noted that the zwitterionic form of Leu-enkephalin was quite labile but only sampled folded, closed forms during their simulation. Smith et al.[64] found at neutral pH their simulated zwitterionic form of Leu-enkephalin had a roughly equal mixture of close and open states. They also found that starting from an open form the peptide rapidly closed, but eventually re-opened on their 10 ns time scale (only once during the entire simulation). Nielsen et al.[65] simulated Met- and Leu-enkephalin in zwitterionic forms and found a rapid one-way transition from open to stable, closed

conformers. Freed et al.[66] compared the results of explicit and implicit solvent simulations and found mostly compact with some open conformations. Karvounis et al. [68] simulated the zwitterionic form of Leu-enkephalin in explicit solvent initiating a number of trajectories from open forms and found a variety of behaviors including a persistent salt-bridge form. Garcıa et al.[69] simulated neutral Met-enkephalin in explicit solvent using the Temperature Replica Exchange Method (TREM) with 17 replicas (we call replicas "systems" in this dissertation) used. Garcia's study demonstrated the enhanced sampling capability of a REM relative to conventional MD, showing that it could surmount the barriers separating non-helical from helical conformations in the simulation interval. One thing noticeable of the above mentioned simulation work is that, although the enkephalins are considered to be highly flexible according to experimental evidence (details are in Chapter 1), for the zwitterionic forms of enkephalins, which should be the predominant forms in water, no study has achieved a decent number of transitions between closed and open forms during their simulation time scales. Assuming the co-existence of the open and closed forms, then, to make meaningful statistical inferences of the enkephalins (required for the evaluation of potentials of mean force and equilibrium constants) in their conformational spaces not only requires sampling both forms but also requires achieving a large number of transitions between the open and closed forms during simulation. The co-existence is highly probable and supported by above-mentioned simulation results and experimental studies listed in Chapter 1. Therefore, in the work of this chapter, a special version of Hamiltonian Replica Exchange Method (HREM) was developed and applied to address the aforementioned sampling issue for Met-enkephalin.

2. Methodology

Hamiltonian Replica Exchange Method

As described in Chapter 1, the REM concept was first introduced as Temperature Replica Exchange Method (TREM), which suffers from the deficiency that the number of system copies needed could be very large for large system[25]. To solve ths problem, the REM concept was generalized to a Hamiltonian Replica Exchange Method[25] (HREM) where the systems differ by their Hamiltonians (in practice, systems usually differ only in their potential energy functions and their kinetic energy parts are left intact). As a matter of terminology, we shall refer to these different Hamiltonians as *systems* (versus replicas), since *replica* connotes a copy of an item. In fact, for HREM, it is more natural to use this terminology and the term replicas will be reserved for the configurations that are present on any particular MD step. In CUKMODY[83,84], the MD program developed in our group and the simulation program used for all the work in this dissertation, a given configuration (replica) is maintained on a particular computer node and the systems (with different potential functions) move onto and out of that node. The Hamiltonian for the i^{th} system within the extended HREM ensemble can be represented as, $H_i(\mathbf{X},\mathbf{P}) = T(\mathbf{P}) + V_i(\mathbf{X})$ where $T(\mathbf{P})$ is the kinetic energy and $V_i(\mathbf{X})$ is the potential energy function for the i^{th} system with phase space coordinates $\mathbf{X},\mathbf{P}$. Exchange attempts are made regularly at certain predetermined MD steps and, for the HREM implementation in this work, the attempt was made every 40 steps. Between exchange attempts, normal MD is run for each system. The exchanges may be thought of either as configuration exchanges or potential energy function exchanges, which will be scale-of-

interaction in this HREM implementation. Computationally, it is much more efficient that only a scale factor for the potential energy function needs to be exchanged, versus exchanging configurations. Exchanges are attempted only between neighboring systems, because for the method to be efficient the overlap between the systems' probability distributions needs to be adequate. (Efficiency can be roughly measured by the exchange acceptance probability, which is the fraction of successful exchange attempts, and is a compromise between the speed of motion and step size through the configuration space.) When system interchanges are attempted, detailed balance equations for pairs of neighboring systems

$$\alpha(\mathbf{X},\mathbf{X}' \to \mathbf{X}',\mathbf{X}) P_i(\mathbf{X}) P_j(\mathbf{X}') = \alpha(\mathbf{X}',\mathbf{X} \to \mathbf{X},\mathbf{X}') P_i(\mathbf{X}') P_j(\mathbf{X}) \quad (2.1)$$

are enforced. Here, $\alpha(\mathbf{X},\mathbf{X}' \to \mathbf{X}',\mathbf{X})$ is the acceptance probability (transition probability) that configuration $\mathbf{X}$ in the i^{th} system and $\mathbf{X}'$ in the j^{th} system before exchange results in configuration $\mathbf{X}'$ in the i^{th} system and $\mathbf{X}$ in the j^{th} system after exchange, and $P_i(\mathbf{X})$ is the Boltzmann distribution at temperature $T = 1/k_B\beta$ for the i^{th} system. The Metropolis rule for exchange between two systems,

$$\alpha(\mathbf{X},\mathbf{X}' \to \mathbf{X}',\mathbf{X}) = \min\left(1, e^{-\Delta(\mathbf{X},\mathbf{X}' \to \mathbf{X}',\mathbf{X})}\right) \quad (2.2)$$

where

$$\Delta(\mathbf{X},\mathbf{X}' \to \mathbf{X}',\mathbf{X}) = \beta\left[\left(V_i(\mathbf{X}') - V_j(\mathbf{X}')\right) + \left(V_j(\mathbf{X}) - V_i(\mathbf{X})\right)\right] \quad (2.3)$$

is used to impose the detailed balance equations to guarantee that Boltzmann equilibrium in the extended ensemble of the product of all the systems' ensembles will result for a

sufficiently long trajectory.[25] If the potential functions differ by a restricted set of degrees of freedom, only those will contribute to equation 2.3.

In the HREM implementation for the Met-enkephalin study, the potential energy is parameterized as

$$V_i(\mathbf{X}) = \lambda_i^2 V_{PP}(\mathbf{x}_P, \mathbf{x}_P) + \lambda_i V_{PS}(\mathbf{x}_P, \mathbf{x}_S) + V_{SS}(\mathbf{x}_S, \mathbf{x}_S), \quad (2.4)$$

where the terms in equation 2.4 denote peptide-peptide, peptide-solvent and solvent-solvent interactions, respectively, and λ_i is a scaling factor for the Lennard-Jones and electrostatic nonbonded interactions. In explicit solvent simulations, the number of degrees of freedom is dominated by the solvent. Thus, the indicated scaling is much reduced relative to the TREM where the global scaling $\beta V_i(\mathbf{X}) = \beta\lambda_i V(\mathbf{X})$ would be used.

The form of scaling in equation 2.4 is a requirement for the use of an Ewald method[15], where the evaluation of energy and force in the reciprocal space is based on a structure factor (a sum over atoms), and consequently necessitates assignment of the scale factor to *atoms,* versus to *atom-atom interactions*. The electrostatic and Lennard-Jones interactions are uniformly scaled; therefore, as λ_i decreases, both softer Lennard-Jones and reduced electrostatic interactions are obtained, permitting sampling enhancement.

One thing noticeable for REM schemes is the number of steps between two adjacent exchange attempts. Because by imposing the correct detailed balance equations, at equilibrium, all the systems in the extended ensemble will have corresponding Boltzmann distribution. The Boltzmann distributions are guaranteed when the whole

extended ensemble is at equilibrium, based only on the following two assumptions. Assumption one, the distributions of the systems within the extended ensemble are independent. Assumption two, the moving process of the whole extended ensemble is a Markov Chain process with only one equilibrium state. Therefore, for different choices of the number of steps between two adjacent exchange attempts, as long as the corresponding extended ensemble is at equilibrium (in practice, this "at equilibrium" is checked by checking whether the trajectories have converged), the systems should sample canonical ensembles. Of course, the optimal choices of the number of steps between two adjacent exchange attempts will be the one that make the trajectories converge most rapidly,

The normal potential energy function is being used in system 1 (λ_1=1), so that the trajectory associated with it samples the normal canonical ensemble. Therefore, all our analysis in the later Results and Discussion section was based on trajectories for system 1, and hence for the normal canonical ensemble.

Molecular Dynamics Simulation settings

The CUKMODY protein molecular dynamics code, which uses the GROMOS96[6] force field, was modified to incorporate the HREM, based on the previous DREM[83] code. The systems were run independently on different nodes of a Linux cluster computer and, when exchanges were attempted, information was passed between different computer nodes using technique conforming to the Message Passing Interface (MPI) standards. SHAKE[8] was used to constrain bond distances enabling a 2 fs time step and temperature was globally controlled using a Berendsen thermostat[10] with relaxation time of 0.2 ps. For the evaluation of the electrostatic and the attractive part of the Lennard-Jones

energies and forces, the PME method[15] was applied with a direct-space cutoff of 8.52 Å, an Ewald coefficient of 0.45, and a $30\times30\times30$ Å^3 reciprocal space grid.

All simulations were carried out in a box with 30.0 Å sides, having 864 waters initially. The starting Met-enkephalin configuration was obtained from an NMR ensemble (pdb 1PLW).[55] In this configuration, the end-to-end distance (nitrogen of the N-terminus to carboxylate carbon of the C-terminus) is 10.5 Å. This distance in the ensemble of 80 lowest energy structures is ~10-11 Å. The peptide was immersed in the water box and 51 overlapping waters were removed. To apply the HREM, five systems were used with the scale factors set to $\lambda_1=1$, $\lambda_2=0.925$, $\lambda_3=0.85$, $\lambda_4=0.775$, $\lambda_5=0.7$. All five systems were started from the same initial configuration and the first 3 ns are considered as the equilibration time. Exchanges were attempted every 40 steps. For an odd number 2n+1 of systems, exchange attempts are alternated among 1◦2,...,2n-1◦2n and 2◦3,...,2n◦2n+1.

Convergence tests for Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) method described in Chapter 1 was extensively used in the studies of the simulated trajectories of Met-enkephalin in this work. The PCA analysis was carried out by using ANALYZER[85], a program written for the purpose of analyzing trajectory data by a wide variety of methods.

One common issue related to using the PCA method is that the slow relaxation of the large fluctuations may prevent the fast convergence of the covariance matrix[41], which results in slow convergence of the essential space spanned by those large fluctuation modes. So, before using trajectories generated from PCA, it will be better to first test the convergence in the essential subspace. (These tests can also be used as a severe tests of

simulation convergence because the larger PCA eigenvalues correspond to the slower motions of the subject peptide or protein.) Several convergence tests have been proposed.[86,87] Amadei and co-workers[86] introduced a root mean square inner product (RMSIP) measure

$$RMSIP = \left[\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{n} \mathbf{m}_k(t) \cdot \mathbf{m}_i(t') \right]^{1/2} \qquad (2.5)$$

that evaluates the overlap of a subset of n $\mathbf{m}_k(t)$ modes obtained from different time intervals of the total trajectory. Here, we take t and t' to be two disjoint time intervals of the trajectory and use the resulting RMSIPs to monitor the stability of these modes.

3. Results and Discussion

Diagnostics

As discussed before, the HREM has the virtue of requiring only a limited number of systems while, if implemented appropriately, still providing robust sampling. In explicit solvent simulations, not scaling the solvent-solvent interactions should provide a substantial reduction in the number of systems required relative to the TREM that scales all the degrees of freedom. As in all REM versions, the choice and optimization of the acceptance probability of attempted exchanges in the HREM is a central issue. There should be an optimal acceptance probability, because for low exchange probability the rate of movement through configuration space is small, while for high exchange probability the movement through configuration space is slow. Table 2.1 lists the

acceptance ratios for the HREM simulation (the ratios are shown for three 3-nanosecond time intervals, 3 to 6 ns, 6 to 9 ns and 9 to 12 ns, the first three nanoseconds are considered as equilibrium time); they are all around 0.43. Predescu and co-workers[9] analyzed the optimization of the TREM acceptance ratio for a multi-dimensional oscillator system, and found that 0.3874 is the optimal acceptance ratio, with the efficiency falling off slowly around this value (acceptance probabilities in the 7-82% range provide rates sufficiently close to optimal sampling rates).. The uniformity of the acceptance probability values observed indicates that the choice of the number and spacing (the specific λ_i values) are appropriate.

Table 2.1. The acceptance ratio for the time span of the HREM simulation (exchanges are attempted every 80fs.)

Potential Index	λ_a	λ_b	Acceptance Ratio for 3 – 6 ns	Acceptance Ratio for 6 – 9 ns	Acceptance Ratio for 9 – 12 ns
1<--->2	1.0	0.925	0.424	0.438	0.420
2<--->3	0.925	0.85	0.451	0.409	0.463
3<--->4	0.85	0.775	0.457	0.433	0.458
4<--->5	0.775	0.7	0.431	0.408	0.423

To examine whether all the configurations (replicas) can visit a particular system (with a particular λ_i value) and whether given configurations are visited by all the systems, time trajectories for three 3-nanosecond intervals are displayed in Figures 2.1 to

2.3. From the plots, it is clear that all the configurations (replicas) can be visited by all the systems (a-e) and, conversely, all the systems can be visited by all the configurations (replicas) (f-j). In Figure 2.1 (a) the points are vertically connected to show that the plane is covered, demonstrating the desired itineration. In all the other plots, simple points are inserted to indicate occupancy. On this scale, all the plots look uniform in time and similar, which supports the desired feature that the systems undergo a random walk in the whole exchange range. Examined at higher resolution (a shorter time interval) plot 2.1 (a), for example, does show that replica 1 is slightly favored by lower-numbered systems, as actually can be inferred from the white space in the plot.

The important migration aspect has been verified, as shown in the above, to make sure that that the exchanges of systems are sufficient. But, whether these exchanges are efficient and effective, in other words, whether these exchanges induce sufficient transitions of states for the normal system (λ=1) is the next important thing that needs to be investigated. In Figure 2.4, from (a) to (c) the end-to-end (Tyr1 terminal nitrogen to Met5 terminal carboxylate carbon) distances are plotted along the time line for the 3 to 6, 6 to 9, and 9 to 12 three 3-nanosecond intervals respectively. From the figure, it is clear that, for each time interval, there are numerous transitions between open and closed states. This observation of sufficient sampling of not only both open and closed states but also transitions is a necessary condition for the following analyses, since statistical estimates are not meaningful unless they are based on enough collected sample data points. In fact, a 18 ns normal MD simulation was conducted using the same starting structure (an open form of Met-enkephalin), but the sampling was always trapped in the open form for the whole simulation period.

Figure 2.1. (a)-(e): Migration of systems into and out of a given configuration for 3-6 ns. (f)-(j): Migration of configurations (replicas) into and out of a given system for 3-6 ns (λ_i scale value). The figures from (a) to (e) correspond to systems 1 to 5 and, similarly, the figures from (f) to (j) correspond to configurations 1 to 5. In (a) the points are vertically connected, and the coverage demonstrates the desired itineration. Note that in view of the number of data points that are plotted, it appears as if, at a particular time, several replicas occupy the same window or several systems visit the same replica– this does not happen, looking at shorter intervals dots will be separate.

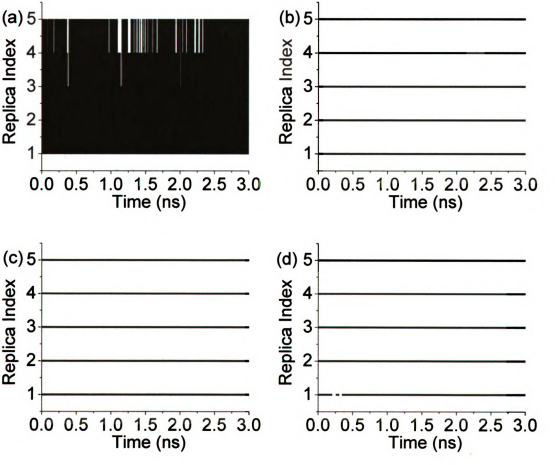


Figure 2.1 (continued)

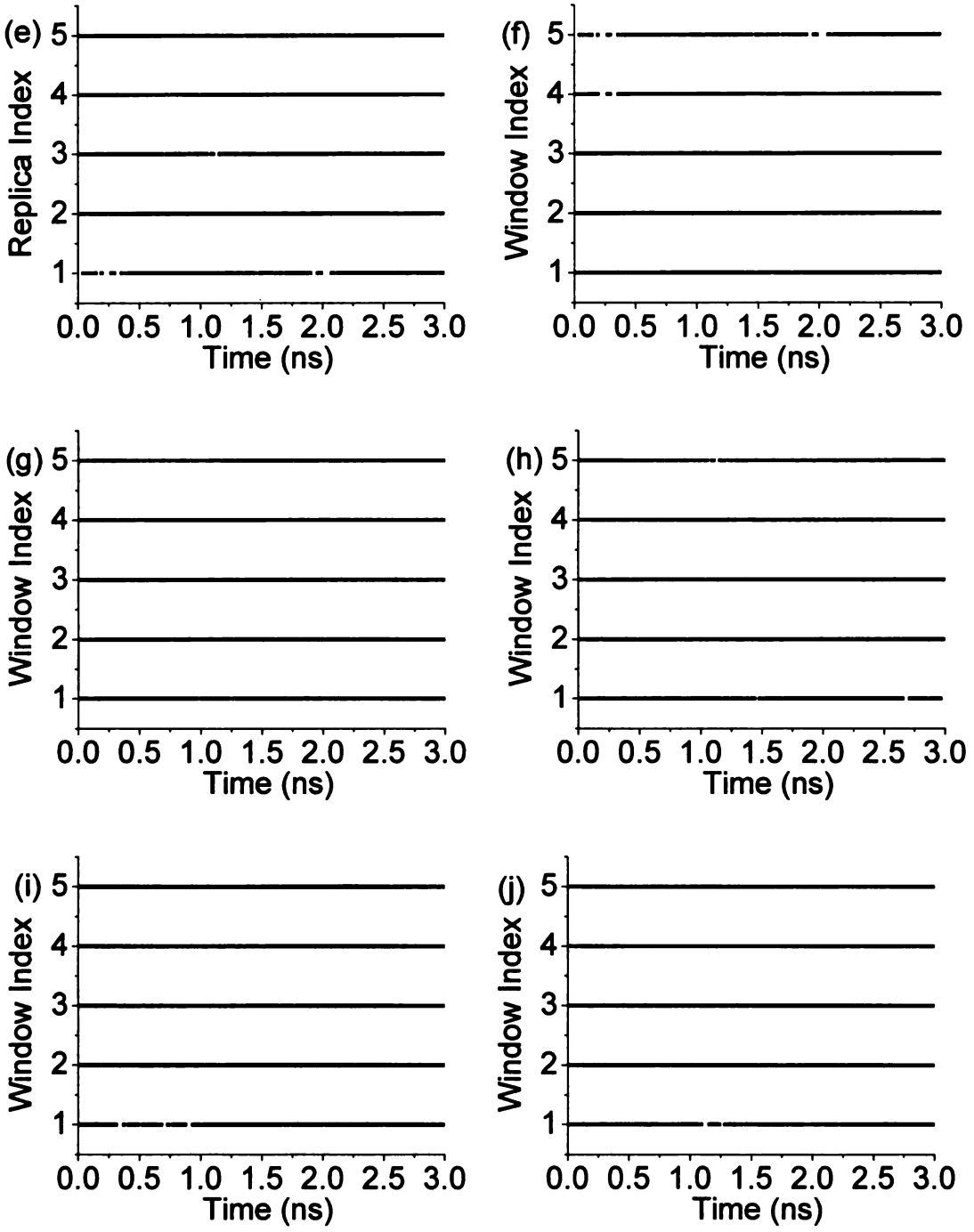


Figure 2.2. (a)-(e): Migration of systems into and out of a given configuration for 6-9 ns. (f)-(j): Migration of configurations (replicas) into and out of a given system for 6-9 ns (λ_i scale value). The figures from (a) to (e) correspond to systems 1 to 5 and, similarly, the figures from (f) to (j) correspond to configurations 1 to 5. Note that in view of the number of data points that are plotted it appears as if, at a particular time, several replicas occupy the same window or several systems visit the same replica– this does not happen, looking at shorter intervals dots will be separate.

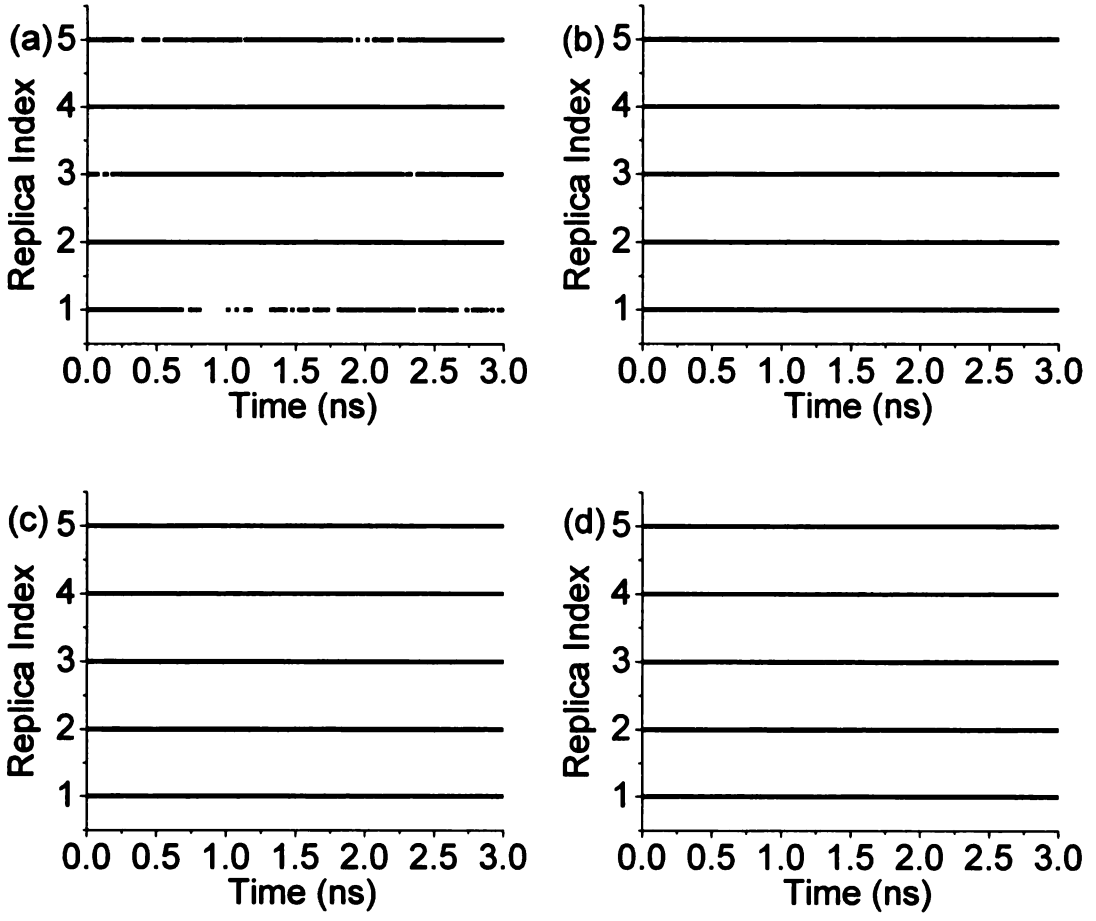


Figure 2.2 (continued)

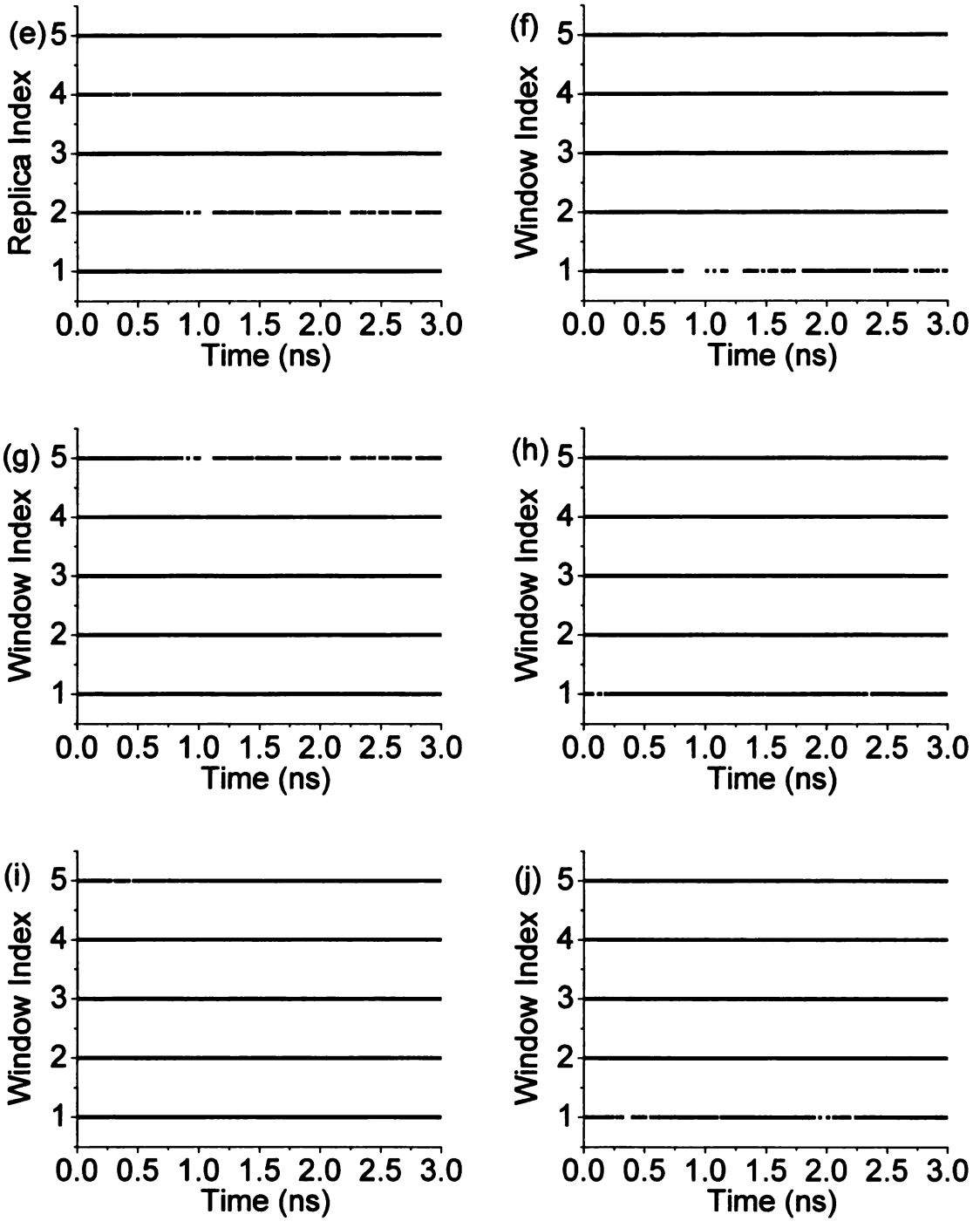


Figure 2.3. (a)-(e): Migration of systems into and out of a given configuration for 9-12 ns. (f)-(j): Migration of configurations (replicas) into and out of a given system for 9-12 ns (λ_i scale value). The figures from (a) to (e) correspond to systems 1 to 5 and, similarly, the figures from (f) to (j) correspond to configurations 1 to 5. Note that in view of the number of data points that are plotted, it appears as if, at a particular time, several replicas occupy the same window or several systems visit the same replica– this does not happen, looking at shorter intervals dots will be separate.

