DEVELOPMENT AND APPLICATION OF THE MULTI-OVERLAP MOLECULAR DYNAMICS METHODS

A Thesis
Presented to the Department of Functional Molecular Science
School of Physical Sciences
The Graduate University for Advanced Studies
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science

by

Satoru Itoh

March 2005

Acknowledgments

My most heartfelt thanks go to my thesis advisor, Professor Yuko Okamoto, for his patient guidance and constant encouragement. I wish to express my gratitude to all the members of the Okamoto Group for stimulating discussions and encouragement. I am grateful to the members of the IMS theory groups for their generous support. I wish to thank the faculty and staff members of the Graduate University for Advanced Studies for their kindness. Finally, I am most thankful to my parents Yasuhide and Aiko Itoh for their constant encouragement and support.

Contents

Chapter 1

General Introduction

Proteins carry out various biological functions *in vivo*. In order for proteins to display unique functions, it is important that the proteins have unique three-dimensional structures (tertiary structures). The native three-dimensional structures seem to be affected by the environment in the cell, which is a very complex system. Anfinsen and co-workers, however, showed experimentally that the native structures of proteins are decided by their amino-acid sequence information (Anfinsen's dogma) [1]. Specifically, they substantiated such a dogma as follows. They denatured completely bovine pancreatic ribonuclease *in vitro* by using denaturants. The proteins lost their enzymatic activity completely and had random-coil conformations. From such unfolded states, the bovine pancreatic ribonuclease refolded back into the native structure and recovered the enzymatic activity when the denaturants were removed. Anfinsen's dogma implies that we just have to deal with the amino-acid sequence information and surrounding solvent, not other molecules which exist in the cell, when we study the protein folding problem. In other words, we can reduce the problem of a protein in the cell to that of a single protein in solution. With such a simplification, it is still very difficult to study the protein folding problem with computer simulations due to Levinthal's paradox [2],[3]: it is essentially impossible to find the native structure among an astronomically large number of metastable conformations. However, this is an apparent paradox because we neglect the fact that protein structures take on various potential energy values. In other words, the thermodynamic consideration is missing in that there exist much less number of important conformations to consider at room temperature. Thus, the protein folding problem can be understood in the thermodynamic framework. Namely, the native structure corresponds to the free-energy global-minimum state and random-coil states rapidly make transitions to the native state. By utilizing computer simulations based on statistical mechanics [4]-[15], we are able to study the protein folding problem theoretically.

In order to understand the protein folding, it is essential that the detailed free-energy landscape of the protein system is obtained. By analyzing the free-energy landscape, we can find the folding pathways and the stability of any structures of the protein. Furthermore, the transition state between two specific stable states can also be discovered. Exploring the transition state, we can gain information about state transitions. From a point of view of molecular modeling or drug design, moreover, it is also very important that the transition state is found. Accordingly, many efforts are devoted to obtain the detailed free-energy landscape by computer simulations.

A canonical-ensemble simulation [4]-[9] is widely used as a conventional method for computer simulations. In the canonical ensemble at a fixed temperature, the probability distribution of the potential energy is given by the product of the density of states and the Boltzmann weight factor, and we have a bell-shaped probability distribution of the potential energy. However, this simulation method is not suitable to be applied to complex systems such as proteins. Because such complex systems have many local-minimum free-energy states, canonical-ensemble simulations tend to get trapped at the local-minimum states. At low temperatures, in particular, the usual canonical-ensemble simulations cannot realize efficient sampling in the configurational space. This is because in canonical simulations energy fluctuations are small at a low temperature and energy barriers cannot be overcome. Therefore, if we employ the usual canonical-ensemble method in complex systems, we may estimate inaccurately the free-energy landscape in the complex systems. To overcome the difficulties, the generalized-ensemble algorithms have been proposed (for a review, see Ref. [10]).

The multicanonical algorithm [11]-[14] is perhaps one of the most well-known methods among the generalized-ensemble algorithms. In the multicanonical ensemble, the probability distribution of the potential energy is expressed by the product of the density of states and a non-Boltzmann weight factor, which we refer to as the multicanonical weight factor, and we have a flat probability distribution of the potential energy. Therefore, multicanonical-ensemble simulations realize a free random walk in the potential-energy space and overcome energy barriers. By such efficient sampling in the configurational space, the multicanonical-ensemble simulations are able to give the broad free-energy landscape in comparison with conventional canonical simulations. Furthermore, because the multicanonical simulations do not get trapped in local-minimum states, we need much less simulation time to get an accurate free-energy landscape than conventional canonical simulations. Therefore, the application of the multicanonical algorithm to the protein folding was proposed [16]. Since then there have been many works based no this method and its variants in protein and related systems (for reviews, see Refs.[17],[18]). This method aims at achieving a wide range sampling in the configurational space. However, because of the very nature of the algorithm, it is difficult to focus on specific configurations. Consequently, the free-energy landscape around or among specific configurations of interest may be incorrectly estimated in multicanonical-ensemble simulations.

To understand protein folding, as discussed previously, we must investigate the stability of specific configurations and the transition state between