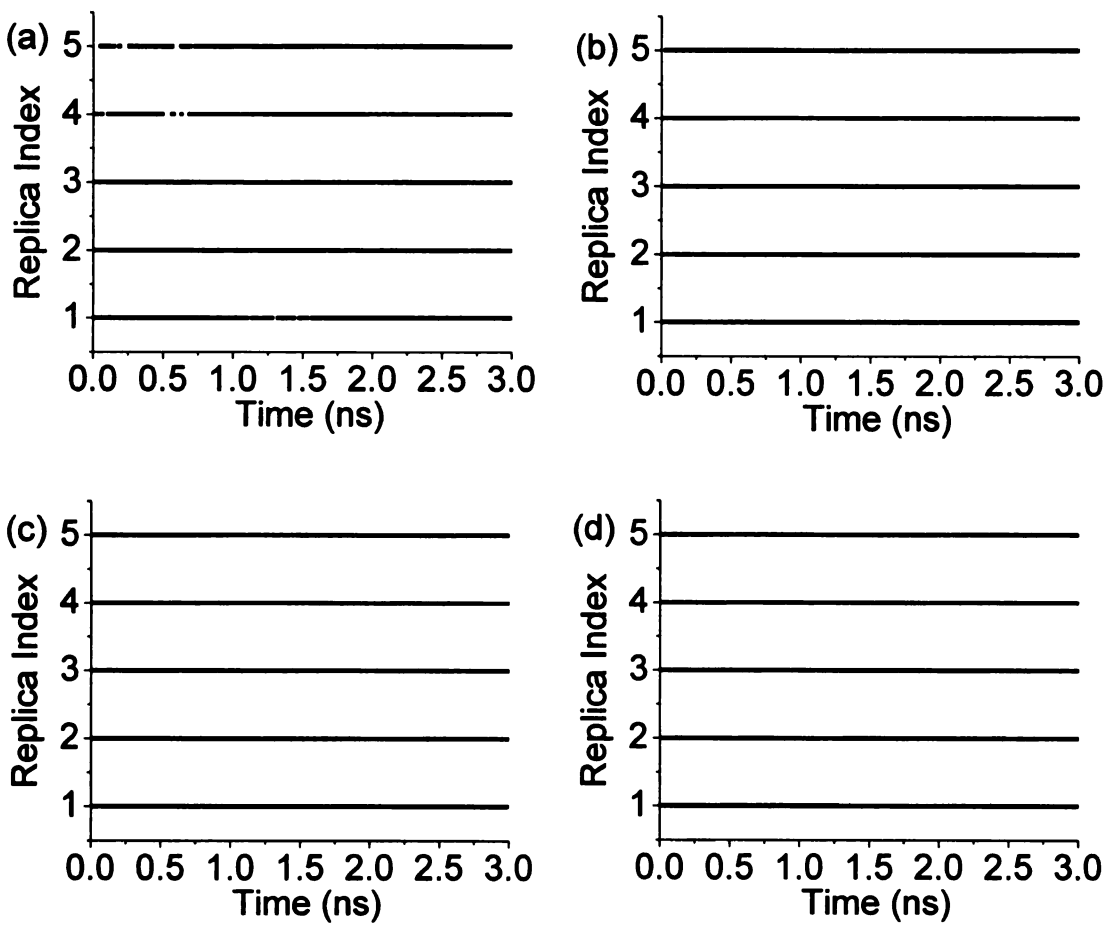


Figure 2.3 (continued)

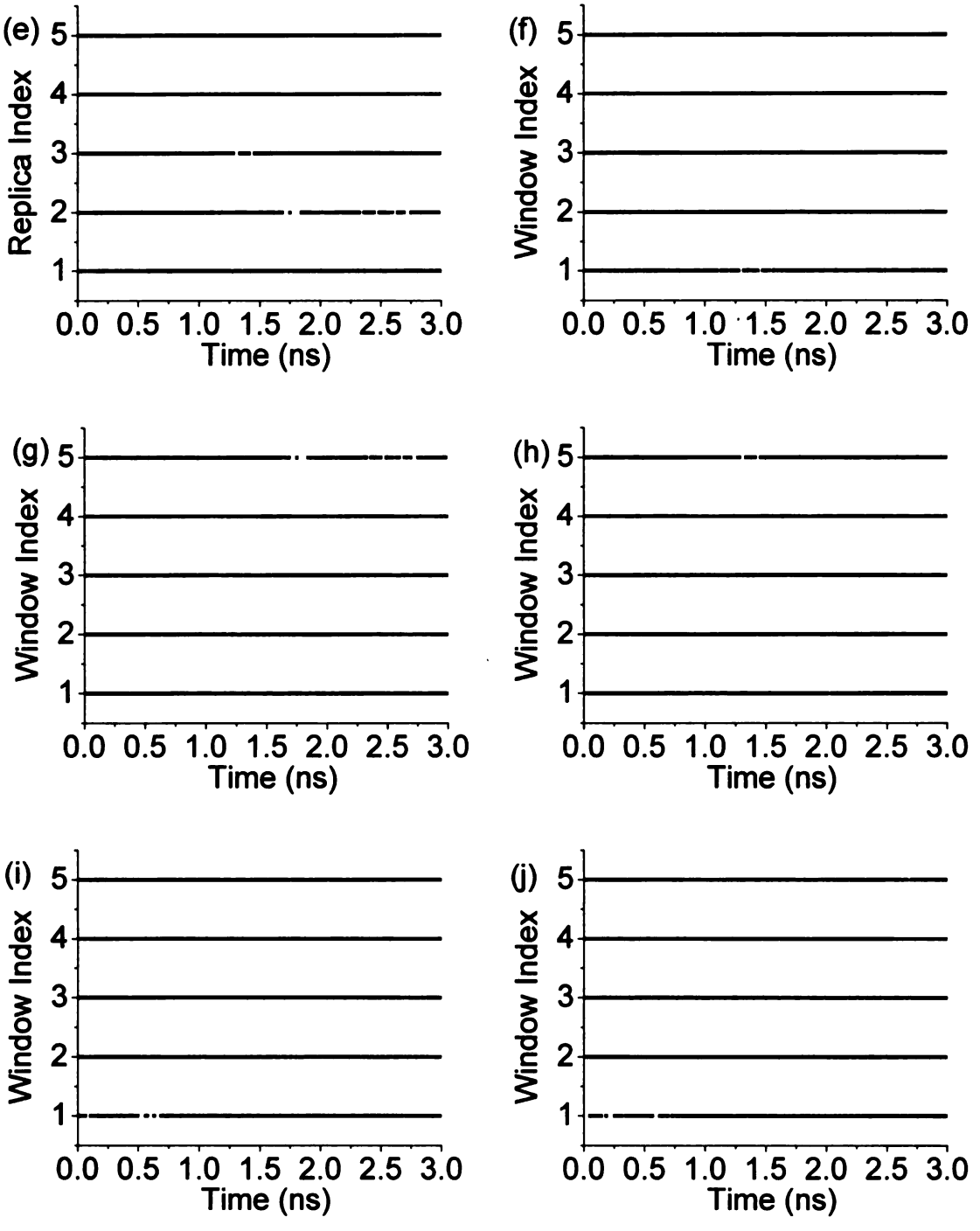
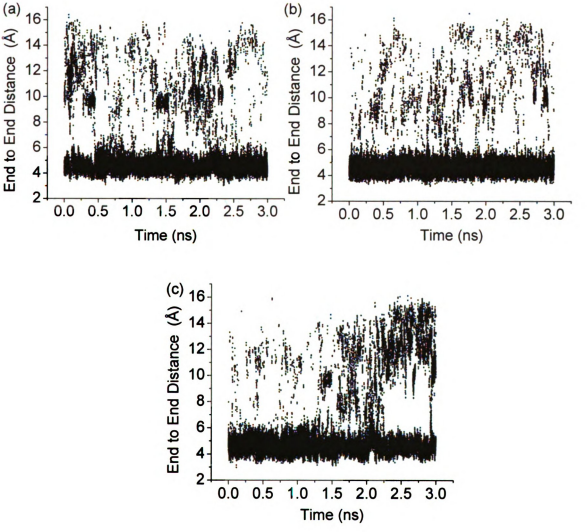


Figure 2.4. The end-to-end distances (Tyr1 terminal nitrogen to Met5 terminal carboxylate carbon) along the time line for: (a) 3 to 6 ns, (b) 6 to 9 ns, and (c) 9 to 12 ns for the λ_1=1 system.



PCA trajectory diagnostics

As discussed in Chapter 1, Principal Component Analysis (PCA) is a highly useful and popular tool for the systematic investigation of the conformational sampling of a peptide[44,88] since it can reduce the high dimensional configuration space to an essential subspace that contains most of the significant, large-scale motions. Only window 1 in our HREM simulation (λ_1=1) corresponds to the normal force field, and all of the following analyses are based on window 1 for this reason. The first 3 nanoseconds of simulation are considered as an equilibration period, and the PCA trajectories are constructed based on the backbone atoms. Over the trajectories, the first mode represents around 40% of the total backbone variance (the total MSF), the first two modes together around 55%, and the first ten modes together around 90%, showing that the HREM simulation of met-Enkephalin does provide a division into essential and remaining subspaces. Because of the possible convergence problem mentioned earlier in this chapter, before diving into any detailed examination of the trajectories using PCA, we first test the convergence in the essential subspace with the RMSIP[86] method (eq 2.5), examining convergence by comparing the overlap of modes constructed from different time intervals. We compare RMSIPs among three time intervals, 3-6 ns, 6-9 ns, and 9-12 ns, for the system 1 trajectories (λ_1=1) for the first mode, the sum of the first two modes (these two modes are the only modes which will be used for the following analysis), and the sum of the first ten modes. The RMSIP results, which are a direct measure of the projection of the basis of one subspace (for one time interval) onto the other subspace (another time interval), are listed in Table 2. All the RMSIP values are close to their limiting values of 1.0. The mode one result indicates that the three PCA first mode vectors constructed from the

three time intervals are essentially the same. That result suggests that there might be a reaction coordinate correlated with the PCA first (slowest) mode, and that 3-nanosecond intervals are long enough to capture this slowest motion. The RMSIP values for the first two and ten modes show a similarly good convergence. Thus, we may be confident that the HREM simulation leads to stable results over the 9-nanosecond interval (3-12 ns) and, that, 3-nanosecond intervals are sufficiently long to capture the fluctuations of Met-enkephalin.

Table 2.2. The RMSIPs (for HREM window 1) for the three 3 ns intervals (3-6, 6-9 and 9-12 ns) using the PCA first mode, first two modes and first ten modes

	3-6 vs. 6-9 ns	3-6 vs. 9-12 ns	6-9 vs. 9-12ns
First Mode	0.972	0.991	0.980
First Two Modes	0.918	0.959	0.909
First Ten Modes	0.926	0.988	0.914

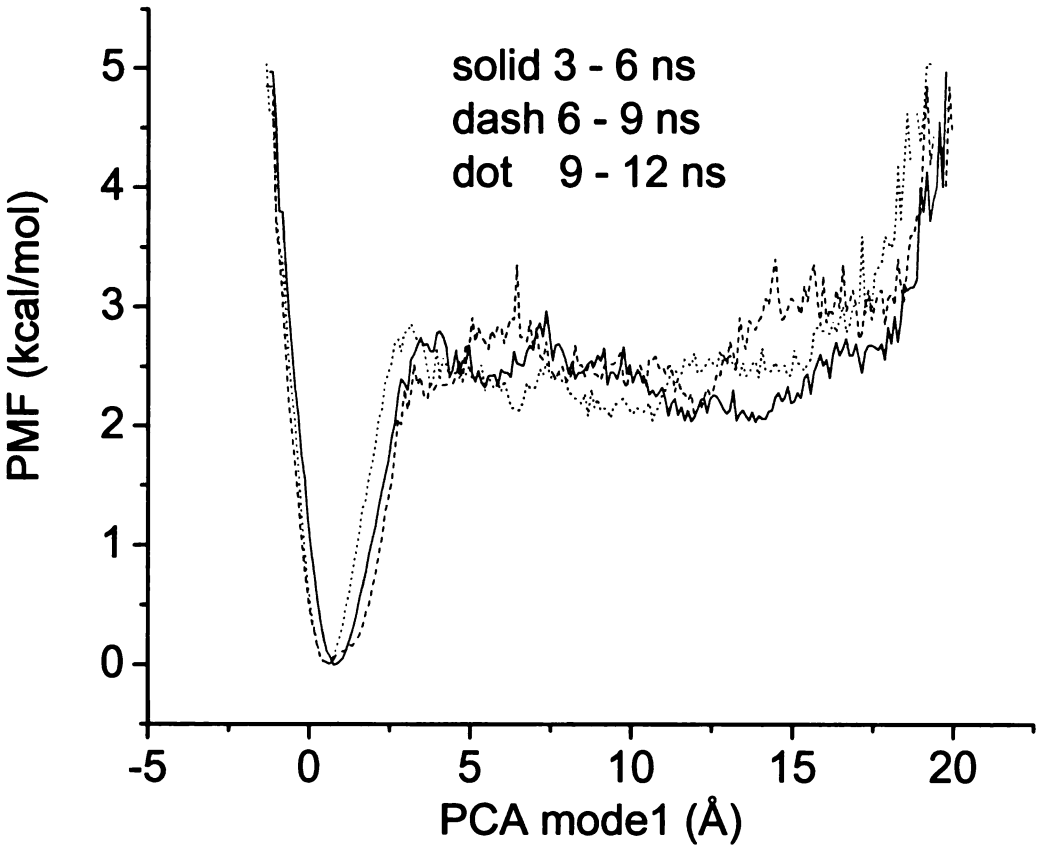
1-dimensional PCA analysis

The observation that the first PCA mode captures about 40% of the total MSF combined with the observation of high convergence of this mode shown by the RMSIP test suggests its potential appropriateness to be used as a reaction coordinate. And, it should be more objective in character than that based on, for example, an end-to-end distance which could otherwise be a natural choice. The first mode displacement

trajectory, $p_1(t)$, defined in eq 1.19 exhibits numerous transitions between its extreme values, which shows that a broad range of distances is being sampled repeatedly during the simulation time. Thus, it is legitimate to construct a potential of mean force (PMF). The PMF, constructed by making a histogram of $p_1(t)$ for the three time intervals (for each 3ns interval, 15000 snapshots are used) of 3 to 6 ns, 6 to 9 ns, and 9 to 12 ns are displayed in Figure 2.5. The convergence among the three PMF profiles is as good as anticpated from the RMSIP test results. Along each PMF profile, there is a well defined well, around 1 Å, and a broad plateau region from 5 to 18 Å. The difference between the lowest points of the two wells is about 2.0 kcal/mol, and the barrier between them is around 3 kcal/mol. The barrier is not high, suggesting that switching between these two states should not be difficult. Although the deep well is about 2 kcal/mol lower then the broad well, it is much narrower, suggesting that there could be an energy/entropy compensation trade-off between the broad and deep wells. These features support the observations of the co-existence of different conformations of Met-enkephalin in water and of a lack of distinguishable secondary structure.

The type of atom displacements that correspond to the first PCA mode can be inferred by calculating the correlation coefficient of the end-to-end distance for the first mode (obtained from eq 2.6 with i=1) and for the true trajectory. The correlation coefficients calculated are 0.968, 0.955 and 0.980 for the three 3-nanosecond time intervals respectively, suggesting that mode one is essentially reporting on the end-to-end distance fluctuations. The PCA has thus succeeded in singling out the principal collective motion that spans the open to closed conformations.

Figure 2.5. The PMF corresponding to the PCA mode 1 displacements for all λ=1 system snapshots for the three time intervals of 3 to 6 ns, 6 to 9 ns, and 9 to 12 ns.

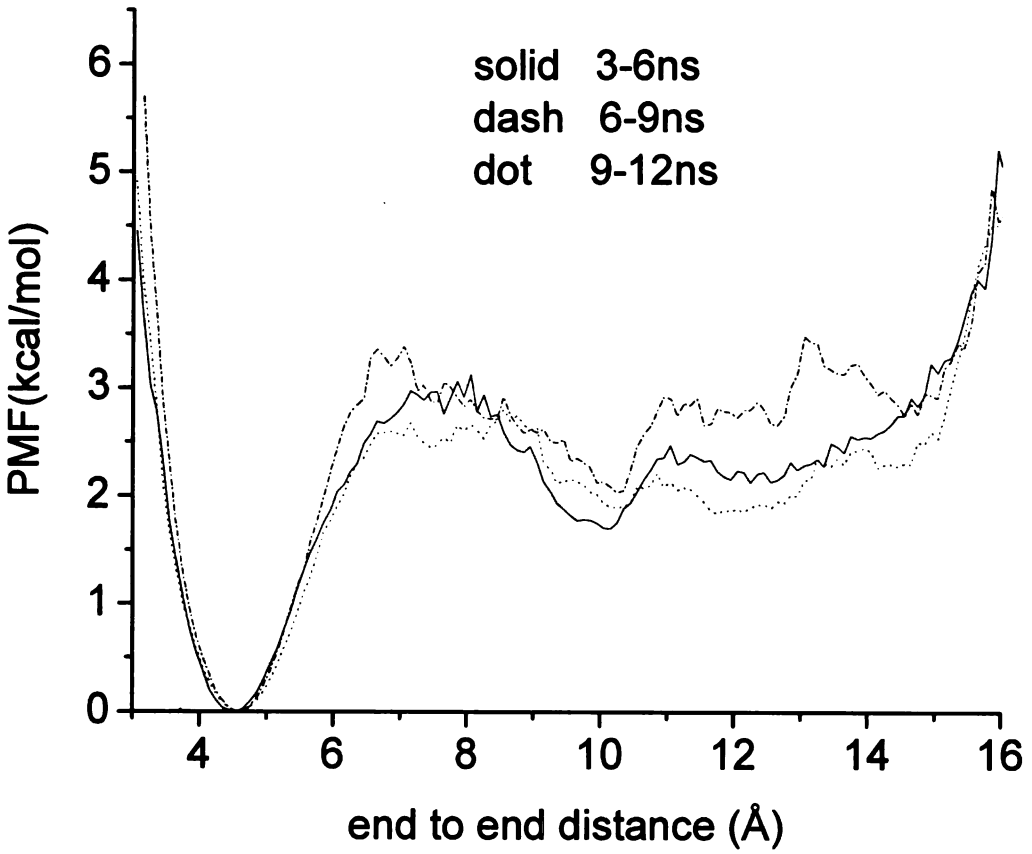


End-to-end (Tyr1 terminal nitrogen to Met5 terminal carboxylate carbon) distance PMF and its comparison with DREM results

As noted above, there exists very high correlation between the end-to-end (Tyr1 terminal nitrogen to Met5 terminal carboxylate carbon) distance fluctuations for the complete trajectory and that found from the first PCA mode, making the end-to-end distance seemingly to be an interesting reaction coordinate. In addition, the consideration that the zwitterionic form of Met-enkephalin should be capable of forming a salt bridge between the terminal Tyr1 amide and Met5 carboxyl groups also makes the end-to-end distance to be a natural choice for a reaction coordinate. Therefore, the 1-dimensional PMF along this end-to-end distance coordinate was calculated from the trajectory of the HREM normal system (system 1) for the three time intervals, 3-6 ns, 6-9 ns, and 9-12 ns respectively and the results were displayed in Figure 2.6. The agreement among the three PMF profiles is farily good with the exception of the region expanded from 9 to 14 Å in which there is a discrepancy about 1 kcal/mol. The poorer convergence in this region is most likely due to the extensive configuration space for these less constrained regions (this point is discussed more in detail in the next part for 2-dimentional PCA results).

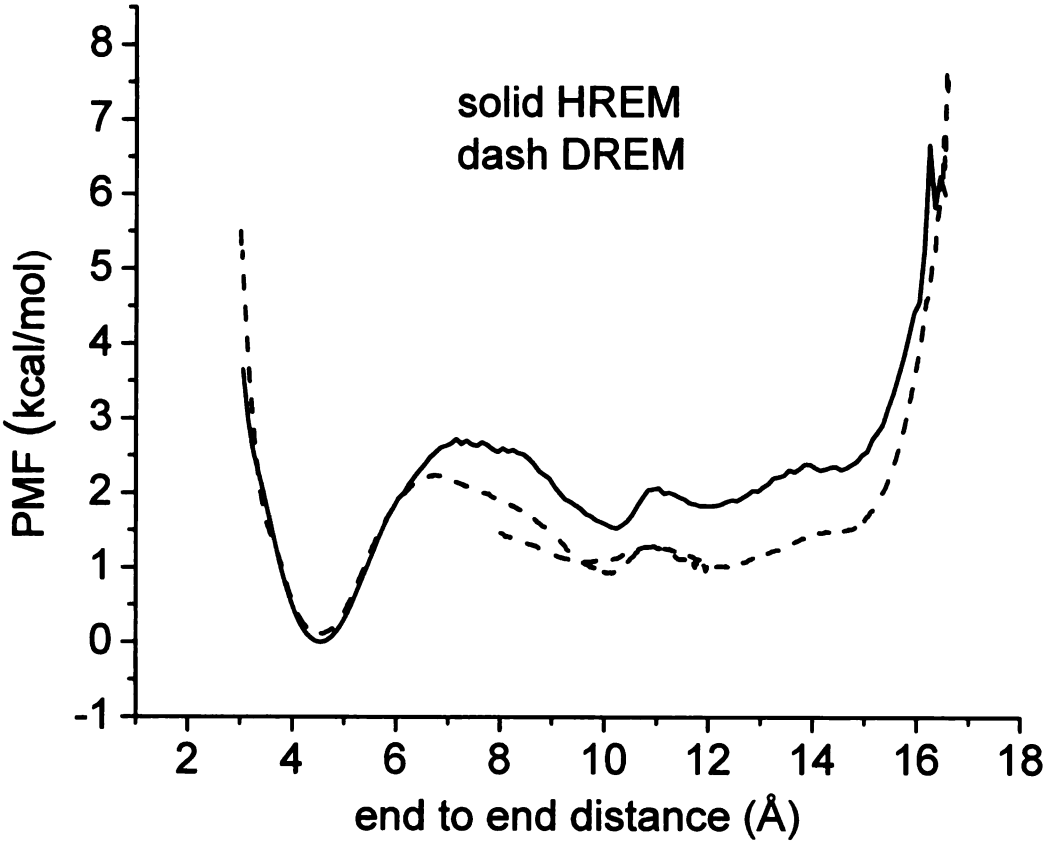
When the end-to-end distance is chosen as the reaction coordinate, an umbrella sampling based technique could be implemented to focus on such a natural reaction coordinate. Usually, an umbrella sampling based technique, if possible to be implemented, should be able to provide a better PMF estimation along a particularly chosen reaction coordinate than would a HREM simulation, due to the former's focus on a specified reaction coordinate. Of course, for many problems, it is hard to determine a proper

Figure 2.6. The PMF for the end-to-end distance generated by the HREM simulation for all systems with λ=1 for the three time intervals of 3 to 6 ns, 6 to 9 ns, and 9 to 12 ns.



reaction coordinate beforehand, and some chosen reaction coordinates are hard to implement. For example, as for this Met-enkaphalin study, the end-to-end distance was chosen mostly because it was observed to have extremely high correlation with the PCA first mode and it would be very hard, if at all possible, to use a PCA mode as reaction coordinate for implementing an umbrella sampling like techniques. The umbrella sampling based technique chosen here is the Distance Replica Exchange Method (DREM)[89] , which can also be considered as a special type of HREM method. The Hamiltonian for the i^{th} system of the DREM approach has the form: $H_i(X) = H_0(X) + W_i(r)$, where, $W_i(r) = \frac{1}{2} k_i (r - r_0^i)^2$ is the standard harmonic window potential, r is the chosen reaction coordinate distance and, r_0^i is the equilibrium distance for the system's harmonic constraints used along with the force constant k_i. The main difference between DREM and a regular umbrella sampling method is that, for DREM, like all other replica exchange methods, after certain steps (100 steps for this work), exchange attempts will be made between neighboring systems for their window potentials. It was shown[89] that DREM trajectories have much better convergence than for trajectories generated with normal umbrella sampling. Of course, since DREM biases the true Hamiltonian, as does the regular umbrella sampling method, some un-biasing procedure is needed to correct the corresponding probability distribution. The Weighted Histogram Analysis Method (WHAM)[37,39] was used to do the un-biasing and to combine the trajectory data obtained from the different systems. Comparing the DREM PMF with the corresponding HREM PMF (constructed by combining 3ns time interval trajectories to form a long time trajectory from 3 to 21ns to get a smoother estimation in the broad 8

Figure 2.7. The PMF for the end-to-end distance generated by the HREM simulation and DREM simulations. The PMF line for DREM is obtained by combining two separate lines for the 3-11 Å and the 8-16 Å simulations.



to 16 Å region) in Figure 2.7, it is clear that the patterns are very similar, but the DREM PMF simulation is shifted down around 1 kcal/mol in the region 7 to 15 Å, which covers the broad well and the barrier between the two wells. The total time over all processors in the presented simulation, excluding the equilibration times, was 136 ns for the DREM and 90 ns for the HREM simulations. Although the 136 ns vs 90 ns difference seems not to be too substantial, note that they do not represent the real time scale of the simulated trajectories. For REM schemes, because dynamics are accelerated, it is hard to estimate what the real time scale is for a given simulated trajectory. Therefore, the actual motions sampled by the DREM scheme and HREM scheme could be a little bit different, resulting in the corresponding pmf profiles to have some differences. Futhermore, although the DREM simulation was designed to focus on the particular end-to-end distance reaction coordinate, the quality of the resulting pmf depends on how well the other dimensions orthogonal to the reaction coordinate are sampled. Of course, the quality of the pmf calculated from the HREM results also depends on the sample quality of the HREM simulation. Therefore, it is hard to conclude which method provides a more accurate estimation of the pmf profile along end-to-end distance for Met-enkephalin. But, because the difference between the two profiles are not large, and their patterns are very similar, it seems to suggest both schemes are appropriate to producing a pmf. Taking into consideration that HREM does not require any explicit guiding potential for a reaction coordinate, the HREM simulation does a remarkably good job.

2-dimentional PCA Analysis

Besides the first PCA mode, the second PCA mode also has a significant contribution (15%) to the overall motion, motivating construction of a 2-dimensional

Figure 2.8. The 2D PMF for the first and second PCA modes for all λ=1 system snapshots. Backbone CA stick plots of the configurations in the dense places are shown with the distance between Tyr1 backbone nitrogen and Met5 carboxyl carbon shown.

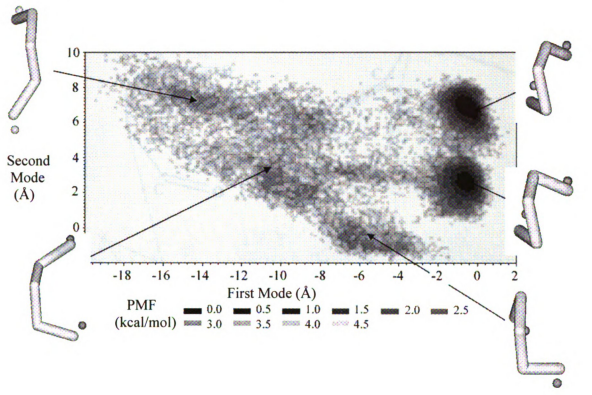


Figure 2.9. CA wire plots for the two salt bridge conformers shown in Figure 2.8, with the Gly-2 and Gly-3 backbone atoms shown explicitly that illustrates the Ψ(2) and Φ(3) dihedral angle compensation mechanism.

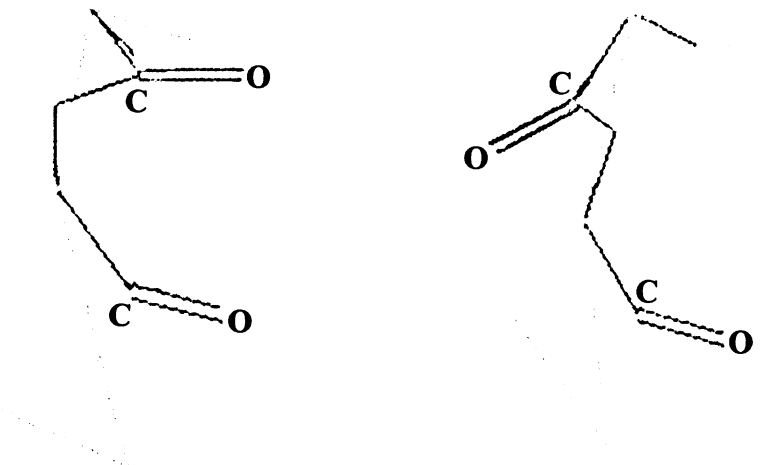
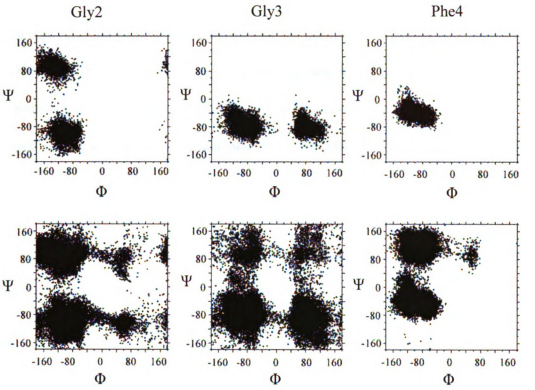


Figure 2.10. Ramachandran plots for Gly2, Gly3 and Phe4 of Met-enkephalin; The three plots in the upper row show results for snapshots that are in the PCA first mode deep well (–0.6 to –0.4 Å); The lower row for all snapshots.



PMF displayed in Figure 2.8 (Since we have already shown good PCA convergence in the above discussions, for the 2 dimensional analyses only the 3-6 ns interval trajectories was focused on for convenience). There are two wells around (–0.5 Å, 2.3 Å) and (–0.5 Å, 7.0 Å) in the figure that correspond to the deep well in the one dimensional PMF plot. Representative alpha Carbon (CA) pictures of the dominant backbone conformations are also shown in the figure. Configurations in the two wells have end-to-end distances corresponding to salt-bridged conformers, versus the more extended states of all the other displayed configurations. Figure 2.9 shows representative CA wire frame backbone structures with the Gly-2 and Gly-3 atoms explicit for the two wells. The differences in the backbone structures, with the parallel and anti-parallel carbonyls, come from the differences of the Gly-2 Ψ and Gly-3 Φ dihedral angles. The first row plots of Figure 2.10 shows the Ramachandran plots for Gly2, Gly3 and Phe4, with snapshots picked from the PCA first mode within the range from –0.6 to –0.4 Å, and the bottom row plots for all snapshots are shown for comparison purposes. Although the configurations corresponding to the two deep wells in the 2-dimensional PMF have large differences in their Gly-2 Ψ and Gly-3 Φ dihedral angles, the main patterns of their backbones are similar, illustrating the known[90] feature that the $\Psi(i)$ and $\Phi(i+1)$ values of residues i and $i+1$ can compensate and still lead to overall similar structures. This is the only local mechanism in peptides (or proteins) that can lead to a structure with essentially the same overall conformation. The Ramachandran plots in the first row of Figure 2.10 are quite similar to those generated by van der Spoel and Berendsen[63] in their study of zwitterionic Leu-enkephalin solvated by water. Their simulations were started from a salt-bridge and an open form that rapidly closes and remains mostly closed, so they mainly sample salt

bridge conformers. They also note a Gly2 Ψ and Gly3 Φ compensation for these salt-bridged conformers.

The finding that the Gly2 and Gly3 are concentrated in two distinct regions of Ψ and Φ suggests that these residues may be participating in hydrogen bonding. For the "parallel" carbonyl arrangement of Figure 2.9 there can be Gly3 N-H to Met5 O=C hydrogen bonding/salt-bridge interactions. For the "anti-parallel" arrangement there can be Gly2 C=0 Met5 N-H as well as Gly2 N-H Met5 O=C hydrogen bonding/salt-bridge interactions. Selected snapshots of the parallel and anti-parallel conformers do show these interaction patterns that most likely aid in stabilizing the direct terminal salt-bridge interaction. However, hydrogen bond analysis of the trajectories shows that none of these additional interactions persists. The relevant distances are constantly fluctuating and are mainly beyond the distance that would permit such a conclusion.

In the remaining region of the 2-dimensional PMF plot in Figure 2.8, there is a less concentrated but much larger area of relatively low free energy, than in the area around the two deep wells. The barrier between the deep and broad areas is only around 3 kcal/mol, and the difference between them is roughly within 2 to 3.5 kcal/mol, both of which are not very large on a thermal scale but, significantly, the pathway between two areas is very narrow. This observation suggests that although Met-enkephalin samples both regions, one favored by energy and the other favored by entropy, the rate of transition between them may be slower than might be inferred from the small free energy barrier. Starting from an extended configuration, there are many configurations that correspond to the extended states area to first sample, and then a restricted region in configuration space must be found that corresponds to the small transition area in Figure

2.8, to enter the salt-bridge region. This feature may explain the discrepancy between the experimental results[53-56,58,60] that Met-enkephalin or Leu-enkephalin shows great flexibility and a lack of definite conformations in water, and the difficulty of going from the extended to closed conformations found through conventional MD simulattions[19,32,61-69]. Indeed, for the zwitterionic Met-enkephalin simulated here, an 18 ns normal MD simulation with the same starting configuration and conditions as for the HREM system 1 (λ=1) is trapped in extended states and never samples configurations corresponding to the two deep wells.

4. Concluding Remarks

The HREM approach was successfully implemented to improving the MD sampling of zwitterionic Met-enkephalin in explicit solvent. The implementation of HREM was quite efficient in the sense that only five windows were required to generate results that, as monitored by the PCA modes, show good convergence properties in the simulation time. More importantly, it was shown that the system 1 (normal system with λ=1) trajectory repeatedly sampled configurations that correspond to all the relevant end-to-end distances between the open and close states, providing an essential condition for properly exploring the configuration space of Met-enkephalin and a meaningful calculation of free energy quantities, such as a potential of mean force. The PCA was effective in singling out the dominant end-to-end distance fluctuation motion in the first mode.

The 1-dimensional PMF profile along the end-to-end distance for the Met-enkephalin calculated from using the HREM approach was also compared to the one

calculated from using a DREM approach, which was implemented to use window potentials to focus on the particular end-to-end distance reaction coordinate. The two PMF profiles are quite similar, especially in pattern. The DREM result might be more accurate since it is carefully designed to focus on the reaction coordinate. But, the HREM has the advantage that it does not require a commitment to a particular reaction coordinate, or a particular dimensionality of reaction coordinate. The objectivity of the HREM in this regard is an argument in its favor relative to the DREM since in many cases it is hard to specify or know beforehand a proper reaction coordinate. Of course, the TREM is also objective because all degrees of freedom are thermally excited. However, for a system with many degrees of freedom, the chain of replicas required may become impractically large. The HREM can be viewed as an attempt to excite only important degrees of freedom, but that requires a decision as to which are the important degrees of freedom. For explicit solvent simulations, not having to excite the solvent degrees of freedom is a great advantage since their number is an order of magnitude larger than the number of peptide (and protein) degrees of freedom in typical MD simulations.

The simulated trajectory, analyzed with the PCA decomposition, also revealed the presence of another significant PMF dimension that is mainly composed of a correlated local change of the Gly2 Ψ and Gly3 Φ dihedral angles. That two very different regions of dihedral space for these residues are stably occupied, even though the overall structure, as monitored by the end-to-end distance, corresponds to a salt-bridge conformer, results from this local $\Psi(i)$ and $\Phi(i+1)$ compensation. Though Met-enkephalin is small and the barriers between the open and close states is not large, the configuration space of important states of the peptide seems to be very broad and there seems to be only a

narrow bridge between open and closed states. This observation probably suggests that properly sampling the Met-enkephalin configuration space is more an issue of having to explore a large space with small potential energy barriers than of having to overcome a large barrier along a reaction coordinate for some transformation as in a folded, stable protein. The result is a free energy profile that does not exhibit a large barrier, yet requires a great deal of simulation time or implementing special MD methods such as the HREM used here.

Chapter 3 A specific Hamiltonian Replica Exchange Method for the study of EcHPPK and YpHPPK

1. Introduction

As described in Chapter 1, HPPK (6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase), a key enzyme in the folate biosynthetic pathway, which catalyzes the transfer of pyrophosphate from ATP to 6-hydroxymethyl- 7,8-dihydropterin (HP)[72], is a protein of great interest, since it could possibly be targeted to design new antibiotics. In this work, an HREM implementation (details are in the following "methodology" section), similar to the one which helped enhance the sampling of Met-enkephalin, as described in Chapter 2, was used to study two homologs of HPPK, *E. coli* HPPK[75] (EcHPPK) and *Y. pestis* HPPK[80] (YpHPPK). EcHPPK is the most studied HPPK, and these studies have shown that it has several well-defined, biologically important conformational states along its catalytic cycle.[77,78,74] YpHPPK is a causative agent of septicemic, pneumonic and bubonic plague, and is one of the most virulent pathogens known.[91] The main reason to study two homologs is, hopefully, to find appropriate targets for which species-selective inhibitors could be successfully designed. If this goal were accomplished, certain new species-selective inhibitors might be designed to be used as new narrow band antibiotics effective against targeted microorganisms, helping to counteract the dangerously wide spread antibiotic resistance.

One important basis for this work is related to current suggestions that conformational change occurs, to a greater or lesser extent, when proteins and ligands form complexes. A progression of models has been introduced that in essence increase the role of protein plasticity in the binding of ligands.[92-95] The "lock and key" model

asserts that a protein has a cavity that a ligand (or another protein) can fit into with minor rearrangements of protein and ligand. However, this model cannot account for proteins that can bind differently shaped substrates. In contrast, the "induced fit" model tries to account for this by suggesting that a ligand induces a conformational change at the binding site, shifting it toward an active state. The pre-existing equilibrium hypothesis[96], based on more recent funnel energy landscape protein-folding theories[97,98], asserts that the native state of a protein can exhibit an ensemble of conformations that can span apo to more ligand-bound-like conformations. The ligand can select a pre-binding conformation from the ensemble of protein conformers, and then bias the equilibrium toward the catalytically competent binding conformation. Some studies have shown that a number of ligand-free proteins (staphylococcal nuclease[99,100], calbindin[101], adenylate kinase[83,102], and calmodulin[103]) do sample conformations required for ligand binding. Conformational plasticity might be more advantageous for ligand binding by requiring smaller protein rearrangements (with a reduced energetic requirement) than would be necessary in the more open forms.

Adopting the reasonable assumption that proteins can fluctuate to produce conformations suited for ligand binding, in this work the focus is on *near-closed* (close to what is referred to as the closed structure that characterizes the ligand-protein complex, as determined by crystallography) conformations that are relatively stable (i.e. significantly populated) for the reason presented in the rest of this paragraph. Active-site residues are often well preserved in a given enzyme among different species due to the fact that the fundamental catalytic mechanism and substrates are typically the same. Consequently, the closed conformation usually represents a challenging template for

developing *species-selective* inhibitors. Compared with closed conformations, near-closed conformations may provide greater opportunities for two homologous proteins to discriminate between different ligands. The identification of new, low-energy protein conformations that have not been trapped by crystallographic or NMR studies may provide targets for species selective inhibitors.

To carry out the identification of near closed conformations, there are two tasks that need to be completed. First, near closed conformers need to be generated; for this purpose, the HREM is used. Second, near closed conformers need to be distinguished and then categorized. For this purpose, a clustering algorithm is used along with a qualitative method to define binding cavities for the substrate. Implementing HREM in the way described in the next methodology section for a protein requires information about which parts are most rigid (a core region) and which parts are more flexible. By comparing the closed (the ternary complex with substrates HP and ATP)[77,78] and open (apo)[74] forms of EcHPPK it is clear that the regions of the protein that have the largest differences between the two forms are Loops 2 and 3. Furthermore, for both EcHPPK[33] and YpHPPK[75,80], these loops are among the most flexible regions based on xray B-factors. Therefore, the HREM method implemented in this work focused on these two loops. Compared with using normal MD, we found that HREM helps to sample the low-energy states neighboring the closed conformations of the crystal structures for both EcHPPK and YpHPPK so that broader simulated ensembles of dominant conformations were obtained. Based on these simulated ensembles, clustering methods were applied to identify well-populated conformations that differ significantly in HP binding pocket conformations between the two species. Differences among some near closed

conformations in the EcHPPK and YpHPPK HP binding pockets can be identified that are potential targets for designing ligands with enhanced specificity.

2. Methodology

Hamiltonian Replica Exchange Method

The HREM implementation used in the work summarized in this chapter is similar to the one used for Met-enkaphalin described in detail in last chapter. The only difference is that for Met-enkephalin the whole small peptide is scaled, while, in this chapter, only the most flexible regions of the protein are scaled with the aim of further reducing the number of systems required. In more detail, for the HREM implementation used for the HPPK studies, the potential energy is parameterized as

$$V_i(\mathbf{X}) = \lambda_i^2 V_{LL}(\mathbf{x}_L, \mathbf{x}_L) + \lambda_i V_{LR}(\mathbf{x}_L, \mathbf{x}_R) + V_{RR}(\mathbf{x}_R, \mathbf{x}_R), \qquad (3.1)$$

where the "L" subscript in eq 3.1 stands for "Loop" and denotes the HPPK Loop 2 (residues 44-53 for both EcHPPK and YpHPPK) and Loop 3 (EcHPPK residues 82-93, YpHPPK residues 87-94). The R subscript stands for "Rest" which denotes the rest of the atoms in the system, including the remaining parts of HPPK: ATP, the two magnesium ions, and the explicit solvent. Thus, the three terms in eq 3.1 represent interactions between atoms of the two loops, between a loop atom and an atom within the "Rest" parts, and between atoms within the "Rest" parts. λ_i is a scaling factor for the Lennard-Jones and electrostatic nonbonded interactions. It is clear by the construction used in eq 3.1 that the number of degrees of freedom is dominated by the Rest parts; thus, the

indicated scaling is much reduced relative to TREM for which the global scaling $\beta V_i(\mathbf{X}) = \beta\lambda_i V(\mathbf{X})$ would be used.

Again as described in last chapter, the normal potential energy function is being used in system 1 (λ_1=1), so that the trajectory associated with it samples the normal canonical ensemble. Therefore, all our analysis in Section 3 is based on trajectories for system 1.

Molecular Dynamics Simulations

Similar to Chapter 2, the modified CUKMODY program for HREM is used to carry out the simulations. Again, the systems were run separately on different nodes of a Linux cluster computer and, when exchanges were attempted, information was passed using MPICH, an implementation of the Message Passing Interface technique.

For all simulations, bond distances were constrained using SHAKE[8], enabling a 2 fs time step, and temperature was globally controlled at 298 K using Berendsen thermostat[10] with a relaxation time of 0.2 ps. For the evaluation of the electrostatic and the attractive part of the Lennard-Jones energies and forces, the PME method was applied with a direct-space cutoff of 9.0 Å, an Ewald coefficient of 0.45, and a $72\times72\times72$ Å^{-3} reciprocal space grid. All simulations were carried out in a cubic box with sides of 64.1 Å. For the EcHPPK (YpHPPK) simulations, 7671 (7644) waters were included, after waters overlapping the protein were removed. Five sodium cations were added to neutralize the systems for EcHPPK.

The MD starting structures were obtained from the crystal ternary complex structures for EcHPPK (PDB entry 1q0n) and YpHPPK (PDB entry 2qx0) with AMPCPP (an ATP analog that prevents turnover), two associated magnesium cations, and HP (6-

hydroxymethyl-7,8-dihydropterin). For our simulations, HP was removed, and AMPCPP replaced by ATP, with the ATP modeled using the GROMOS96[6] force-field parameter values. The two Mg^{2+} cations were each covalently linked to an oxygen phosphate atom of ATP with one linked to an alpha phosphate oxygen and the other to a gamma phosphate oxygen, in accord with their placement in the crystal structures. The ATP phosphates were assumed fully deprotonated (ATP thus has a total charge of –4) in agreement with the ATP protonation state when ligated to Mg^{2+}.[104] The protonation states of all the ionizable residues were set to their normal ionization states at pH 7.

One thing noticeable is that the available crystal structure of YpHPPK is a dimer and the issue of whether dimerization is catalytically relevant is not clear according to experiments.[80] From the crystal dimer structure, it is evident that each monomer's Loops 2 and 3 (especially Loop 2, which is longer) extends close to the other monomer, and seems to strongly interact with the other monomer. Therefore, in the dimer, these active site loops might spend a substantial amount of time bumping into each other, suggesting that the dimer is not the physiologically relevant state. For these reasons, the monomer of YpHPPK is simulated.

In the HREM simulation of both EcHPPK and YpHPPK, six systems were used with the scale factors set to λ_1=1, λ_2=0.83, λ_3=0.73, λ_4=0.65, λ_5=0.57, λ_6=0.5. They were started from their respective initial configurations, and the first 2 nanoseconds are considered as equilibration time. Exchanges were attempted every 40 steps. For the conventional MD simulations, the same initial configurations as the HREM were used and the first 4 nanoseconds are considered as equilibration periods.

The HREM scaling used does include interactions between loops 2 and 3 and the rest of the protein (and the solvent). Thus, there should be and there are some modest enhancements of the rest of the protein RMSFs relative to the MD results. The interaction of loops 2 and 3 with the rest of the protein is somewhat stronger in EcHPPK than in YpHPPK as monitored by the RMSFs. The weaker interaction for YpHPPK might be due to the shorter length of its loop 3. But, even for EcHPPK, using the HREM approach, the RMSFs for core residues are mostly below 2 Å. An interesting feature, especially evident for EcHPPK, is that though loop 1 is not scaled its RMSF has increased substantially. This observation supports the crystal structure evidence that there may be strong interactions between loop 1 and some parts of loops 2 and 3. As desired, by the conservative HREM scaling in eq 2.3 the core protein structure is very well maintained.

Deviation and fluctuation Measures

The standard structural deviation measure used is the $RMSD_j$ (root mean square deviation) defined as follows:

$$RMSD_j = \sqrt{\left(\int_0^T dt \left(\mathbf{r}_j(t) - \mathbf{r}_j^0 \right) \cdot \left(\mathbf{r}_j(t) - \mathbf{r}_j^0 \right) \right) \Big/ T} \,. \qquad (3.2)$$

where $\mathbf{r}_j^0$ is the base position vector (usually the crystal structure) for atom j.

The conventional $RMSF_j$ (root mean square fluctuation for atom j) measure of protein fluctuations (that can be compared to B-factors, by calculating B-factor using RMSF through $B - factor = (8\pi^2 / 3) * RMSF^2$)[16] is defined as

$$RMSF_j = \sqrt{\left(\int_0^T dt \left(\mathbf{r}_j(t) - \bar{\mathbf{r}}_j\right) \cdot \left(\mathbf{r}_j(t) - \bar{\mathbf{r}}_j\right) \right) \Big/ T} \,. \qquad (3.3)$$

where $\bar{\mathbf{r}}_j = \int_0^T dt\, \mathbf{r}_j(t)/T$ is the trajectory averaged position vector for atom j.

Clustering method and HP binding pocket profile calculation

The clustering is done using the "g_cluster" routine of GROMACS which is implemented in the software package using the algorithm as described in Daura et al.[47] For each input structure, the algorithm counts the number of neighbors within an RMSD cut-off (for the work summarized in this chapter, all the clustering is done using a 2 Å cut-off), takes the structure with the largest number of neighbors, along with all its neighbors, as a cluster, and eliminates all the structures within this cluster from the pool of structures. These steps are repeated for the remaining structures in the pool until no structures are left.

The HP binding pocket profile calculation is carried out as follows. First, the conformation of a protein snapshot is superimposed onto the starting crystal structure. Then, using the initial position of the mass center of HP in the crystal structure as the center, and a radius of 5 Å as a search sphere, a grid-point search is done within this sphere with an oxygen test atom for points which have Lennard-Jones interaction potential energy with protein atoms no greater than zero, and record them. Finally, connect those points to define the HP binding pocket profile.

3. Results and Discussion

HREM Replica Exchange Diagnostics

As usual, the most important issue for successful application of an HREM method to a certain protein simulation problem is to limit the number of systems required while still providing robust configurational sampling. In explicit solvent simulations, not scaling the solvent-solvent interactions should provide a large reduction in the number of systems relative to the temperature REM that scales all the degrees of freedom. The work for Met-enkaphalin summarized in Chapter 2 provides a good example of the above statement. In addition, since it can be expected that within a reasonable time span, the relatively rigid core and Loop 1 of HPPK should not change their conformations substantially, the interactions among those atoms can be left unscaled to reduce further the number of systems required. As also highlighted in the last chapter, the choice and optimization of the acceptance probability of attempted exchanges in the HREM is a central issue because there should be a "reasonable" acceptance probability, since for too low an exchange probability the rate of movement through configuration space is too small, while for too high an exchange probability the movement through configuration space is too slow. Table 1 lists the acceptance ratios for the HREM simulation for EcHPPK and YpHPPK. It is clear that all the acceptance ratios are within the optimal range (7-82%) suggested by Predescu and co-workers.[26] By using same system λ values for EcHPPK and YpHPPK, the acceptance ratios for YpHPPK seems to be much larger which is probably because YpHPPK is more flexible than EcHPPK.

Table 3.1. Acceptance ratios for the HREM simulations; (a) for EcHPPK and (b) for YpHPPK

Potential Index[(a)]	λ_a	λ_b	Acceptance Ratio
1<--->2	1.0	0.83	0.13
2<--->3	0.83	0.73	0.094
3<--->4	0.73	0.65	0.14
4<--->5	0.65	0.57	0.080
5<--->6	0.57	0.5	0.12

Potential Index[(b)]	λ_a	λ_b	Acceptance Ratio
1<--->2	1.0	0.83	0.64
2<--->3	0.83	0.73	0.50
3<--->4	0.73	0.65	0.43
4<--->5	0.65	0.57	0.41
5<--->6	0.57	0.5	0.37

To examine whether all the configurations (replicas) can visit a particular system (with a particular λ_i value) and whether given configurations are visited by all the systems, we display these time trajectories in Figures 1 and 2. (Since the plots for the third and fourth ns are similar to the plots for the fifth ns, only the plots for the fifth ns are displayed for convenience.) From the plots, it is clear that for the simulation of YpHPPK, all the systems (windows) can be visited by all the configurations (replicas) (a-f), and, conversely, all the configurations (replicas) can be visited by all the systems (windows) (g-l). Although the plots for the simulation of EcHPPK are not as uniform as the ones for YpHPPK, it is still true that configurations 2 and 3 can be visited by all the systems and systems 2 and 3 can be visited by all the configurations while all the systems have configurations with both higher and lower indices visiting them.

Comparing RMSD and RMSF for HREM trajectories with those for normal MD

As noted in the Introduction and Methodology sections in this chapter, the most flexible regions of HPPK are Loops 2 and 3, and these two loop regions comprise the atoms whose interactions were scaled. To see whether the HREM simulations improve sampling relative to the normal MD simulations, the RMSD and RMSF of the two loops are compared. The system 1 HREM trajectories are used for the comparison since it is system 1 that samples from the normal MD force field's (λ=1) canonical ensemble. The RMSD and RMSF are evaluated based on the backbone atoms and on all the atoms for each residue in the two loops. All trajectories used are ones obtained after the RMSD stabilized.

Figure 3.1. Migration Check for EcHPPK. (a)-(f): Migration of configurations (replicas) into and out of a given system (λ_i scale value). (g)-(l): Migration of windows (systems) into and out of a given configuration . The figures of a-f correspond to systems 1-6 and the figures of f-j correspond to configurations 1-6. Note that in view of the number of data points that are plotted it appears as if, at a particular time, several replicas occupy the same window or several systems visit the same replica–this does not happen,–looking at shorter intervals, the dots will be separate.

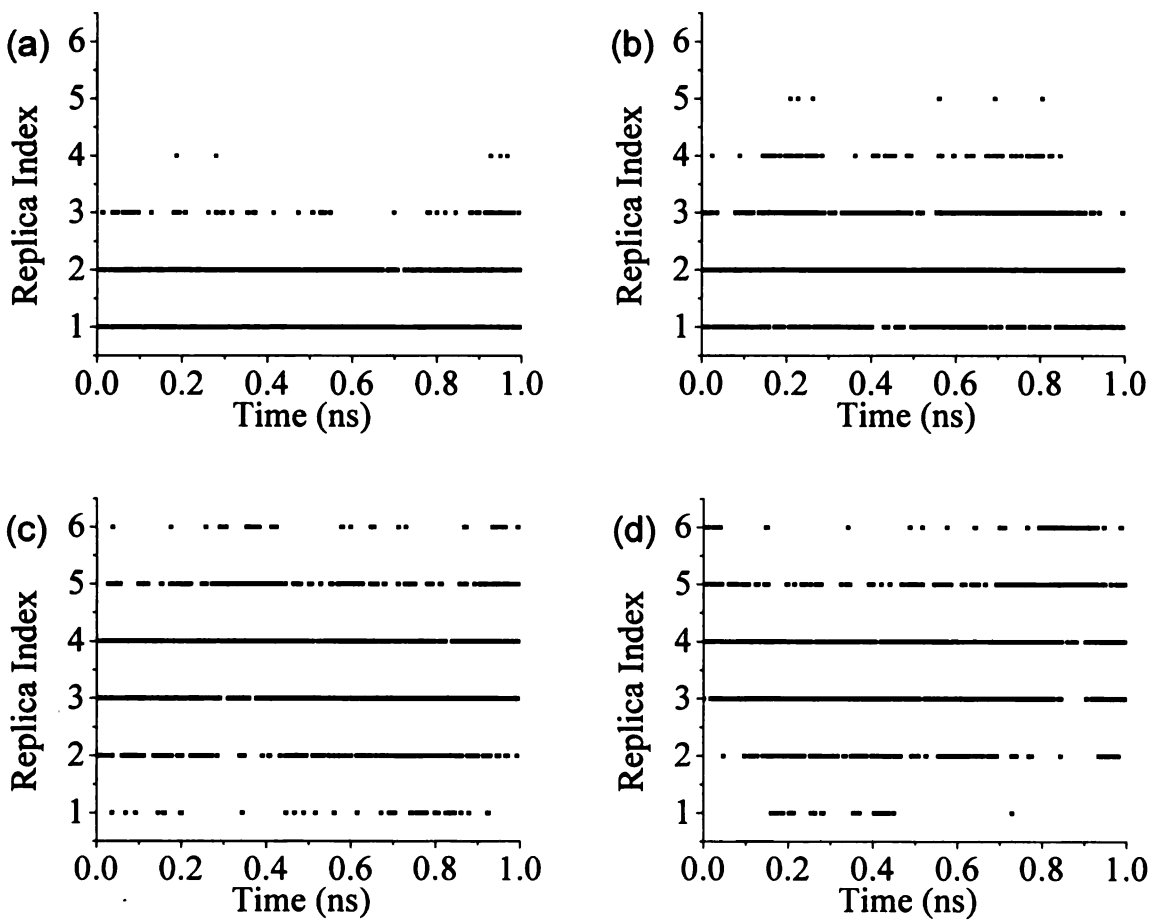


Figure 3.1 (continue)

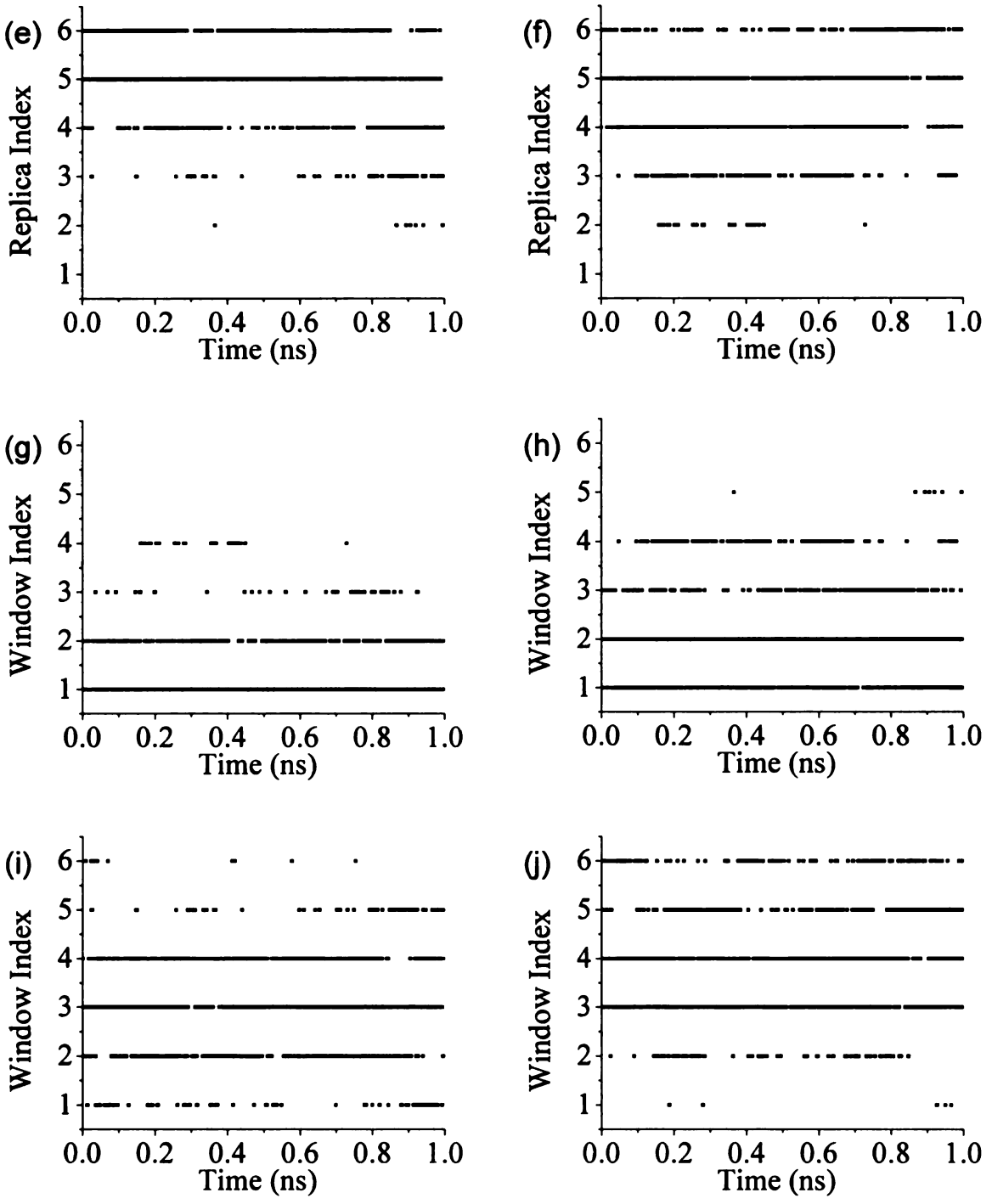


Figure 3.1 (continue)

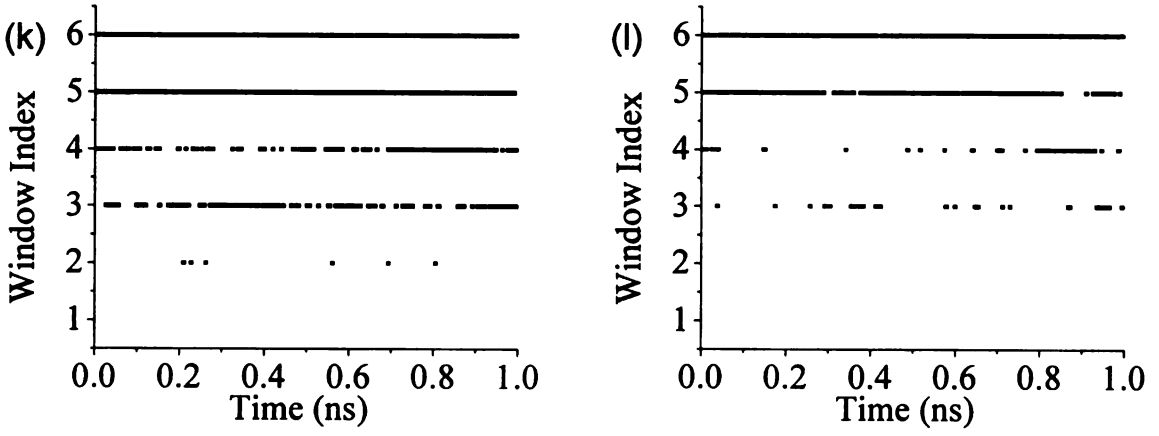


Figure 3.2. Migration Check for YpHPPK. (a)-(f): Migration of configurations (replicas) into and out of a given system (λ_i scale value). (g)-(l): Migration of windows (systems) into and out of a given configuration . The figures of a-f correspond to systems 1-6 and the figures of f-j correspond to configurations 1-6. (Note that in view of the number of data points that are plotted it appears as if, at a particular time, several replicas occupy the same window or several systems visit the same replica– this does not happen. Looking at shorter intervals, the dots will be separate.

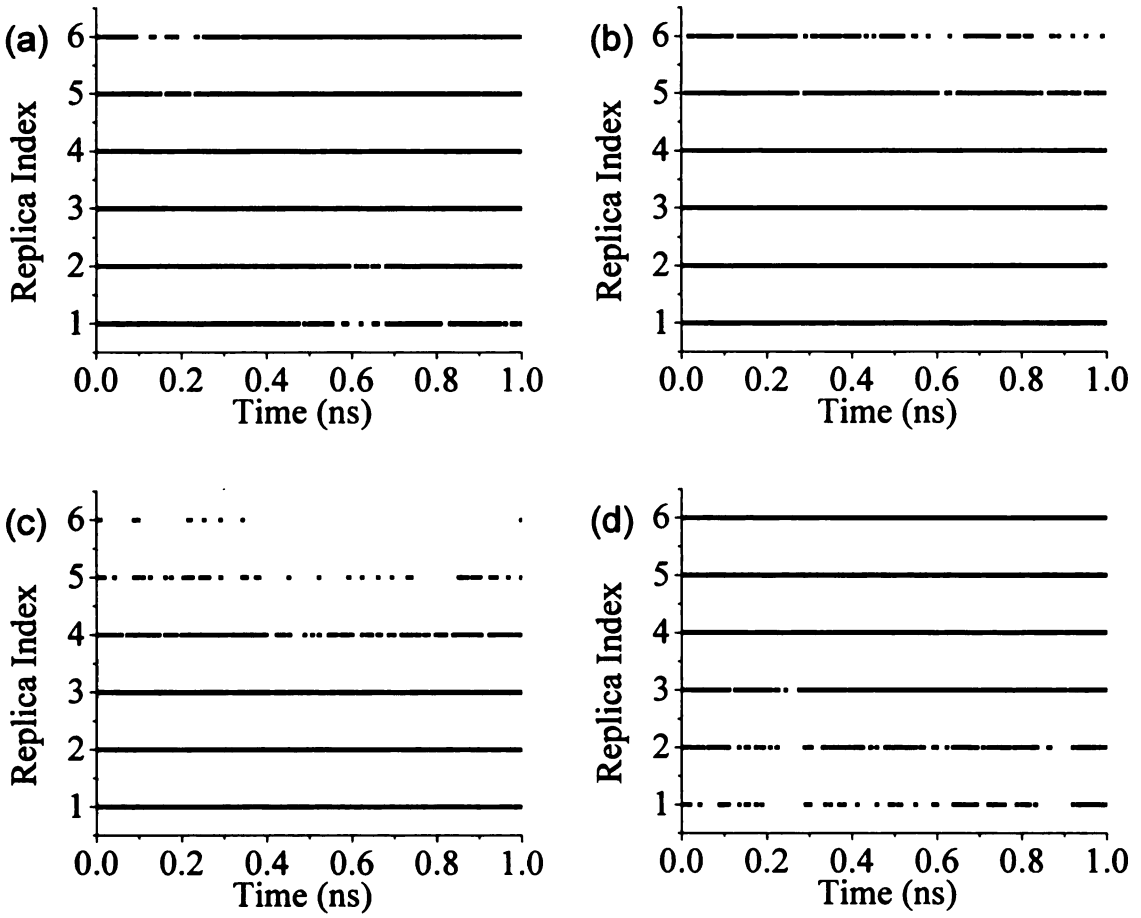


Figure 3.2 (continue)

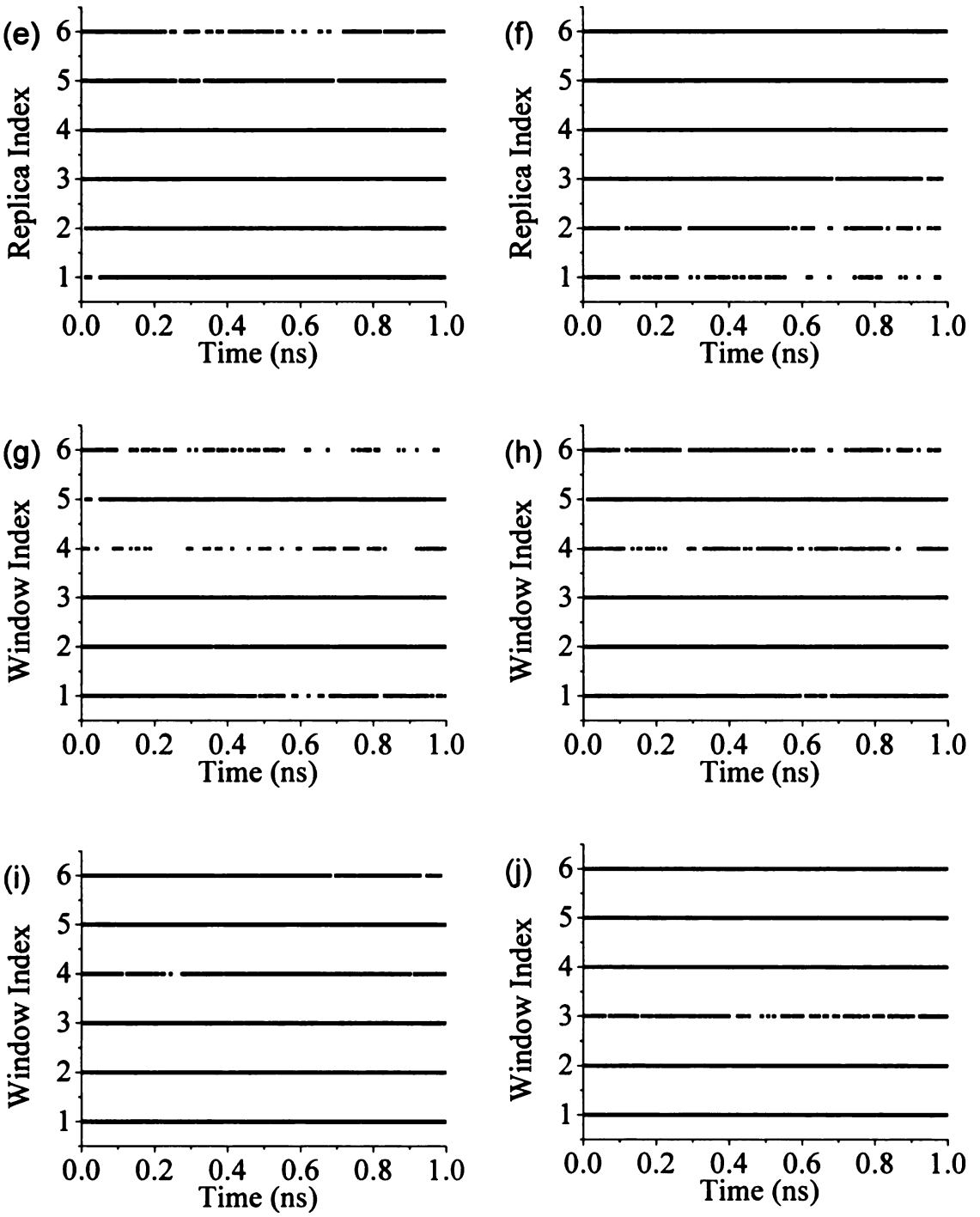
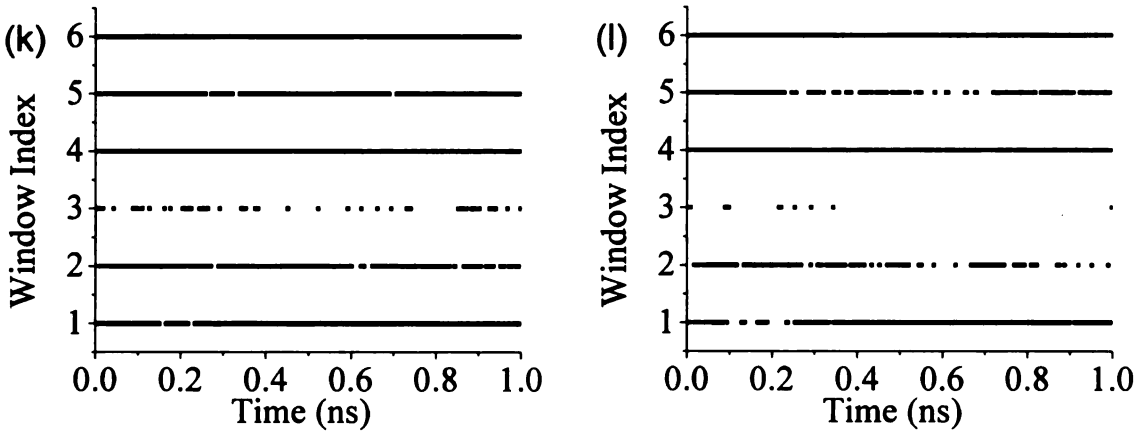


Figure 3.2 (continue)



The RMSD comparisons are displayed in Figures 3.3 and 3.4. Based on EcHPPK backbone atoms, the RMSD calculated for the normal MD trajecotry is around 3 Å, showing that the simulated conformations did not deviate far from the original crystal structure. The RMSD for most residues in Loop 2 and Loop 3 based on the HREM trajectories has a value between 3 and 5 Å with residues at the ends of two loops having smaller RMSD while two residues in the center of Loop 3 having a bit larger RMSD. The simulated conformations based on HREM deviate more from the starting structure than those based on normal MD trajectories from the starting structure. A larger deviation from the starting structure after simulation for a couple of nanoseconds has a better chance to occur since the starting structure was made by removing HP from the ternary crystal structure. Based on YpHPPK backbone atoms, using either the normal MD or HREM method, the resulting RMSD plots show that the simulated protein loop 2 and 3 conformations are substantially different from the starting crystal structure after several nanoseconds of simulation. Basing the RMSD calculations on all atoms of Loops 2 and 3, the changes of conformations for most residues are larger for both EcHPPK and YpHPPK either by regular MD or by HREM. This result is expected because the side chains are usually more flexible than backbones. The larger RMSD difference for YpHPPK versus EcHPPK may be due to the fact that the starting crystal structure for YpHPPK was obtained by simulating one monomer from the dimeric crystal structure.[80] As discussed in this chapter's methodology section about the starting crystal structure for YpHPPK, in the dimer crystal structure each monomer's Loop 2 interacts strongly with residues in the other monomer, and it is Loop 2 that shows the largest RMSDs from the crystal structure.

Figure 3.3. The RMSD for residues of Loop 2 and Loop 3 in EcHPPK: (a) and (b) are for RMSD based only on backbone atoms for Loop 2 and Loop 3, respectively, while (c) and (d) are based on all atoms. The RMSD based on normal MD are shown with solid lines and the RMSD based on HREM are shown with dashed lines. The RMSDs are average RMSDs calculated for all the snapshots in the two three nanosecond long trajectories based on normal MD and HREM picked when the RMSDs and RMSFs for the trajectories have been stable.

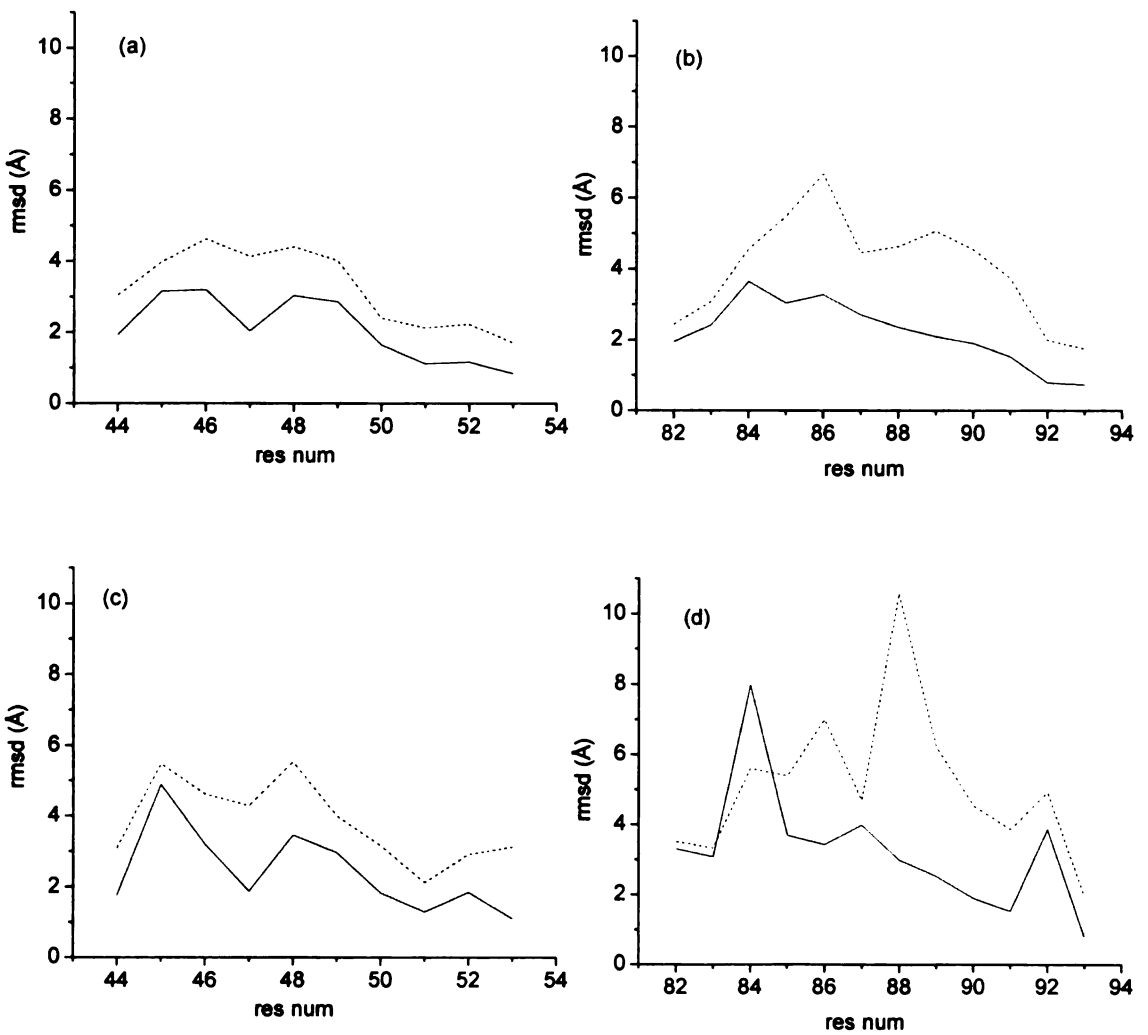
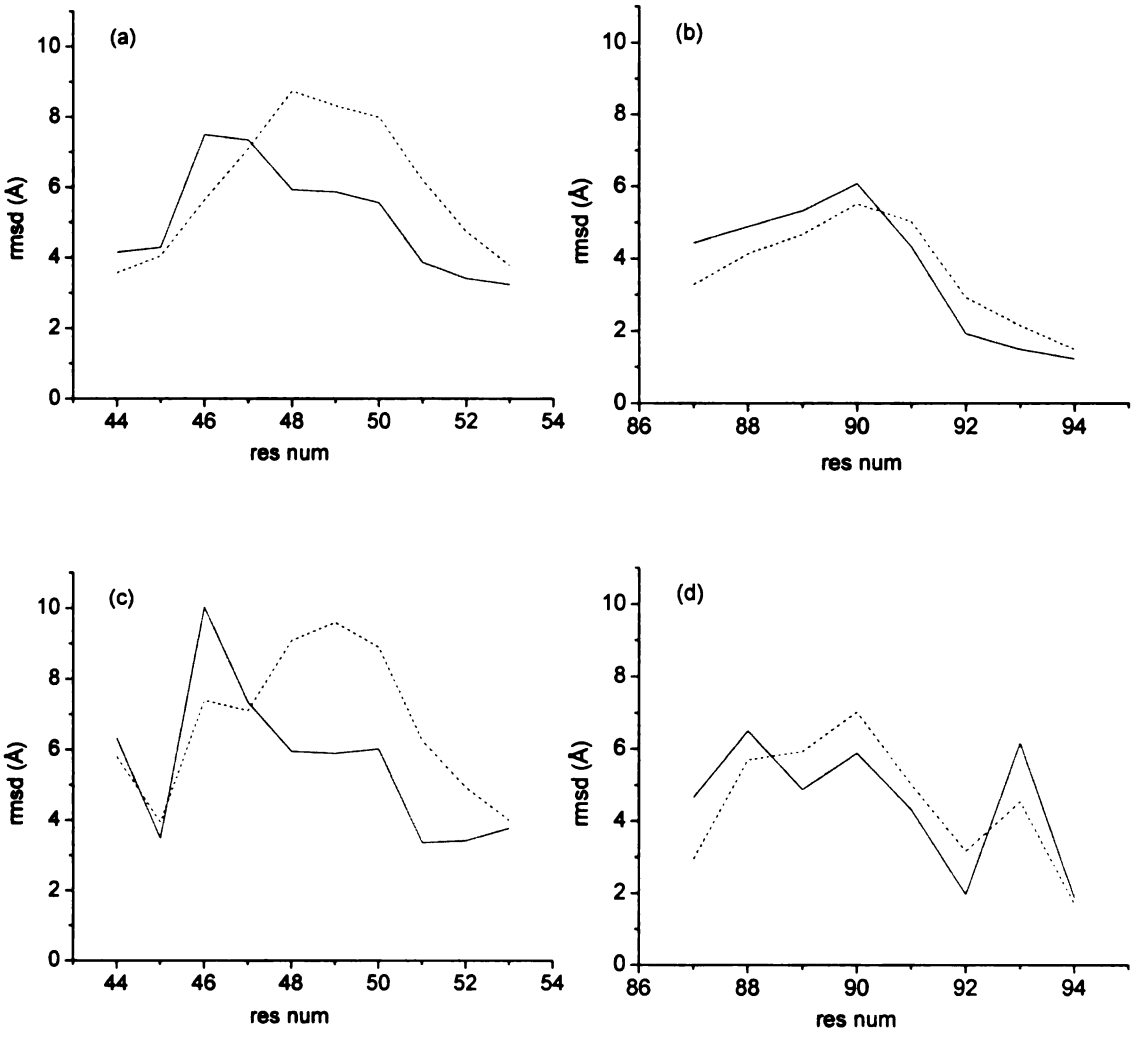


Figure 3.4. The RMSD for residues of Loop 2 and Loop 3 in YpHPPK: (a) and (b) are for RMSD based only on backbone atoms for Loop 2 and Loop 3, respectively, while (c) and (d) are based on all atoms. The RMSD based on normal MD are shown with solid lines and the RMSD based on HREM are shown with dashed lines. The RMSDs are average RMSDs calculated for all the snapshots in the two three nanosecond long trajectories based on normal MD and HREM picked when the RMSDs and RMSFs for the trajectories have been stable.



The Loop 2 residues of YpHPPK deviate more from the starting crystal structure than those of EcHPPK. Figure 3.5 shows that the Loop 2 conformations for the starting crystal structures of the two proteins are quite similar, while from the ternary structures it is clear that the HP binding pocket in HPPK is close to its loop 2. Therefore, the difference of the magnitudes of the deviations from the crystal structures observed for the Loop 2 residues in the two proteins suggests the possible existence of different HP binding pocket conformations for the two proteins.

The differences between the RMSDs based on HREM and normal MD trajectories are in the main not large. But, RMSD only shows the deviation from the starting crystal structure and does not contain direct information about protein flexibility. In particular, if the protein is trapped in a small area far from the starting point within the configuration space, the RMSD can be large. Therefore, RMSF comparisons should be more revealing of the extent of configurational sampling. If they turn out to be small, then the RMSD measures indicate that the various trajectories are confined to basins in configuration space that are displaced by varying degrees from the crystal structures. Figures 3.6 and 3.7 display the RMSF comparisons, which show that the RMSF based on HREM trajectories is consistently much larger than the RMSF based on normal MD trajectories for both EcHPPK and YpHPPK. The normal MD RMSF based on the backbone atoms is less than 2 Å for each residue of the two loops for the two proteins with the only exception that the Gly47 of YpHPPK has an RMSF of 2.2 Å. Even when side chains are included, only Arg84 has an RMSF greater than 3 Å. It is clear that the normal MD trajectories were trapped around some basin in the configuration space over the simulation time scale, after a stable RMSD was achieved. The RMSF obtained

Figure 3.5. Starting crystal structures for the simulations: (a) for EcHPPK (PDB entry 1q0n) (b) for YpHPPK (PDB entry 2qx0)

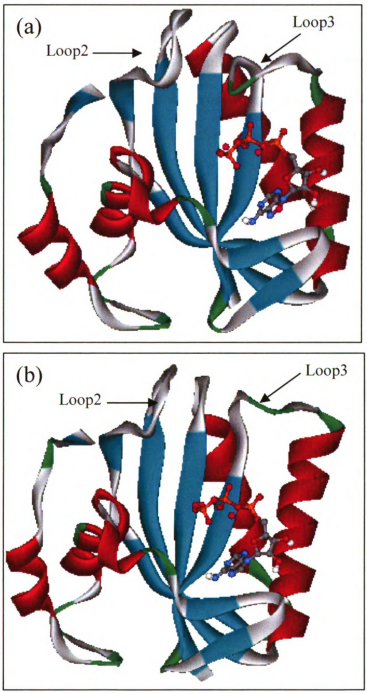


Figure 3.6. The RMSF for residues of Loop 2 and Loop 3 in EcHPPK: (a) and (b) are for RMSD based only on backbone atoms for Loop 2 and Loop 3, respectively, while (c) and (d) are based on all atoms. The RMSF based on normal MD are shown with solid lines and the RMSF based on HREM are shown with dashed lines The RMSFs are calculated for all the snapshots in the two three nanosecond long trajectories based on normal MD and HREM picked when the RMSDs and RMSFs for the trajectories have been stable.

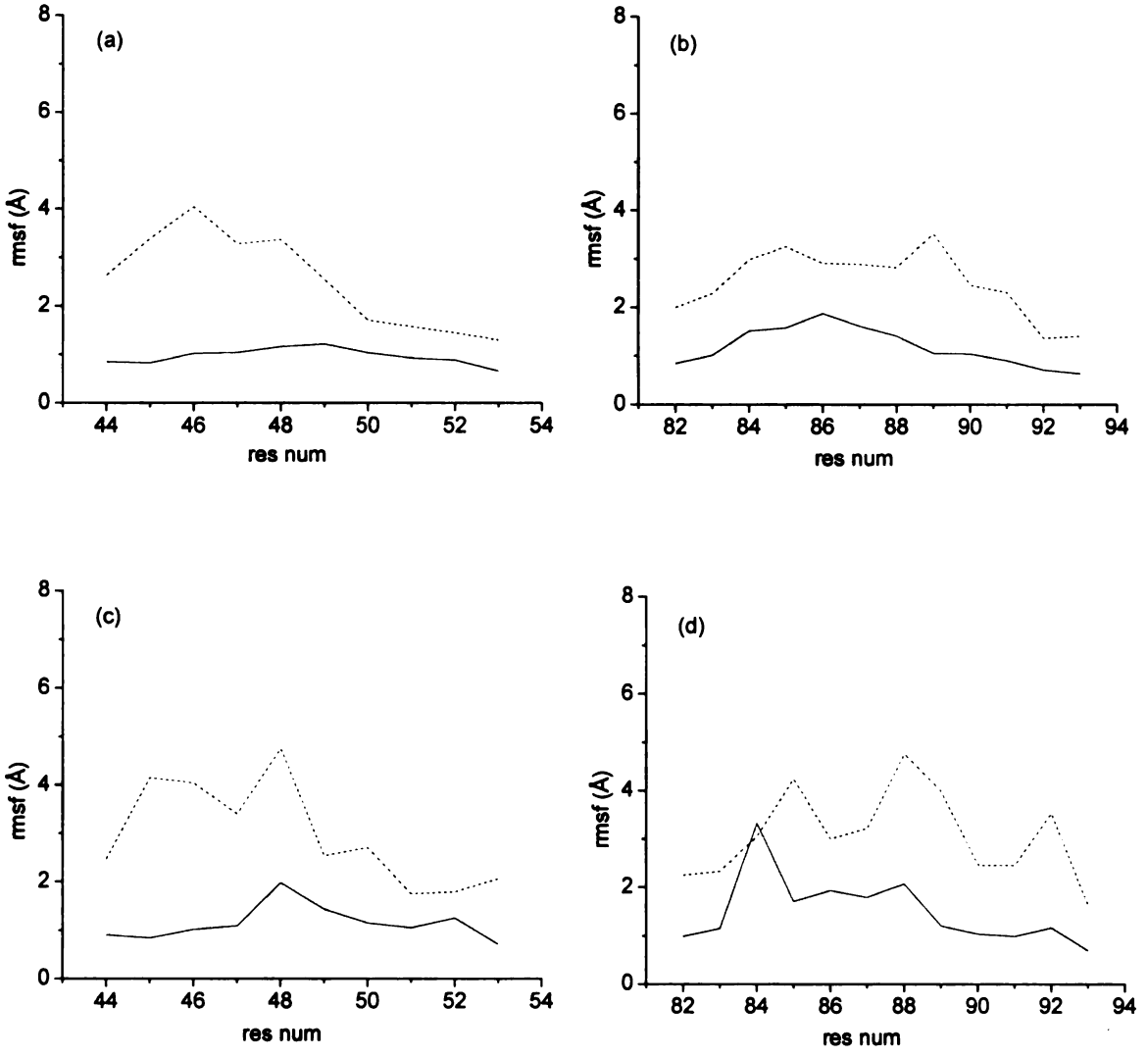
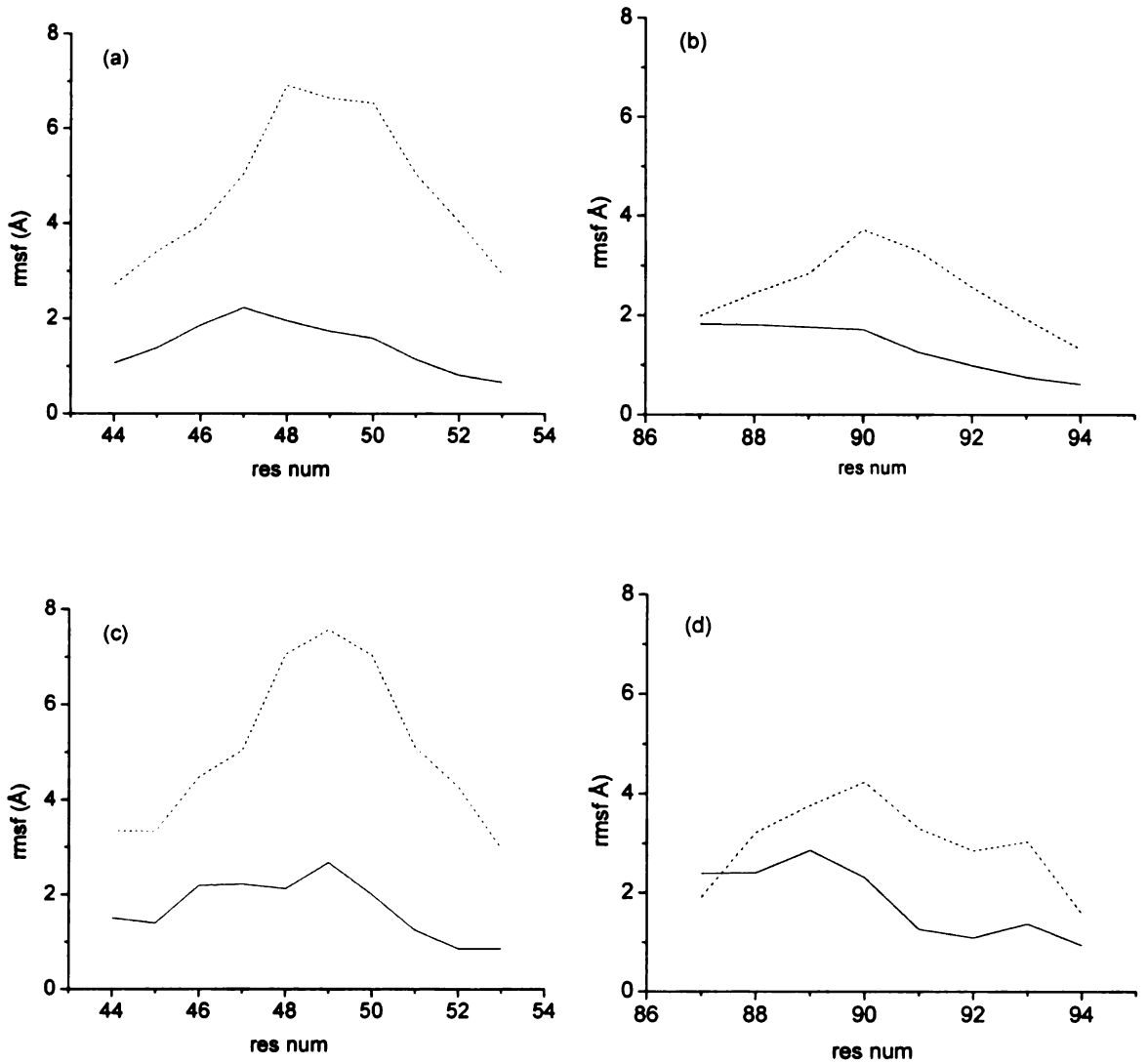


Figure 3.7. The RMSF for residues of Loop 2 and Loop 3 in YpHPPK: (a) and (b) are for RMSD based only on backbone atoms for Loop 2 and Loop 3, respectively, while (c) and (d) are based on all atoms. The RMSF based on normal MD are shown with solid lines and the RMSF based on HREM are shown with dashed lines. The RMSFs are calculated for all the snapshots in the two three nanosecond long trajectories based on normal MD and HREM picked when the RMSDs and RMSFs for the trajectories have been stable.



through HREM simulation is much larger, especially for residues in the middle of the loops that are less susceptible to the influence of the rigid core. Based only on backbone atoms, the residues in the middle of the loops for the two proteins all have an RMSF greater than 3 Å. It is noteworthy that the YpHPPK backbone of Loop 2 exhibits a much larger flexibility compared with that of EcHPPK. The RMSF based on backbone atoms is mostly more than 5 Å for the residues in the middle of Loop 2 of YpHPPK, with the central three having an RMSF close to 7 Å, while the similarly calculated RMSF is mostly between 3 and 4 Å for EcHPPK. Again, since the HP binding pocket is close to Loop 2, this large difference suggests that there might be different configurations of residues contributing to the HP binding pockets for the two proteins. The relatively large RMSF based on backbone atoms from trajectories obtained through the HREM simulations makes it possible to search for the potential existence of different configurations suited for binding ligands. A possible reason for this large HREM loop 2 RMSF difference between the two proteins may be due to the short length of the Loop 3 of YpHPPK. Loop 3 of YpHPPK only contains eight residues, making it difficult to interact with Loop 2, leading to the possibility of enhanced freedom of motion for Loop 2.

Binding pockets for EcHPPK and YpHPPK

To investigate HP binding pocket conformations, and their potential differences between EcHPPK and YpHPPK, residues that are close to HP in the respective crystal structures are selected. These residues include Thr42, Pro43, Pro44, Leu45, Try53, Asn55, and Phe123 for EcHPPK and Thr43, Lys44, Pro45, Leu46, Phe54, Asn55, and Phe123 for YpHPPK. (These residues will be referred to as pocket residues hereafter.) Only the backbone atoms of these residues are included in view of their greater stability

relative to side-chain atoms. The g_cluster method described in the methodology section of this chapter can be used to divide the total 3000 snapshots of the HREM trajectories for each of the two proteins into different groups according to the RMSD calculated based on backbone atoms of the pocket residues. By using a 2 Å cutoff in g_cluster, the HREM trajectories for EcHPPK and YpHPPK can be clustered into 7 and 8 clusters, respectively. (The populations of the clusters are shown in Table 3.2 and Table 3.3). The observation that the YpHPPK trajectory can be separated into more clusters using the same cutoff value may be due to the greater flexibility of Loop 2 of YpHPPK. The g_cluster method was also applied to the normal MD trajectories using the same 2 Å cutoff, but all snapshots turned out to be in the same cluster for EcHPPK while 99.9% of the snapshots were in the same cluster for YpHPPK. The backbone atom conformations of the pocket residues of the central structures are shown in light green in Figures 3.8 and 3.9 for cluster 1 and cluster 4 obtained from the EcHPPK trajectory and for clusters1, 2, 5 and 6 from YpHPPK. The crystal structure conformations are shown in bright yellow in the two figures for comparison. HP is also put back and shown in magenta in the two figures, according to its place in the respective crystal ternary complexes, to give a clearer picture of the binding pocket. Central structures of other dense clusters are not shown since they are either similar to the crystal structures or to the central structures shown in Figures 3.8 and 3.9. From Figure 3.8, it is clear that the central structure of cluster 1 for EcHPPK has residues Thr42 and Pro43 moved downward and residues Pro44 and Leu45 move in to be closer to the HP position in the x-ray ternary complex. And, there exist substantial changes of the Φ and Ψ dihedral angles of Pro44 and Φ dihedral angle of Leu 46 in the structure. These changes are mainly a consequence of the

Table 3.2. The number of snapshots in the clusters constructed for EcHPPK.

Cluster	Number of snapshots out of 3000
1	1719
2	1058
3	183
4	29
5	9
6	1
7	1

Table 3.3. The number of snapshots in the clusters constructed for YpHPPK.

Cluster	Number of snapshots out of 3000
1	1692
2	906
3	220
4	74
5	39
6	36
7	18
8	15

Figure 3.8. EcHPPK backbone atom conformations of pocket residues of the HP binding pocket of the central structures for clusters 1 and 4 of HREM snapshots: The left panel is for cluster 1 and the right for cluster 4. The central structure of each cluster is shown in green, the starting crystal structure is shown in yellow, and HP, in magenta, is placed according to the ternary complex crystal structure.

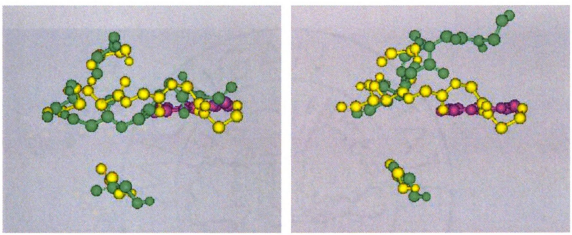
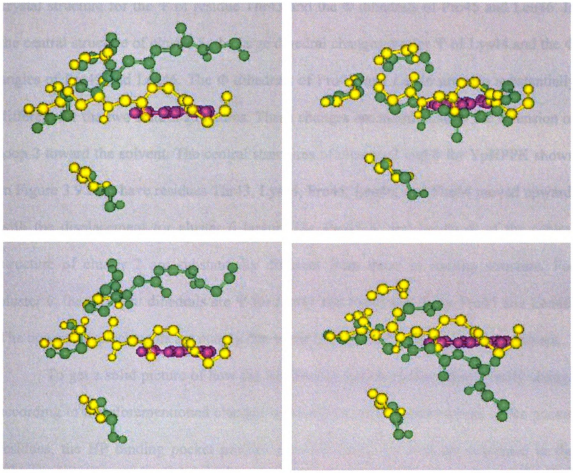


Figure 3.9. YpHPPK backbone atom conformations of pocket residues of HP binding pocket of the central structures for cluster 1, cluster 2, cluster 5, and cluster 6 of HREM snapshots. The left panel is for cluster 1 (top) and cluster 5 (bottom) and the right panels for cluster 2 (top) and cluster 6 (bottom). The central structure of each cluster is shown in green, the starting crystal structure is shown in yellow and HP, in magenta, is placed according to the ternary complex crystal structure.



bending down of the left half of loop 2 (residues 44 – 49) toward the protein core. The central structure of cluster 4 shown in Figure 3.8, has residues, Pro43, Pro44, Leu45, and Try53, moved upward, and there are substantial changes of the Φ and Ψ dihedral angles of Pro43. These changes are mainly due to loop 2 transiting toward the solvent. For YpHPPK, Figure 3.9 shows that the central structures of clusters 1 and 5 both have residues Thr43, Lys44, Pro45, Leu46, and Phe54 displaced substantially upward. In the central structure of cluster 1, there are substantial dihedral changes from the starting crystal structure for the Ψ of residue Thr43 and the Φ dihedrals of Pro45 and Leu46. In the central structure of cluster 5, the large dihedral changes are for Ψ of Lys44 and the Φ angles of Pro45 and Leu46. The Φ dihedrals of Pro45, and Leu46 are also substantially different for the two central structures. These changes are mainly due to the extension of loop 2 toward the solvent. The central structures of clusters 2 and 6 for YpHPPK shown in Figure 3.9 both have residues Thr43, Lys44, Pro45, Leu46, and Phe54 moved upward, with the displacement for cluster 6 larger. The Pro45 Ψ and Leu46 Φ of the central structure of cluster 2 are substantially different from those in starting structure. For cluster 6, the different dihedrals are Ψ for Lys44 and Pro45 and Φ for Pro45 and Leu46. The upward displacements are mainly due to the extension of loop 2 toward the solvent.

To get a solid picture of how the HP binding pocket conformations really change according to the aforementioned changes in backbone atom conformations of the pocket residues, the HP binding pocket profiles obtained using the methods presented in the methodology section in this chapter are shown in red in Figures 3.10 and 3.11 for EcHPPK and YpHPPK, respectively. It is clear from the two figures that when Loop 2 moves downward towards the core compared with the starting crystal structure (cluster 1

Figure 3.10. The HP binding pockets profiles for EcHPPK. The HP binding pocket profiles are shown in bright red in the following plots. (a) shows the profile for the starting crystal structure, and (b) and (c) show profiles for the central structures of cluster 1 and cluster 4, respectively.

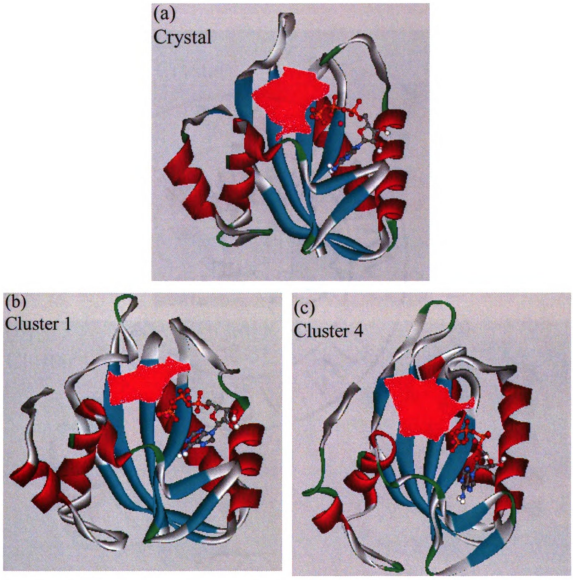


Figure 3.11. The HP binding pockets profiles for YpHPPK. The HP binding pocket profiles are shown in bright red in the following plots. (a) shows the profile for the starting crystal structure, and (b), (c), (d), and (e) show profiles for central structures of cluster 1 cluster 2, cluster 5 and cluster 6, respectively.

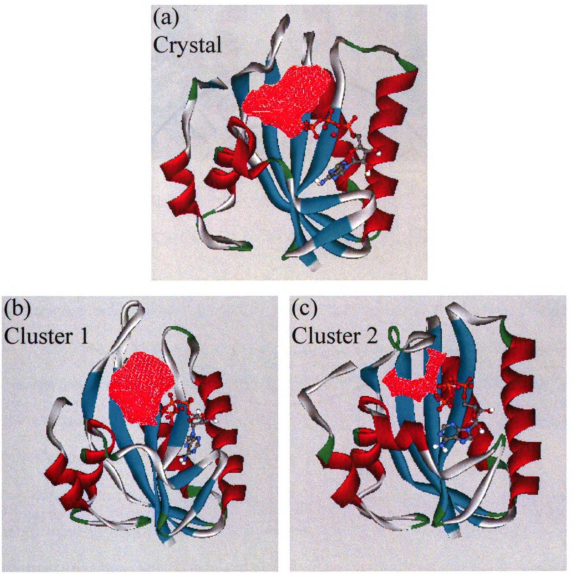
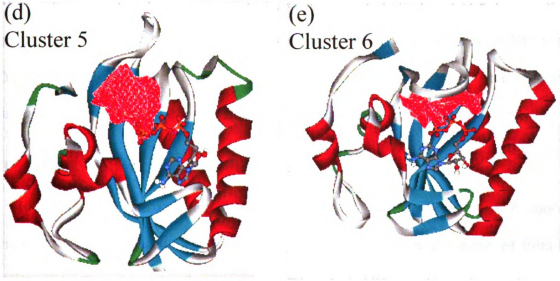


Figure 3.11 (continue)



for EcHPPK and clusters 2 and 6 for YpHPPK), the HP binding pockets get narrower, while they get wider when Loop 2 moves upward towards the solvent (cluster 4 for EcHPPK and clusters 1 and 5 for YpHPPK). Figures 3.10 and 3.11 also show that the HP pocket conformations for the central structures shown are quite different for EcHPPK and YpHPPK. This finding supports the hypothesis that it might be possible to design inhibitors that are effective for only one of the two proteins.

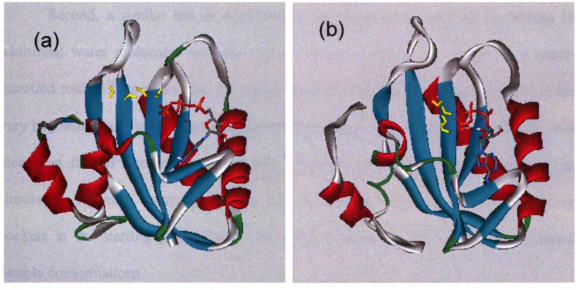
Test of adding water molecules in HP binding pockets

MD simulation trajectories sometimes depend on the starting state. A concern here is that for the results discussed in this chapter, HP was simply removed from the ternary structure to make the starting structure. In addition, from the configuration generated by our start up protocol it was clear that no water molecule was left in the region that was occupied by the HP before its removal. Hereafter, we will refer to this region as HP binding pocket region. The CUKMODY start up protocol consists of placing water molecules on a grid at the density appropriate to the normal water density for T=303. Then, all water molecules that overlap protein atoms are eliminated, based on a van der Waals overlap criterion. Since no waters were left in the HP pocket, it was not clear whether there would be some substantial difference in simulated trajectories if water molecules could be added into the HP binding pocket at the beginning of the MD.

To investigate the issue, extensive tests were carried out by adding water molecules into the HP binding pocket of the starting structure for the EcHPPK simulation, as the following describes. (Note that water molecules can be added into the HP binding pocket region because the region has a somewhat flat profile, a feature that leads to the non-specific grid routine tending to eliminate waters from such a region.) In a first try, four

water molecules were added into the HP binding pocket region as shown in Figure 3.12 (a). (Note: other water molecules remaining after water removal-routine described above were not shown in the figure, because otherwise there would be more than 7000 water molecules in the figure, and the four separately added in water molecules could not be seen. However, in terms of simulation parameters, there were no differences between the water molecules remaining after the removal routine and the four water molecules added separately.) The configurations (distances and orientations) of the four added water molecules were adjusted so that hydrogen bonds (to define a hydrogen bond, the distance between two oxygen atoms is set to be around 2.7 Å and the angle of oxygen-hydrogen-oxygen is set to be close to 180 degrees) are present between any two adjacent water molecules. Then, normal MD was conducted for a nanosecond with the structure shown in Figure 3.12 (a) as starting structure. In Figure 3.13 (b), the configuration after 1 ns of simulation is shown, it is clear that the water molecules originally put inside the HP binding pocket region moved out of the region to some places between Loop 2 and Loop 3. Those places between Loop 2 and Loop 3 would contain water molecules simply by executing the normal start-up procedure. The actual time needed for the water molecules to move away from the HP binding pocket region is only about 200 ps. One possible reason for water molecules to leave the HP pocket region shortly after being added explicitly is that the residues close to the HP binding pocket region (Thr42, Pro43, Pro44, Leu45, Try53, Asn55, and Phe123) are mostly hydrophobic. Another possible reason is that the HP binding pocket region is in part solvent exposed and also close to the highly-charged ATP phosphate tail. The high charge on ATP phosphate tail and hydrophilic property of residues around it may also help to explain the observation that the added

Figure 3.12. (a) The configuration of EcHPPK and four hydrogen bonded added water molecules. (b) The configuration after 1 ns normal MD simulation with the starting structure as shown in (a). The added water molecules are shown in yellow.

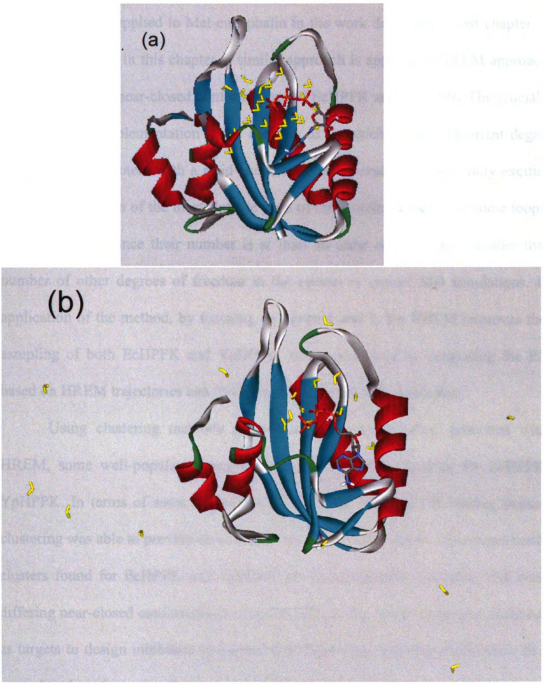


four water molecules first moved to the region between Loop 2 and Loop 3 close to the ATP phosphate tail after leaving the HP binding pocket region. .Therefore, it seems that adding those four water molecules should not have substantial influence on the HREM simulation results of HPPK.

Second, a similar test as described in the above paragraph, but by adding 16 additional water molecules into the region between Loop 2 and Loop3 by a water-insertion routine[105] to make sure the region was fully occupied by water molecules at the very beginning, was tried. The result is shown in Figure 3.13. From Figure 3.13 (b), it is clear that the HP binding pocket region no longer had water molecules after 1 ns simulation. Again, it seems that adding water molecules explicitly into the HP binding pockets at the starting point should not have substantial influence on the simulated sample conformations.

Moreover, the test using 4 hydrogen-bonded water molecules and the 16 additional water molecules was tried again by using smaller van der Waals constants for the added 20 water molecules to reduce possible large repulsion at the beginning. Because 4 of the water molecules were added in manually, relatively large vdw repulsion might exist at the beginning, and cause these water molecules to be bounced out of the HP binding pocket region quickly. However, even with smaller vdw constants, after 1 ns, still no water molecules stayed at the place that was directly under Loop 2 and was occupied by HP in the starting ternary structure. In summary, it seems that there should be no big difference between starting simulations with an HPPK structure with water molecules added into the HP binding pocket region and starting simulation simply with an HPPK structure.

Figure 3.13. (a) The configuration of EcHPPK with 20 added water molecules. (b) The configuration after 1 ns normal MD simulation with the starting structure as shown in (a). The added water molecules are shown in yellow.



4. Concluding Remarks

A Hamiltonian Replica Exchange Method (HREM) molecular dynamics (MD) approach was successfully applied to Met-enkephalin in the work described in last chapter. In the work summarized in this chapter, a similar approach is applied. A HREM approach was used to study the near-closed conformations of EcHPPK and YpHPPK The crucial point of the HREM implementation is the decision as to which are the important degrees of freedom. For a protein with a rigid core simulated in explicit solvent, only exciting the degrees of freedom of the most flexible parts of the protein (usually just some loops) is a great advantage since their number is at least an order of magnitude smaller than the number of other degrees of freedom in the system in typical MD simulations. In our application of the method, by focusing on Loops 2 and 3, the HREM improves the MD sampling of both EcHPPK and YpHPPK, when monitored by comparing the RMSFs based on HREM trajectories and those based on normal MD simulation.

Using clustering methods focused on the conformations generated with the HREM, some well-populated, near-closed structures were obtained for EcHPPK and YpHPPK. In terms of some of the key residues that form the HP binding pocket, the clustering was able to provide some distinct residue conformations. Most importantly, the clusters found for EcHPPK and YpHPPK are distinguishable, indicating that there are differing near-closed conformations suited to HP binding. Those structures could be used as targets to design inhibitors to test the hypothesis that selecting inhibitors to fit some near-closed conformation that is significantly more accessible in the targeted protein than in its homolog is an effective strategy to enhance inhibitor specificity. If successful, those

designed inhibitors might be used as new narrow band antibiotics to help deal with the problem of antibiotic resistance.

Chapter 4. Studies of HPPK-ATP conformation space and ATP binding through using enhanced MD methods

1. Introduction

For the simulation processes conducted for the work described in Chapter 3, one interesting observation is that the ATP is very stable. During all the simulations, ATP always stays in its binding pocket with no substantial change of its conformation. Experimental results show that HPPK binds ATP (in those experiments, the non-reactive analog AMPPCP was used) with high affinity and that ATP binds first, and slowly, followed by very rapid HP binding[106]. Without the presence of ATP, HPPK does not bind HP in measurable quantities. These data suggest that there is a "proto-pocket" for HP binding that is formed by first binding ATP. The slow ATP binding is consistent with the idea that there is conformational flexibility of HPPK, and that ATP selects from a conformational ensemble. Once ATP is bound, the HPPK-ATP complex may stabilize to a set of conformations appropriate for the rapid uptake of HP. This chapter is devoted to the study of the HPPK-ATP conformation space and possibly how ATP can bind to HPPK.

Similar to the last chapter, the HPPK–ATP (only E. coli HPPK is studied in this chapter) is considered in this chapter as the basic species of the investigation (HPPK-ATP notation is used as shorthand for the complex of HPPK, ATP and the two Mg^{2+} ions that are required for the catalytic activity) A natural and simple approach to the investigation, of course, is to use conventional MD simulation to generate trajectories of conformations that can be used to address these issues. However, as shown before in Chapter 2 and Chapter 3, conventional MD often suffers the problem that integrating

Newton's equations of motion with femtosecond scale time step will have difficulty reaching times that can capture substantial conformational transitions and fluctuations that may be occurring on the micro to even seconds time scales owing to the complex configuration space that has substantial barriers in the high-dimensional potential energy surface. A number of methods exist that speed up transitions over potential energy barriers. One class of method introduces restraints that operate on the atoms to direct the trajectory between some initial and final state[9]. The restraints can range from being applied to all the atoms to drive the system between the (known) endpoints with complete conformity. Or, they can be applied to just a particular atom-atom distance as would be appropriate for obtaining a one-dimensional potential of mean force. Another possible approach is to use a reweighting method, which modifies the potential surface to generate a trajectory that more readily surmounts barriers.[27-30,33,34,107] The trajectory is then re-weighted back at each step to restore Boltzmann sampling. (More details about reweighting methods are provided in Chapter 1.) Cukier and Morillo developed a version of reweighting that can be referred to as targeted reweighting.[30] The term targeted implies that rather than modify the entire potential surface, only very specific terms in the potential are to be modified. Targeting can provide the flexibility to address different impediments to overcoming specific barriers on the potential energy surface. Targeted reweighting can be viewed as the opposite extreme of the global, temperature REM. (More details about REM are provided in Chapter 1 and Chapter 2.) The targeted reweighting method was successfully applied by the authors onto a simple testing protein model consisting of a chain of connected beads characterized by dihedral angles and the van der Waals interactions among the beads.

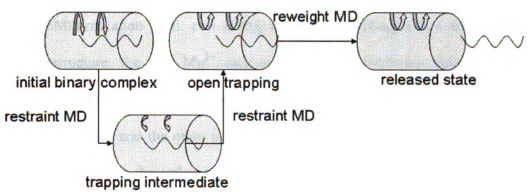
In the work summarized in this chapter, a combination of a restraint method and a targeted reweight method was used to explore the conformation space and mechanism of ATP binding in the HPPK (More details about HPPK are provided in the first introduction chapter and last chapter.) The simulation protocol is summarized in Figure 4.1. The cylinder denotes a set of residues that form a binding pocket for ATP (represented as a wiggly line) with orientations of key residue side-chains that are important for binding ATP denoted by curved arrows. The initial HPPK-ATP binary complex is based on the ternary crystal structure. Restraint simulations are used to direct HPPK-ATP toward the apo crystal structure[74] (pdb entry 1HKA). This restrained simulation was done in the hope that the ATP might move away from its binding pocket during the process. However, the simulations show that ATP remains bound throughout the process even though residues initially trapping ATP gradually move away, through ATP trapping intermediates, until reaching ATP trapping states that have an apo-like HPPK structure. ATP is still held in trap states by a network of hydrogen bonds and salt bridges that operate between ATP-$2Mg^{2+}$ and a number of HPPK core residues. The reweight simulation that starts from this last trap state then shows that ATP can separate from HPPK through a series of breaking and making hydrogen bonds and salt bridges.

2. Methodology

Molecular Dynamics Simulations

As similar in previous chapters, the CUKMODY protein molecular dynamics code was used to carry out the simulations. The code was modified to incorporate the

Figure 4.1. A schematic representation of the MD simulation protocol. The cylinder denotes residues that form a binding region for ATP with key residue side-chain orientations denoted by curved arrows. ATP is represented as a wiggly line. The initial HPPK-$ATPMg_2$ binary complex is constructed from the ternary crystal structure. Ten consecutive restraint simulations direct HPPK-$ATPMg_2$ toward the apo crystal structure. $ATPMg_2$ remains bound throughout the process even though residues initially trapping $ATPMg_2$ gradually move away, through trapping intermediates, until reaching open trapping states. The reweight simulation then shows that $ATPMg_2$ can separate from HPPK through a series of breaking and making hydrogen bonds and salt bridges.



reweight method as discussed below.

All simulations were run at 303 K under fixed number, volume and temperature (NVT) conditions.[15] The simulations were carried out in a cubic box with sides of 64.1 Å, having 7671 waters added after waters overlapping the protein were removed. For the evaluation of the electrostatic and the attractive part of the Lennard-Jones energies and forces, the PME method was applied with a direct-space cutoff of 9.0 Å, an Ewald coefficient of 0.32, and a $72\times72\times72$ reciprocal space grid. Five sodium cations were added to neutralize the system. Bond lengths were constrained with the SHAKE algorithm[8] allowing a 2 fs time step and the temperature was globally controlled with a Berendsen thermostat[10] with relaxation time of 0.2 ps.

The starting structure for simulation of HPPK-ATP was the same as the one used for HREM simulation of E. coli HPPK described in Chapter 3. For constructing the starting structure, the two Mg^{2+} cations were each covalently linked to an oxygen phosphate atom of ATP. One magnesium, designated as Mg1, is linked to an alpha phosphate oxygen and the other is linked to a gamma phosphate oxygen, designated as Mg2, in accord with their placement in the crystal structure. The ATP phosphates were assumed fully deprotonated (ATP thus has a total charge of –4) in agreement with the ATP protonation state when ligated to Mg^{2+}.[104]

Restraint method

For the restrained MD simulations that were used to transit the protein configuration between two desired conformations, restraints on distances between mass centers of paired atom groups were introduced. Harmonic restraint potentials $V=(k/2)(x-x_0)^2$ with force constant $k=2$ kcal/mol/Å^2 were used to restrain the

current distance x to be around a desired target distance x_0. A small restraint force constant was used to permit the HPPK-ATP complex to fluctuate extensively without being dominated by the effects of the restraints. The restraints were applied in a step-wise manner, by dividing the transition between the initial and final states into 10 sequential windows. Each consisted of 500 ps of MD simulation with restraints set as stated above with $x_0 = n\left(x_{final} - x_{initial}\right)/10 + x_{initial}$, n the window number running from 1 to 10, and x_{final}=x_{apo} and $x_{initial}$=x_{close} denoting a distance in the apo and closed (the ternary complex with HP removed) crystal structures, respectively.

Reweighting method

The reweighting method[27-31] to accelerate configurational sampling is described in detail in Chapter 1. The essential idea underlying reweighting method is that a modified potential surface can be used during actual simulation to reduce barriers and then, by reweighting back when analyzing trajectory data, the correct ensemble average can be obtained. In extreme, if the total potential is modified multiplicatively then it simply corresponds to changing the temperature of the simulation. The essence of targeted-reweighting method developed by Cukier and Morillo[30] is that only specific terms in the potential energy are targeted, and these terms are chosen based on specific information such as the presence of strong electrostatic interactions corresponding, for example, to salt bridges between residues or what we still will refer to as salt bridges between charged portions of ATP and the magnesium cations and ionized residues.(These are the potential terms targeted in the work summarized in this chapter).

For specific potential term V_{ij} for certain pair of atoms i and j to be targeted, a linear scaling is used in the work summarized in this chapter, $V_{ij}^* = \lambda V_{ij}$. A number of λ values were tried. The actual results are quite sensitive to the λ value chosen with too large a value (closer to 1) leading to no effect and too small leading to a release that is so rapid that the sampling is very poor. The choice of λ=0.71 led to escape on the ns time scale. The stages that are found took about 1 ns for this choice, and the simulation was run for a total of 2 ns to make sure that ATP, once unbound, did not rebind.

As also mentioned in Chapter 2 and 3, the Ewald[9] method, used in our CUKMODY program, requires the system to be electrically neutral. Thus, in general, a direct inclusion of the scaling factor into the Ewald calculation is hard to implement because it would require finding an overall neutral set of interactions to scale. (The Ewald method calculates energy and forces in reciprocal space that is atom-based versus the real space part that is interaction-based.) Since the modified potential surface does not have to correspond to a real surface this is not, in principal, a problem. However, as in any modified potential method, if the trajectory becomes too distorted from the true potential trajectory, the reweighting method becomes counterproductive. To deal with this issue, because we will only target interactions within the protein and/or ATP, and these interactions are dominant within the same MD image cell, the potential modification will only be applied to the primary cell and the real space energy and forces. For example, to target an electrostatic interaction (the only type of interaction targeted in the work summarized in this chapter) between atoms labeled 1 and 2, the interaction energy V_{12} and corresponding atom forces $\mathbf{F}_1$ and $\mathbf{F}_2$ are calculated using the Ewald method with $\lambda = 1$, which is just the normal Ewald MD procedure. Next, the differences

between using $\lambda = 1$ and using the scale factor λ are calculated directly from Coulomb's law in the primary cell as $V_{12}^{diff} = (\lambda - 1)(q_1 q_2 / r_{12})$ and the corresponding forces $\mathbf{F}_1^{diff}$ and $\mathbf{F}_2^{diff}$ are obtained by differentiation. Then, the above differences are added to the Ewald V_{12}, $\mathbf{F}_1$ and $\mathbf{F}_2$ calculated before to obtain the net values $V_{12}^{net} = V_{12} + V_{12}^{diff}$, $\mathbf{F}_1^{net} = \mathbf{F}_1 + \mathbf{F}_1^{diff}$ and $\mathbf{F}_2^{net} = \mathbf{F}_2 + \mathbf{F}_2^{diff}$. These net forces are used to advance the system configuration and the net energy is used to obtain ΔV. The other (van der Waals and internal) interaction terms can be obtained by a similar scheme.

3. Results and Discussion

Restraint MD of HPPK-ATP

As described in the methodology section, the simulations are based on the *E. coli* ternary complex of HPPK, AMPCPP with two magnesium cations and HP, with HP removed and AMPCPP replaced by ATP. The fully deprotonated ATP phosphates (total charge –4) with the two ligated Mg^{2+} cations provides a neutral, though highly polar, ligand for HPPK. A primary 7 ns conventional MD simulation of this binary complex led to little change in conformation from the ternary (closed) structure. Furthermore, we also carried out simulations using the HREM approach similar to the one used for the work of last chapter to see if HPPK by itself, again started from the ternary crystal structure with both ATP and HP removed, would take on a more apo-like conformation but these simulations also led to modest rearrangements. For these reasons, the HPPK-ATP binary complex was chosen to be the focus and a particular set of restraints were picked to drive the loops towards a conformation similar to the apo structure. Restraints were picked

partly based on the observation that long side-chains of some of the Loop 3 residues (principally Arg82, Arg88 and Trp89) drape over the ATP when HPPK is closed, forming a cylinder whose front face is composed of these residues, as schematized in Figure 1. These side-chains should be pushed out toward the solvent in conformity with the apo crystal structure. Other restraints were chosen to make the backbone conformation of Loop 2 and Loop 3 similar to that of the apo crystal structure. To achieve this transformation, harmonic restraints (with force constant k=2 kcal/Å^2) on 15 distances, measured between the mass centers of the atom pairs listed in Table 4.1, were introduced. The restraints were applied in a step-wise manner, by dividing the transition from the HPPK-ATP closed to open form into 10 sequential phases, with each phase consisting of 500 ps of MD, to interpolate between the endpoints.

Figure 4.2 displays a superposition of representative structures selected from three of the above ten 500 ps time intervals corresponding to windows 1, 5 and 10. From the figure, it is clear that the cores of the three structures are well-aligned, which suggests that even under the forces arising from converting HPPK-ATP from its closed to open forms, there is minimal perturbation of the core, again showing the stability of the protein core. By comparing the loop conformations (including both backbone and important long side-chain conformations, such as Arg82, Arg88 and Trp89), it is clear that Loop 2 and Loop 3 were pulled gradually from closed-oriented to open-oriented conformations. In particular, the side chains of Arg82, Arg88, and Trp89 that point inward to seal ATP in the binding pocket of the closed form progressively move to point out to the solvent in the open form. Another key feature to note is that the ATP position is very stable along the trajectory, with all its parts hardly moving. Our hope was that once the side-chains

Table 4.1. The atom(s) whose distances are used for the restraint simulations

Atom(s) 1	Atom(s) 2
CA[(a)] Gly46	CA Asp97
CA Pro47	CA Asp97
CA Asp49	CA Asp97
CA Arg82	CA Asp97
CA Arg84	CA Asp97
CA Ala86	CA Asp97
CA Arg88	CA Asp97
CA Trp89	CA Asp97
CA Arg92	CA Asp97
CZ[(b)] Arg82	CA Asp97
CZ Arg84	CA Asp97
CZ Arg88	CA Asp97
CZ Arg92	CA Asp97
CZ[(c)] Arg92	CA Leu78
CE2[(d)] Trp89	CA Asp97

(a) CA denotes a C-alpha atom; (b) CZ stands for the ε carbon and amino groups of the Arg side-chain; (c) This restraint was not used during the ten-window restraint procedure, but was later realized to be important to ATP separation from HPPK. It was added later, while holding the loops open in the tenth window; (d) CE2 denotes a carbon atom of the 6-member aromatic ring.

Figure 4.2. Superposition of representative structures from three windows along the restraint pathway from closed to open. The structure for window 1, which is quite similar to the closed form, is colored in yellow; the structure for window 5 is colored in purple; the structure for window 10 is colored using the normal convention. ATP in each window is shown in "ball and stick" mode, and residues Arg82 and Trp89 are shown in "stick" mode. It is clear that ATP remains bound to HPPK throughout the restraint simulation.

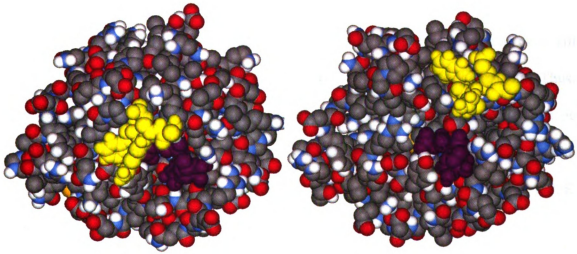
that cover part of ATP in the closed form move away, ATP might be able to get separated from HPPK or at least show substantial change in its conformation. That this did not happen suggested that there must be other interactions responsible for maintaining ATP bound to HPPK in the absence of HP.

To further the study, a 9 ns MD simulation was carried out starting from the end configuration of the above-mentioned step-wise restraining procedure. The restraints as listed in Table 4.1 were still maintained with equilibrium distances set to be equal to the ones measured from the apo HPPK crystal structure. This simulation was carried out to see if ATP would leave the active site automatically when HPPK was restrained to its apo-like state. In Figure 4.3 (left panel), a CPK view of the HPPK x-ray closed structure is shown for comparison with the CPK view of the protein structure after holding the restraints for 9 ns Figure 4.3 (right panel). Two points are worth noting. One is that from the viewpoint of the figure, it is clear from the right panel picture that there is nothing blocking the pathway for ATP to leave the binding pocket, while the path for ATP to leave in the left panel picture is clearly blocked by Ala86, Glu87, and the side-chains of Arg88 and Trp89. For clarity, these four residues are colored yellow. The other point is that there is minimal difference between the two pictures with respect to the ATP center of mass position relative to the protein core, though the conformations of ATP in the two figures are a bit different. (The conformation for the MD snapshot shown in the right panel is more compact in comparison with the x-ray structure shown in the left panel).

Reweighting MD of HPPK-ATP

It is clear from the restraint MD study that the ATP-HPPK interactions are

Figure 4.3. (left panel) CPK view of the starting MD simulation structure derived from the ternary complex. (right panel) CPK view of the MD structure after maintaining the restraints for 9 ns subsequent to the restraint opening protocol. ATP is colored purple. The yellow residues, 86-89, which cover part of ATP in the initial ternary complex, are mainly replaced by solvent interactions in the MD-opened structure.



sufficiently strong that on an MD time scale (here 9 ns) ATP remains bound, even though the interactions present in the closed structure that are certainly designed to hold ATP in position are absent in the MD-open form. To attempt to get the ATP separated from the HPPK, one possible and common scheme is similar to what was just used to open HPPK-ATP. However to do so either requires initial and final states as given by e.g. crystal structures, or if an obvious path does exist the ligand could be pushed out with restraints in an appropriate direction. However, here, an appropriate path is not clear from e.g. the end of the restrained open simulation. More importantly, picking a particular path will certainly introduce a prejudice into the simulation that may or may not be realistic. Thus, as noted in the methodology section in this chapter, when specific interactions can be identified targeted reweighting is an appropriate and more objective choice to accelerate the exploration of configuration space that here is characterized by the ATP-HPPK interaction.

It is clear that in both the closed form crystal structure and the restraint MD opened structure the two magnesium cations are very close to the carboxylate groups of two HPPK core residues, Asp95 and Asp97. For example, in the restraint MD opened structure the distances of Mg1 (associated with the α phosphate of ATP) to the closest carboxylate oxygens of Asp95 and Asp97 are 2.5 and 3.0 Å, and the corresponding distances for the Mg2 (associated with the γ phosphate of ATP) are 3.3 and 4.6 Å. Even though the side-chains of Arg82, Arg88 and Trp89 no longer block the path for ATP to leave its pocket, clearly there still are strong interactions with the protein; notably with the carboxylate groups of core residues Asp95 and Asp97, and these strong "salt bridges" are good candidates for important barriers restraining $ATP^{4-}(Mg^{2+})_2$. Unscreened charge

interactions that are on the scale of about 100 kcal/mol are essentially as strong as covalent bonds and will clearly prevent the movement of ATP on any normal MD time scale. Therefore, these salt bridges are good candidates to subject to the targeted reweighting scheme discussed in the methodology section in this chapter. The targeted atoms are the carbon and two oxygens of the carboxylate groups in these two residues and both magnesium cations; thus, the Mg1 and Mg2 to Asp95 and Asp97 electrostatic interactions are targeted. A uniform scaling value of $\lambda = 0.71$ (see the methodology section for details about the reason for picking this value) was used. The modified trajectory was started up from an endpoint configuration of the restrained MD simulation.

The targeting leads to ATP separating from HPPK on the nanosecond time scale. Six stable states were found along the separating process. The stable states are defined by examining the reweight energy – the electrostatic energy between the two magnesium cations and the two residues – along the trajectory. They correspond to plateaus in this energy. The initial modified energy of interaction V^* is about 300 kcal/mol while at the final state it is about 30 kcal/mol. This energy scale is so large that we were not successful in carrying out the reweighting required to obtain the correct Boltzmann populations along the trajectory. That would be necessary if a potential of mean force were the objective, but the modified trajectory still provides legitimate states along the separating process. Six representative structures from the trajectory were picked for display to show different states as ATP separates, and their superposed pictures are shown in Figure 4.4. The first structure (ATP in yellow) is the starting structure (endpoint conformation of the restrained MD simulation) and the sixth one (ATP in red) is where

Figure 4.4. $ATPMg_2$ for six different conformations as it separates from HPPK are shown with yellow, green, blue, purple, orange, and red indicating stages one through six. The gradual separation of ATP is evident as a combination of reorientational and translational motion.

ATP has moved out of the binding pocket. The middle four structures (the ATP is colored green, blue, purple, and orange, respectively) are stable states along the path of separation.

The gradual exit of ATP is evident from the Figure 4.4, and it is accomplished through a combination of mass center movement and rotation of the ATP, although it is difficult to see details of the motion from this figure. Therefore, the relevant hydrogen bonds and salt bridges were analyzed for the first 5 representative structures, and the results for structures 1, 3 and 5 are shown in Figure 4.5 that displays hydrogen bonds (left panels) and salt bridges (right panels) and are listed in Table 4.2 (hydrogen bonds) and Table 4.3 (salt bridges). Between the starting structure (structure 1) and structure 2, the mass center position of ATP does not change much, while its conformation does change significantly so that three new hydrogen bonds were formed and the two salt bridges involving one of the two Mg ions, Mg2, were broken. Between structure 2 and structure 3, ATP moved along its path away from the binding pocket somewhat; whereby three old hydrogen bonds and one salt bridge were broken, while one new hydrogen bond and one new salt bridge were formed. It is noteworthy that ATP has changed its conformation somewhat so that some old hydrogen bonds and salt bridges can be preserved with the change of its mass center position. This aspect of ATP motion is in evidence during the whole process. Throughout the movement from structure 3 to 4, though ATP underwent a large mass center position change, three of the four old hydrogen bonds were preserved, and the one that was broken was replaced by a new one with similar character. For the salt bridges, the one between Mg1 and Asp97 was lost, while the one between ATP and Arg110 changed a little bit. Again, though ATP moved a large step down its path (monitored by the energy of interaction with Asp95 and Asp97) of moving away from the

Figure 4.5. Details of the hydrogen bond and salt bridge interactions as ATP separates from HPPK. Left panels, from top to bottom, are hydrogen bonds (dotted lines) for stages 1, 3 and 5. Right panels, the corresponding salt bridges (solid lines). See Table 4.2 for hydrogen bond and Table 4.3 for salt bridge particulars.

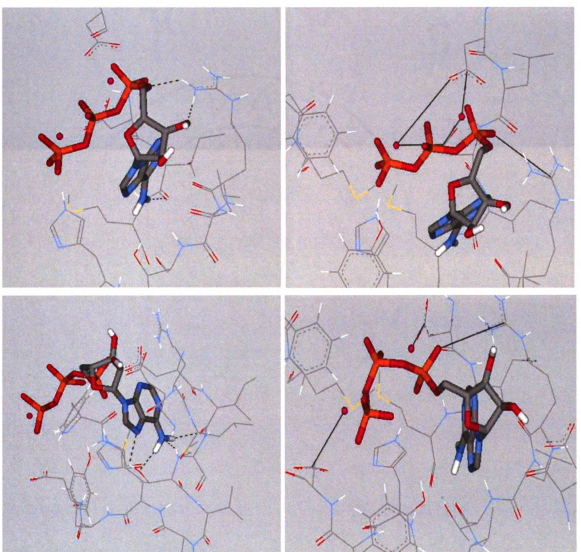


Figure 4.5 (continued)

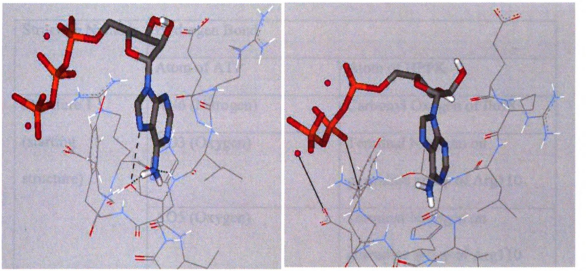


Table 4.2. The hydrogen bonds that are present in the stages of ATP separation from HPPK.

Structure Number	Hydrogen Bond	
	Atom of ATP	Atom of HPPK
Structure 1 (starting structure)	AN6 (Nitrogen)	Carbonyl Oxygen of Ile98
	AO3 (Oxygen)	Terminal Nitrogen on guanidine group of Arg110
	AO5 (Oxygen)	Terminal Nitrogen on guanidine group of Arg110
Structure 2	AN6 (Nitrogen)	Carbonyl Oxygen of Ile98
	AN6 (Nitrogen)	Carbonyl Oxygen of Thr112
	AN7 (Nitrogen)	Backbone Nitrogen of Thr112
	AO3 (Oxygen)	Terminal Nitrogen on guanidine group of Arg110
	AO5 (Oxygen)	Terminal Nitrogen on guanidine group of Arg110
	AO2PG (Oxygen)	Oxygen of the phenol group of Tyr116
Structure 3	AN6 (Nitrogen)	Carbonyl Oxygen of Ile98
	AN6 (Nitrogen)	Oxygen of the hydroxide of Thr112
	AN6 (Nitrogen)	Backbone Nitrogen of Thr112

Table 4.2. (continued)

	AN7 (Nitrogen)	Oxygen of the hydroxide of Thr112
Structure 4	AN6 (Nitrogen)	Carbonyl Oxygen of Met99
	AN6 (Nitrogen)	Oxygen of the hydroxide of Thr112
	AN6 (Nitrogen)	Backbone Nitrogen of Thr112
	AN7 (Nitrogen)	Oxygen of the hydroxide of Thr112
Structure 5	AN6 (Nitrogen)	Oxygen of the hydroxide of Thr112
	AN6 (Nitrogen)	Backbone Nitrogen of Thr112
	AN7 (Nitrogen)	Oxygen of the hydroxide of Thr112

Table 4.3. The salt bridges that are present in the stages of ATP separation from HPPK.

Structure Number	Salt Bridge[(a)]	
	Group 1 (positively charged)	Group 2 (negatively charged)
Structure 1 (starting structure)	Mg1	Carboxylate group of Asp95
	Mg1	Carboxylate group of Asp97
	Mg2	Carboxylate group of Asp95
	Mg2	Carboxylate group of Asp97
	Guanidine group of Arg110	α-phosphate group of ATP
Structure 2	Mg1	Carboxylate group of Asp95
	Mg1	Carboxylate group of Asp97
	Guanidine group of Arg110	α-phosphate group of ATP
Structure 3	Mg1	Carboxylate group of Asp97
	Mg2	Carboxylate group of Asp117
	Guanidine group of Arg110	α-phosphate group of ATP
Structure 4	Mg2	Carboxylate group of Asp117
	Guanidine group of Arg121	β-phosphate group of ATP

Table 4.3. (continued)

Structure 5	Mg2	Carboxylate group of Asp117
	Guanidine group of Arg110	β-phosphate group of ATP

(a) Mg1 (Mg2) refers to the one associated with the beta (gamma) phosphate of ATP

binding pocket, during the transition from structure 4 to structure 5, only one old hydrogen bond was broken, while all 3 other hydrogen bonds and all the salt bridges were preserved. Finally, the ATP-HPPK energetic interaction decreased to its smallest value and ATP moved to positions (see Figure 4.4 with ATP in red) where it is essentially out of the HPPK binding pocket. The successive and staged breaking of the original hydrogen bonds and salt bridges and formation of new ones during the process of separation show those salt bridges and hydrogen bonds are important in the simulated separation process. The ATP leaving process discussed in the above could be interesting for the following two reasons. First, when thinking about the ATP binding (versus leaving) process, the binding process might be similarl to the ATP leaving process simulated in the above. Then, for example, those salt bridges and hydrogen bonds, which seemed to be important in the simulated ATP leaving process, might play important roles in the binding mechanism of ATP. Second, the release process of the product ADP also might be similar to the simulated ATP leaving process, due to the similarity of ATP and ADP. Of course, the ATP leaving path shown in the above discussion was generated by targeting the interactions of the two magnesium cations with Asp95 and Asp97. When other interactions are targeted, it is possible that the generated ATP leaving path could be different.

We were not successful in reweighting back the trajectory to obtain the correct Boltzmann equilibrium sampling. Said otherwise, a potential of mean force was not obtained because the large range in energies of the modified potential surface along the trajectory did not permit the collection of sufficient data for the required accurate. This problem stems from the exponential dependence on the reweight potential. The only way

to address the problem would be to scale many more degrees of freedom and possibly obtain a smaller range of energy values. However, those more implicit degrees of freedom (in contrast with the explicit salt-bridge interactions that we identified) would be very hard to identify. Ultimately, the difficulty is the very high-dimensional surface that is being explored.

One thing notable is that a reweighting scheme does not directly provide dynamical information on the original surface. Once the potential surface has been modified, a time scale cannot be directly extracted from the simulation. So, one should not get the impression from the reweighting simulation that ATP can be released from HPPK on the ns time scale. According to transition state theory, the real time step can also be calculated by reweighting the simulation time step by the reweighting factor $\exp(\beta\Delta V)$, but this calculation requires that the energy difference ΔV to be zero for areas other than the energy basins, which is non-practical implementation requirement for complex systems.[108] To obtain dynamic information from simulation of complex system based on modified energy surface, Hamelberg and McCammon introduced.[108] an estimation method based on their special form of ΔV while Chen and Horing derived [109] a formula to calculate rate of transition out of the initial state for simulations using Langevin dynamics.

4. Concluding Remarks

In the work summarized in this chapter, a set of restraints were found that did lead to HPPK-ATP opening to apo-like conformations. One scenario that could have resulted during the restraint simulation was that at some point along the ten stages used to span

the closed to apo-like conformations ATP would have separated from HPPK. This was not found. Instead, as shown in Figure 4.2, the relative center of mass position between HPPK and ATP hardly changed, even though several residues present in the ternary structure that clearly block ATP from solvent exposure have moved away from the ATP to be essentially solvated in the last restraint window, as in the apo crystal structure (see Figure 4.3).

To investigate the interactions possibly responsible for holding ATP in its binding pocket, a 9 ns conventional MD run was carried out (with the restraints maintained at the last window values) starting from the endpoint of the HPPK-ATP restraint MD simulations and the simulation trajectory showed that ATP still remained bound to HPPK. Examination of the simulation data, illustrated by Figure 4.3, shows that there are several residues (Asp95 and Asp97) that still interact strongly with $ATP^{4-}(Mg^{2+})_2$. These core residues are clearly part of the binding pocket for ATP and provide a large, mainly electrostatic, interaction with $ATP^{4-}(Mg^{2+})_2$. In view of the specific salt-bridge-like interactions between these charged residues and the magnesium cations that in turn have a strong electrostatic interaction with ATP^{4-}, some of these interactions were the natural candidates to use in a targeted reweighting scheme. The targeted reweighting does lead to the separation of ATP from HPPK as shown in Figures 4.4 and 4.5 and detailed in Tables 4.2 and 4.3. The successive and staged breaking of the original hydrogen bonds and salt bridges and formation of new ones dominate the interactions for ATP binding. In terms of energetics, there are six stages of energetic interaction that can be identified over about 1 ns of simulation time and then, over the remaining 1 ns of the simulation, the ATP

remains separated from HPPK. Thus, these salt bridges play a key role in the binding mechanism.

Chapter 5. Some potential useful modifications of the HREM approach

In this chapter, several modifications of the HREM approach described in previous chapters are discussed that are under development and will be useful to other protein problems.

1. HTREM (Hamiltonian Temperature REM)

HTREM[110] is developed through reformulating HREM by also scaling the kinetic energy of the protein, thereby introducing an effective temperature $\overline{T}_\lambda = 1/k_B\overline{\beta}_\lambda$, where

$$\frac{1}{\overline{\beta}_\lambda} = \left(\frac{N_S}{N_P + N_S}\right)\frac{1}{\beta} + \left(\frac{N_P}{N_P + N_S}\right)\frac{1}{\beta\lambda}, \tag{5.1}$$

with N_p and N_s the number of degrees of freedom of protein and solvent, respectively, and λ is the same scaling factor used in the HREM approach, described repeatedly in previous chapters. It can be shown that by appropriate implementation, similar to the HREM approach, the Boltzmann equilibrium will result in the extended ensemble of the product of all the systems' ensembles for a sufficiently long trajectory for the HTREM approach. The main usage of the HTREM approach is that by using these effective temperatures, a pseudo melting curve can be constructed. That is, now each system's trajectory is of use and corresponds to a simulation at a series of effective temperatures. This method can be used for predicting the stability of dimerized proteins (fraction bound as a function of the effective temperature) and finds use to compare the stability of wild type and mutant forms[110].

2. HRTREM: (Hamiltonian Restraint TREM)

The HRTREM is formulated by introducing restraint potentials $V(\mu)$ to the HTREM Hamiltonian $H(\lambda)$, so that the resulting Hamiltonian for HRTREM is $H(\lambda,\mu)=H(\lambda)+V(\mu)$. The HRTREM is introduced in the hope of addressing two HTREM possible deficiencies when applied to constructing melting curves. First, it is probable for higher temperature systems to, having just exchanged, exchange back because they are not so different (in truly different conformations). Second, the sampling of separated monomers in a finite MD simulation box might be prejudiced toward less-than-separated distances. For the above mentioned purpose, in implementation, only the HRTREM systems at the highest temperature consist of $H(\lambda,\mu)$ while systems at successively lower temperatures only consist of $H(\lambda)$. $H(\lambda,\mu)$ systems and $H(\lambda)$ systems exchange at the highest temperature, and the chain of $H(\lambda)$ systems exchange down to the lowest temperature, building a cycle where $H(\lambda,\mu)$ systems may help to break the undesired correlation among the higher temperature $H(\lambda)$ systems and this break in correlation could propagate down to the lowest temperature system. The restraint will also tend to keep the monomers separated as desired at the higher temperatures. Of course, only the $H(\lambda)$ systems are used for melting curve construction. (Note: in the implementation of HRTREM in CUKMODY described in the Appendix Chapter, the HRTREM is implemented by a simple HRTREM routine and a more general (allowing variations in force constants and the form of restraints) HRTREMGEN routine. Details are given in the Appendix, Chapter.)

3. HTREM_RMS (HTREM_RootMeanSquare)

The HTREM_RMS approach is formulated similar to the HRTREM approach. It is introduced to help deal with the following problem. For a HTREM simulation, while the high effective temperature systems could be effective in overcoming barriers, they sometimes have the downside of tending to explore extensively in places that are not important in the true sample space. Therefore, in HTREM_RMS, $H(\lambda=1)$ is used for the experimental system (as described repeatedly in previous chapters, $\lambda=1$ indicating the true system Hamiltonian) and $H(\lambda,\mu)$ are used for all other systems with higher effective temperatures. (Note, the main difference from HRTREM is here: for HRTREM, only a few systems with the highest temperature are using $H(\lambda,\mu)$.) Then the restraints $V(\mu)$ can be set to help to keep the systems with high effective temperature close to some desired places in the sample space. For example, to keep a dimer from separating too readily at some higher temperature and thus permit better sampling of side chain interactions between monomers.) Of course, for HTREM_RMS, only the experimental system ($H(\lambda=1)$) contains ensemble information that can be used directly. (Note: for the routines described in the Appendix Chapter, both HTREM_RMS and its modification, HTREM_RMS_gradual, are for the HTREM_RMS implementation. The difference is that the later routine applies restraints more gradually. Details are provided in the Appendix Chapter.)

4. HREM_Ion

The conformations of highly-charged proteins might be influenced by the ionic atmosphere formed by explicit ions in the system. However, bound ions near charged residues could be quite sticky in an MD context due to the strong electrostatic salt-bridge-like interactions. A scheme[111] whose modifications here will be referred to as HREM_Ion was recently introduced to help increase the sampling efficiency of the ions around highly charged DNA. A base system with normal potential energy $V^{(0)}$ is coupled to a set of N other hypothetical systems, each with a unique potential $V^{(i)} = V^{(0)} + (\mu - 1)v^i$ (i=1,..N). Here v^i denotes the interactions of the ith group of ions, and system i interacts with a modified potential term μv^i. Different systems (with different groups of scaled ions) will have the possibility of exchanges that interchange their interaction strength scales, 1 and μ. Again, although the hypothetical systems could help to accelerate exploration, only the base system provides information that can be used directly. In our implementation oppositely charged ions need to be paired in each system in order to maintain overall neutrality in the Ewald energy and force evaluation. Any number of paired ions can be incoporated in $V^{(i)}$.

5. HREM_IonWater

The scheme for HREM_IonWater has a similar goal to HREM_Ion, though there is an important difference in the underlying concept. All the systems of HREM_IonWater approach are normal systems (all systems contain information that can be used directly).

For HREM_IonWater, instead of scaling groups of ions with μv^i as in HREM_Ion, a water molecule is considered as an ion whose charges on its three sites have been modified so that a $v^i(\mu)$ that is parameterized with water charges (q_H, q_O, q_H) instead of ion charges $(0, q_I, 0)$ can be introduced. For a certain exchange attempt between two systems with potentials $V^{(i)}$ (water i and ion j) and $V^{(j)}$ (water j and ion i), if the attempt is successful, then the ions and water molecules will exchange positions and then there is a large change in configuration. Of course, ions used to exchange with water must be chosen to have similar vdw parameters to water molecules to avoid bad vdw contacts.

Appendix. Description of implementation of HREM and several modifications

1. HREM

The HREM approach is implemented in the CUKMODY program based on the former DREM implementation[112]. The following is some description of how HREM is implemented in CUKMODY.

Initial Loading

When constructing the main MD simulation object of the "Solute_polar" class, a "setHChange" function is called in the constructor to load the parameters needed for HREM and store the parameters in a certain structure member of "Solute_polar" class. Then the "remMachine" class object, which will be used to conduct HREM procedures, is constructed and the HREM parameters are copied from "Solute_polar" class object to the "remMachine" class object into a structure for storing HREM parameters. Note that the correct parameters and configurations are loaded to each MPI process according to its rank. (Each MPI process will contain a certain configuration and the HREM parameters (which are characteristics of systems) will be exchanged according to HREM scheme when attempting exchange).

Exchange Attempt

The first thing to be noted is that there are two pthread threads being used, one thread will take care of normal MD simulation (in this pthread thread, the MD simulation may also be parallelized by using multiple openmp threads) and another thread will take care of the work involving making exchange attempts, including exchanging information

among MPI processes (again, note that each MPI process contains a particular configuration, and a set of HREM parameters according to which system it is at current time.)

For a certain MD step, after forces and energies are calculated normally, if the current time step is a multiple of the predefined number of steps for exchange attempt (say, 20 or 40), then a set of commands will be executed in the driver to conduct exchange attempt. First, functions will be called in each MPI process to calculate the energy for the current configuration of the process under its current HREM system parameters ("HREM_energy_one") and the energy for the configuration under the new HREM system parameters if the exchange attempt would be successful ("HREM_energy_another"). (Note: first, one exchange attempt is only made between two adjacent systems. Each system, with the exception of the first and the last each of which has one adjacent system, only have two adjacent systems and the information of these two adjacent systems is already stored as part of the system's HREM parameters structure. For system i, the exchange attempts are alternated between being made with system i+1 and with system i-1.Therefore, according to step number (by which the system on a certain MPI process can know which one of the two adjacent systems it would attempt to exchange with), in each MPI process the "HREM_energy_another" can be calculated independently. Second, the "HREM_energy_one" and "HREM_energy_ another" only include PME reciprocal space energy and the part of PME direct space energy that can not be canceled out for the later Metropolis Delta calculation. The definition of Metropolis Delta is provided in the methodology section of Chapter 2, Eq. 2.3.)

After obtaining "HREM_energy_one" and "HREM_energy_another" which are stored as part of HREM parameters structure, the thread responsible for exchange attempt on each MPI process will put its HREM parameters structure into an array created on the MPI process with its rank to be zero. Calculations of Metropolis Deltas for all the attempts are then made on that MPI process in its thread responsible for exchange attempt. If any exchange attempt is determined to be successful, corresponding changes will be made to the array containing all the HREM parameters structures. In the end, the thread responsible for exchange attempt on the MPI process with rank zero will pass the correct HREM parameters structure (again, note, the HREM parameters structure is the characteristic of a system) to the MPI process containing the corresponding configuration left intact. Then the new HREM parameters (could be the same ones if exchange attempt was determined to be not successful) will be used to calculate MD forces and energies starting from next MD step.

Code excerpt of the driver of HREM implementation of CUKMODY

The following is some excerpt of the c++ codes of the driver file for the HREM implementation of CUKMODY program. The excerpt is chosen to contain important lines relevant to making exchange attempt and some lines close to those important lines which may help understanding. (The sign "...", as usual, means some lines are omitted, so if no "..." between two lines listed then those two lines are also consecutive in the real driver file.)

```
...

////creating the simulation
```

```
  Simulation *ptr;
  // this will report the num threads used if omp used

  Solute_polar pt_ch2cl2(numProcessPerNode);
// "setHChange" function was called in the constructor for Solute_polar;
//In the function, the parameters needed for HREM were read in and
//temporarily stored in "pHChnge".(A substantial lines of codes here
//are used to take care of "exchange going down" or "exchange going up"
//details.)

  ptr = &pt_ch2cl2;

////set up index
  ptr->setReactCoordIndex(rank);
//At the start, since it is a new start (even for restart, we have re-
//ordered back) so configurations and systems are matched at this
//moment(say, configuration 0 must be in system 0).So we simply, at
//this moment, label each node according to the rank (which is an
//integer starting from 0) being assigned to the MPI process on it by
//the MPI library. And the corresponding configuration will be loaded
//latter accordingly.(The correct HREM parameters were already loaded
//in setHchange using the same trick as here.)

/////create the RemMachine
  RemMachine remMachine(STEP_REM,ptr,&HREMProb);
//Here,as usual data members of remMachine are initialized in the
//constructor and the HREM parameters were read into the remMachine
//object from the copy constructed in the previous "setHchange"
function
```

```
//in the Solute_polar constructor by calling "pSim->getPHChange()"
//function in the RemMachine constructor

  int index = rank;
// Here, the "index" is the copy of index for driver and the previous
//"ptr->setReactCoordIndex(rank)" initializes the copy of index
//essentially for the Solute_polar object pt_ch2cl2. The usage of the
//index and oldIndex variables here in the main function of driver is
//for later output.purposes.

  int oldIndex = index;

////initially scale q c6 c121 c122 c123
  ptr->scale_q_c6_c12_initially();
//Since we need to calculate some energies before any exchange attempt
//can be made, we need to set up the electrostatic and vdw constants to
//be in accordance with the corresponding (scaled) system.

  remMachine.turnOn();
//Here, we spawn off a second pthread thread to take care of the things
//related with exchange (including sending and receiving information,
//and arbitrating, which is only done on the node with rank=0). The
//original pthread thread will still run normally to execute codes for
//MD calculation.

//The following three lines prepare for the replica exchange attempt at
//the first step, so that later we can "setHREMEnergyOne"
  ptr->compute_energy();
  ptr->compute_energy_solute();
```

```
  ptr->evaluateTemperature();

////The following is the MD simulation loop
  while(i<total_time_steps)
    {

      ptr->initialize();
      if(is_solvent_polarizable)
      ptr->drude_dipole(10,0.01);
      ptr->build_list(rl,i);
      ptr->compute_PME(i);
      ptr->compute_force3(rc);
      ptr->compute_force_solute(rc);
      ptr->compute_cobond(i);
      ptr->compute_cobond_solute(i);
      ptr->compute_angle();
      ptr->compute_angle_solute(i);
      ptr->compute_dihang_solute(i);
      ptr->compute_force_internal_solute(rc);
      ptr->compute_inter_solute(rc);
//The above lines for MD force and energy calculation

////The following in "if" brackets are for exchange attempt
      if(i% STEP_REM == 0)
      {
       ptr->setHREMEnergyOne();
       remMachine.applyHChangeForCalc();
       ptr->compute_PME_another();
       remMachine.setHREMEnergyAnother();
```

```
        remMachine.setbackHChange();
      //The previous five lines set up the HREM energy before exchange
//(HREM energy one) and the HREM energy after exchange if exchange
//would be successful (HREM energy another) for the latter calculation
//in the remMachine object for Metropolis Delta. The HREM energy
//comprises the whole reciprocal space energy (calculated by pme method)
//and part of the direct space energy.(The not-included part of the
//direct space energy will be canceled out in the Delta calculation.)
//Due to the requirement of
//pme implementation, to calculate the reciprocal space part of "HREM
//energy another" we need to use "remMachine.applyHChangeForCalc"
//function to set the electrostatic and vdw constants to be proper
//values according to the system which the current system is attempting
//to exchange with. In terms of direct space energy, since it is
//calculated pairwise, we only need to include the energies involving
//atoms with their electrostatic and vdw constants scaled (others will
//be canceled out). Those energies are only calculated in the
//following three functions: "compute_force_solute",
//compute_force_internal_solute", "compute_inter_solute". During the
//execution of the above-mentioned three functions,the
//"vdwE_scale","vdwE_scale_2","elecE_scale" and "elecE_scale_2" were
//obtained. "_scale" means only one of the two atoms in the pair is to
//be scaled and "_scale_2" means both of the two atoms in the pair are
//to be scaled. Therefore, the included direct part of "HREMEnergyOne"
//is simply the sum of the above four terms, while the calculation of
//the included direct part of "HREMEnergyAnother" could be done by
//multiplying the corresponding ratios calculated by dividing the new
//scaling factor by the old factor (see "comments in "Rem.cpp" for
//function "HChange::setHREMEnergyAnother").Since we have changed
```

```
//electrostatic and vdw constants for calculating the
//"HREMEnergyAnother" while we are not sure at this moment whether the
//exchange will be successful; we need to change back those constants
//to original values through the execution of "setbackHChange" function.

       remMachine.updateInfo();
      //The function puts the previously calculated information stored
//in the "Hchange" class object, which is a member of the "remMachine"
//class object, into a struct "info" which is a member of "pData",
//which itself is a member of the "RemMachine" class object.

       remMachine.arbitrate(i);
      //In this function, the thread spawned off for sending and
//receiving messages and doing arbitrating is awakened. The thread will
//do things (see after " varForArbitrator.wait();" inside "Rem.cpp"
//file) related to exchange attempts, including calling functions to
//use the calculated "HREMEnergyOne" and "HREMEnergyAnother" to
//calculate delta and then use the delta to judge whether the
//exchange attempt is successful or not.

       remMachine.applyHChange(i);
      //After the exchange, the correct window information is stored in
//the "HChange" class object and is used here to scale the
//electrostatic and vdw constants of the atoms which are in the
//initially specified group including atoms that will be scaled. (Note:
//When the exchange attempt is rejected, then the ratio used to scale
is actually 1.)

       remMachine.reverseIsUp();
```

```
//As is usual in a REM approach, there are two patterns of exchange:
//referred to as "up" and "down" (see Chapter, xxx) For HREM, the
//information involved is a bit different. So we need to tell which
//pattern it is for a given exchange attempt. For the first exchange
//attempt, the "down" pattern is in use.
      }
//end of the "if" brackets for exchange attempt

      oldIndex = index;
      index = remMachine.getIndex();
      if(i%STEP_SYSLOG==0)
      {
        static int oldIndex = index;
        if(oldIndex != index)
          {
            sysLogStream << cfgIndexEnd;//put the index end sign
            sysLogStream << cfgIndex(index);//put the index sign
            oldIndex = index;//update index
          }
        sysLogStream << "\ntime_step " << i <<endl;
      }
      //The above lines are used to set marks for the end of the
//application of a previous system and the number of the system which
//will be applied next. "oldIndex" and "index" are local variables of
//main function and are used in the "if(oldIndex != index)" to check
//whether some changes have occurred after exchange through updating
//index with the line "index = remMachine.getIndex();".Note that each
//node has its own copy of driver so, of course, each node has its own
//copies of "oldIndex" and "index". In other words, the values for
```

```
//"oldIndex" and "index" are different for different nodes.

...
}
//end of "while" loop for MD simulation steps
```

2. The implementation of some possibly useful modifications based on HREM

In chapter 5, several possibly useful approaches based on modifying the HREM approach were described. A simple description of the implementation of those modifications (focusing on the implementation difference) is listed in the following.

HTREM

The main implementation modification made on HREM to get HTREM is as follows. In the driver file, the function "scale_temperature_HTREM" is called. The function will assign an effective temperature (which is the temperature for the system calculated as a weighted average as defined in Chapter 5, eq 5.1) to every node based on the lambda value (scaling factor for the system applied) specified in an input file,"rewFactor.txt".

HRTREM

The HRTREM implementation is modified based on the implementation of HTREM and has the exactly same feature as HTREM in terms of "effective temperature". The addition for HRTREM is that some restraints specified in the input file "rewFactor.txt" are applied to the systems that have "mu"=1 ("mu" is a member of Solute_polar class and values for different systems are specified in "rewFactor.txt" file) at

every step. Note, for all the systems with "mu"=1, the same set of restraints are used (namely, same pairs of atoms, same equilibrium distances, same force constants).For the "HREMEnergyOne" and "HREMEnergyAnother" used for Metropolis Delta calculation, the corresponding restraint energies are included. The nature of the restraint is easy to change in the "applyRestraint" function of Solute_polar class, which is called in the driver.

HRTREMGEN

The HRTREMGEN, as can be told from its name, is an extension of HRTREM. The added flexibility for HRTREMGEN includes the following. First, now different systems can have different equilibrium distances and force constants. Second, there is a switch implemented between one-sided harmonic restraint (with switch value 1) and harmonic restraint (with switch value 2). If needed, more switches can be easily added. (Note that for all systems of HRTREMGEN, the restrained pairs of atoms still have to be the same.)

HTREM_RMS

The HTREM_RMS is modified based on HTREM and has the exactly same feature as HTREM in terms of having effective temperatures. The input file "rewFactor.txt" containing the Hamiltonian scaling factors for different systems is exactly the same as the one needed for HTREM, but some additional information needed for HTREM_RMS is specified in another input file "restraint_rms.txt". In file "restraint_rms.txt", restraint atom groups, and force constants are specified. Note that the final column of the restraints part in the file repeats the system number and this number

will be read into the variable "actualIndex". There is no real difference between "windowIndex" and "actualIndex" and the only reason for both of them to be used is to keep the structure of some existing routine for implementation simplicity. The "lambda" values specified next in "restraint_rms.txt" are used to set the 2-d array (the array constructed from substantiating the global pointer "pointerToDistMap_global") containing the equilibrium distances for all pairs of restraints and for all systems. The equilibrium distances are calculated in the "setHchange" function using two protein structures (one specified in "coord_gbp_initial.dat" file and another specified in "coord_gbp_final.dat" file) by applying simple linear interpolation on the distances between restraint atom group mass centers from the two structures. The "lambda" values are the factors for linear interpolation conducted for different systems. Note: first, although "lambda" values are often set to be the same for all systems (say, one for each system, meaning that the distances of the structure contained in "coord_gbp_final.dat" file are used as equilibrium distances for all the systems), they can be set to be different. Second, although the lambda for the first system may be set as some real value, no constraints are applied in the system due to the fact that the force constant for the system is always set to zero. Third, the two structures specified in files "coord_gbp_initial.dat" and "coord_gbp_final.dat" are only used to simplify the initial setting of the 2-d array for equilibrium distances; in other words, the program could have been implemented without using the two structures but loading in a directly specified matrix of equilibrium distances, though a large input matrix needs to be made manually then.

HTREM_RMS_gradual

This is a variation of the HTREM_RMS. The difference here is that the target equilibrium distances stored in the 2-d array ("pointerToDistMap_global") are only used as equilibrium distances of the restraints at the last time step. For other time steps, the linear interpolation (using a factor equals to the quotient of current time step over total time step) results of the distances from starting structure and the target equilibrium distances stored in the 2-d array, are used as equilibrium distances of the restraints. Note that this method was implemented to make the change more gradual so that the exchange between the normal system and the system with applied restraints might be enhanced. Normally, using this method, a restart is not expected.

HREM_Ion

The implementation of HREM_Ion is modified based on the implementation of HREM. The basic idea for HREM_Ion, as also described in Chapter 5, is that, with the exception of the first system (the normal system), each system has a neutral group of ions (the number of ions in the group must be same for each system, and no two systems can have the same ion in their groups) on which a uniform lambda value is used to scale their electrostatic interactions (vdw interactions are not scaled but can be easily added in). Note that each system has the same ions in it, the difference is which ions are in the group to be scaled by the lambda (of course, no ions are scaled for the first system). The ions needed to be scaled for the systems are loaded from file "HREM.txt". Once we think those different groups of ions to be the HREM characteristic of systems, the implementation is almost the same. One more difference noticeable is that, for HREM_Ion, a roulette kind of method is used to mix systems with the exception of the

first system, in this implementation. That is done to not privilege any scaled parameter system. So, all systems after the first do not have any useful sampling information.

HREM_IonWater

The implementation of the HREM_IonWater is based on the HREM_Ion implementation. However, as also pointed out in Chapter 5, there is an important difference in the underlying concept. All systems for HREM_IonWater are normal systems. For HREM_IonWater, a 2-d array (substantiated "pWaterTable") is loaded according to the "HREM.txt" file. The array contains places at which there are additional waters in the "list of ions and additional waters" for the systems (the list for the first system is stored in "parameters.dat"). Because the same "coord_gbp.dat" and "parameters.dat" (set for the first system) are loaded for all the systems in the constructor of "Solute_polar", in function "set_c6_c12_initially", according to the "pWaterTable" table, the systems are permuted for where the added waters (those that can turn into ions) are. Similarly, for conducting an exchange, it is the system index that needs to be exchanged so that the old and new indices can be used to check the "pWaterTable" for the old and new water places to be permuted in the "applyHChange" function. (Note that since ions are assumed to move swiftly, the term "neighbor" is only perfunctory here.) For implementation simplicity, the ions are considered to be three-site molecules with appropriate charges, and the vdw radii must be at least very close to a that of a water molecule, to lead to reasonable exchange probabilities. The vdw parameters for those ions are reloaded in function "set_c6_c12_initially" from "ionVdw.txt".

References

(1) Alder, B. J.; Wainwright, T. E. *Journal of Chemical Physics* **1959**, *31*, 459.

(2) Bernal, J. D. *Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences* **1964**, *280*, 299.

(3) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.

(4) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; III, T. E. C.; Wang, J.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsue, V.; Gohlke, H.; Radmer, R.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, C.; Weiner, P.; Kollman, P. A. AMBER7; University of California: San Francisco, 2002.

(5) Mackerell, A. D.; Feig, M.; Brooks, C. L. *Journal of Computational Chemistry* **2004**, *25*, 1400.

(6) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P. *Biomolecular simulation: the GROMOS96 manual and user guide*; Vdf hochschulverlag AG an der ETH: Zürich, 1996.

(7) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles*; McGraw-Hill: New York, 1981.

(8) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(9) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic: San Diego, 1996.

(10) Berendsen, H. H. C.; Postma, J. P. M.; Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(11) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523.

(12) Andersen, H. C. *Journal of Chemical Physics* **1980**, *72*, 2384.

(13) Karasawa, N.; Goddard, W. A. *Journal of Physical Chemistry* **1989**, *93*, 7320.

(14) York, D. M.; Darden, T. A.; Pedersen, L. G. *Journal of Chemical Physics* **1993**, *99*, 8345.

(15) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *Journal of Chemical Physics* **1995**, *103*, 8577.

(16) McCammon, A.; Harvey, S. C. *Dynamics of proteins and nucleic acids*; Cambridge University Press: Cambridge, 1987.

(17) Swendsen, R. H.; Wang, J. S. *Physical Review Letters* **1986**, *57*, 2607.

(18) Geyer, C. J. Markov Chain Monte Carlo Maximum Likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* Keramidas, E. M., Ed.; Interface Foundation: Fairfax Sta tion,.1991.

(19) Sugita, Y.; Okamoto, Y. *Chemical Physics Letters* **1999**, *314*, 141.

(20) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.

(21) Wang, J. S.; Swendsen, R. H. *Progress of Theoretical Physics Supplement* **2005**, 317.

(22) Hukushima, K.; Nemoto, K. *Journal of the Physical Society of Japan* **1996**, *65*, 1604.

(23) Lei, H. X.; Duan, Y. *Current Opinion in Structural Biology* **2007**, *17*, 187.

(24) Sugita, Y. *Front Biosci.* **2009**, *14*, 1292

(25) Fukunishi, H.; Watanabe, O.; Takada, S. *Journal of Chemical Physics* **2002**, *116*, 9058.

(26) Predescu, C.; Predescu, M.; Ciobanu, C. V. *Journal of Physical Chemistry B* **2005**, *109*, 4189.

(27) Liu, Z. H.; Berne, B. J. *Journal of Chemical Physics* **1993**, *99*, 6071.

(28) Wenzel, W.; Hamacher, K. *Physical Review Letters* **1999**, *82*, 3003.

(29) Hornak, V.; Simmerling, C. *Proteins-Structure Function and Genetics* **2003**, *51*, 577.

(30) Cukier, R. I.; Morillo, M. *Journal of Chemical Physics* **2005**, *123*.

(31) Shen, T.; Hamelberg, D. *J Chem Phys* **2008**, *129*, 034103.

(32) Hansmann, U. H. E. *Eur. Phys. J. B* **1999**, *12*, 607.

(33) Hamelberg, D.; Mongan, J.; McCammon, J. A. *Journal of Chemical Physics* **2004**, *120*, 11919.

(34) Hamelberg, D.; McCammon, J. A. *Journal of the American Chemical Society* **2005**, *127*, 13778.

(35) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

(36) Torrie, G. M.; Valleau, J. P. *Chemical Physics Letters* **1974**, *28*, 578.

(37) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *Journal of Computational Chemistry* **1992**, *13*, 1011.

(38) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *Journal of Chemical Theory and Computation* **2007**, *3*, 26.

(39) Souaille, M.; Roux, B. *Computer Physics Communications* **2001**, *135*, 40.

(40) Jolliffe, I. T. *Principal Component Analysis*, Second ed.; Springer Science: New York, 2004.

(41) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *Journal of Physical Chemistry* **1996**, *100*, 2567.

(42) García, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696.

(43) García, A. E.; Blumenfeld, R.; Hummer, G.; Krumhansl, J. A. *Physica D* **1997**, *107*, 225.

(44) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Structure, Function, and Genetics* **1993**, *17*, 412.

(45) Cox, T. F.; Cox, M. A. A. *Multidimensional scaling*, 2nd ed. ed.; Chapman & Hall: Boca Raton, 2001.

(46) T. Hastie, R. T. a. J. F. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*; Springer: Berlin, 2001.

(47) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angewandte Chemie-International Edition* **1999**, *38*, 236.

(48) Scheffe, H. *Biometrika* **1953**, *40*, 87.

(49) Johnson, S. C. *Psychometrika* **1967**, *32*, 241.

(50) Hughes, J.; Smith, T. W.; Kosterlitz, H. W.; Fothergill, L. A.; Morgan, B. A.; Morris, H. R. *Nature* **1975**, *258*, 577.

(51) Schiller, P. W. *Conformational analysis of enkephalin and conformation-activity relationships*; Academic Press: Orlando, 1984; Vol. 6.

(52) Spanagel, R.; Herz, A.; Shippenberg, T. S. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 2046.

(53) Graham, W. H.; Carter, E. S.; Hicks, R. P. *Biopolymers* **1992**, *32*, 1755.

(54) Higashijima, T.; Kobayashi, J.; Nagai, U.; Miyazawa, T. *European Journal of Biochemistry* **1979**, *97*, 43.

(55) Marcotte, I.; Separovic, F.; Auger, M.; Gagne, S. M. *Biophysical Journal* **2004**, *86*, 1587.

(56) Surewicz, W. K.; Mantsch, H. H. *Biochemical and Biophysical Research Communications* **1988**, *150*, 245.

(57) Khaled, M. A.; Long, M. M.; Thompson, W. D.; Bradley, R. J.; Brown, G. B.; Urry, D. W. *Biochemical and Biophysical Research Communications* **1977**, *76*, 224.

(58) Takeuchi, H.; Ohtsuka, Y.; Harada, I. *Journal of the American Chemical Society* **1992**, *114*, 5321.

(59) Spirtes, M. A.; Schwartz, R. W.; Mattice, W. L.; Coy, D. H. *Biochemical and Biophysical Research Communications* **1978**, *81*, 602.

(60) DAlagni, M.; Delfini, M.; DiNola, A.; Eisenberg, M.; Paci, M.; Roda, L. G.; Veglia, G. *European Journal of Biochemistry* **1996**, *240*, 540.

(61) Hansmann, U. H. E.; Okamoto, Y. *Journal of Computational Chemistry* **1997**, *18*, 920.

(62) Hansmann, U. H. E.; Okamoto, Y.; Eisenmenger, F. *Chemical Physics Letters* **1996**, *259*, 321.

(63) vanderSpoel, D.; Berendsen, H. J. C. *Biophysical Journal* **1997**, *72*, 2032.

(64) Aburi, M.; Smith, P. E. *Biopolymers* **2002**, *64*, 177.

(65) Nielsen, B. G.; Jensen, M. O.; Bohr, H. G. *Biopolymers* **2003**, *71*, 577.

(66) Shen, M. Y.; Freed, K. F. *Biophysical Journal* **2002**, *82*, 1791.

(67) Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F. *Journal of Physical Chemistry B* **2003**, *107*, 1685.

(68) Karvounis, G.; Nerukh, D.; Glen, R. C. *Journal of Chemical Physics* **2004**, *121*, 4925.

(69) Sanbonmatsu, K. Y.; Garcia, A. E. *Proteins-Structure Function and Genetics* **2002**, *46*, 225.

(70) Spadaccini, R.; Temussi, P. A. *Cellular and Molecular Life Sciences* **2001**, *58*, 1572.

(71) Murray, B. E. *Advances in Internal Medicine, Vol 42* **1997**, *42*, 339.

(72) Blakley, R. L., Benkovic, S. J. *Chemsitry and Biochemistry of folates*; John Wiley & Sons, Inc.: New York, 1984; Vol. 1.

(73) Hennig, M.; Dale, G. E.; D'Arcy, A.; Danel, F.; Fischer, S.; Gray, C. P.; Jolidon, S.; Muller, F.; Page, M. G. P.; Pattison, P.; Oefner, C. *Journal of Molecular Biology* **1999**, *287*, 211.

(74) Xiao, B.; Shi, G. B.; Chen, X.; Yan, H. G.; Ji, X. H. *Structure* **1999**, *7*, 489.

(75) Blaszczyk, J.; Shi, G.; Yan, H.; Ji, X. *Structure* **2000**, *8*, 1049.

(76) Xiao, B.; Shi, G. B.; Gao, J. H.; Blaszczyk, J.; Liu, Q.; Ji, X. H.; Yan, H. G. *Journal of Biological Chemistry* **2001**, *276*, 40274.

(77) Li, Y.; Gong, Y. C.; Shi, G. B.; Blaszczyk, J.; Ji, X. H.; Yan, H. G. *Biochemistry* **2002**, *41*, 8777.

(78) Blaszczyk, J.; Shi, G. B.; Li, Y.; Yan, H. G.; Ji, X. H. *Structure* **2004**, *12*, 467.

(79) Li, G. Y.; Felczak, K.; Shi, G. B.; Yan, H. G. *Biochemistry* **2006**, *45*, 12573.

(80) Blaszczyk, J.; Li, Y.; Cherry, S.; Alexandratos, J.; Wu, Y.; Shaw, G.; Tropea, J. E.; Waugh, D. S.; Yan, H. G.; Ji, X. H. *Acta Crystallographica Section D-Biological Crystallography* **2007**, *63*, 1169.

(81) Lescop, E.; Lu, Z. W.; Liu, Q.; Xu, H. M.; Li, G. Y.; Xia, B.; Yan, H. G.; Jin, C. W. *Biochemistry* **2009**, *48*, 302.

(82) Su, L.; Cukier, R. I. *Journal of Physical Chemistry A* **2009**, *113*, 2025.

(83) Lou, H. F.; Cukier, R. I. *J. Phys. Chem. B* **2006**, *110*, 12796.

(84) Su, L.; Cukier, R. I. *Journal of Physical Chemistry B* **2007**, *111*, 12310.

(85) Lou, H. Analyzer; 1.0 ed. East Lansing, 2005.

(86) Amadei, A.; Ceruso, M. A.; Di Nola, A. *Proteins-Structure Function and Genetics* **1999**, *36*, 419.

(87) Hess, B. *Physical Review E* **2002**, *65*, 031910.

(88) Romo, T. D.; Clarage, J. B.; Sorensen, D. C.; Phillips, G. N. *Proteins-Structure Function and Genetics* **1995**, *22*, 311.

(89) Lou, H. F.; Cukier, R. I. *Journal of Physical Chemistry B* **2006**, *110*, 24121.

(90) Korn, A. P.; Rose, D. R. *Protein Engineering* **1994**, *7*, 961.

(91) Brubaker, R. R. *Clinical Microbiology Reviews* **1991**, *4*, 309.

(92) Kumar, S.; Ma, B. Y.; Tsai, C. J.; Sinha, N.; Nussinov, R. *Protein Science* **2000**, *9*, 10.

(93) James, L. C.; Tawfik, D. S. *Trends in Biochemical Sciences* **2003**, *28*, 361.

(94) Goh, C. S.; Milburn, D.; Gerstein, M. *Current Opinion in Structural Biology* **2004**, *14*, 104.

(95) Bahar, I.; Chennubhotla, C.; Tobi, D. *Current Opinion in Structural Biology* **2007**, *17*, 633.

(96) Ma, B. Y.; Kumar, S.; Tsai, C. J.; Nussinov, R. *Protein Engineering* **1999**, *12*, 713.

(97) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins-Structure Function and Genetics* **1995**, *21*, 167.

(98) Dill, K. A.; Chan, H. S. *Nature Structural Biology* **1997**, *4*, 10.

(99) Taniuchi, H.; Anfinsen, C. B. *Journal of Biological Chemistry* **1966**, *241*, 4366.

(100) Markley, J. L.; Williams, M. N.; Jardetzky, O. *Proceedings of the National Academy of Sciences of the United States of America* **1970**, *65*, 645.

(101) Chazin, W. J.; Kordel, J.; Drakenberg, T.; Thulin, E.; Brodin, P.; Grundstrom, T.; Forsen, S. *Proceedings of the National Academy of Sciences of the United States of America* **1989**, *86*, 2195.

(102) Zhang, H. J.; Sheng, X. R.; Niu, W. D.; Pan, X. M.; Zhou, J. M. *Journal of Biological Chemistry* **1998**, *273*, 7448.

(103) Wilson, M. A.; Brunger, A. T. *Journal of Molecular Biology* **2000**, *301*, 1237.

(104) Wang, P. M.; Izatt, R. M.; Oscarson, J. L.; Gillespie, S. E. *Journal of Physical Chemistry* **1996**, *100*, 9556.

(105) Cukier, R. I. *Biochimica Et Biophysica Acta-Bioenergetics* **2005**, *1706*, 134.

(106) Bermingham, A.; Bottomley, J. R.; Primrose, W. U.; Derrick, J. P. *Journal of Biological Chemistry* **2000**, *275*, 17962.

(107) Figueirido, F.; Levy, R. M.; Zhou, R. H.; Berne, B. J. *Journal of Chemical Physics* **1997**, *106*, 9835.

(108) Hamelberg, D.; Shen, T.; McCammon, J. A. *Journal of Chemical Physics* **2005**, *122*.

(109) Chen, L. Y.; Horing, N. J. M. *Journal of Chemical Physics* **2007**, *126*.

(110) Li Su, R. I. C. *J. Phys. Chem. B* **2009**, *113*, 9595.

(111) Min, D. H.; Li, H. Z.; Li, G. H.; Berg, B. A.; Fenley, M. O.; Yang, W. *Chemical Physics Letters* **2008**, *454*, 391.

(112) Lou, H. Development of a Molecular Dynamics based method to accelerate sampling of large domain motions in proteins: Applications to Adenylate Kinase, Michigan State University, **2006**.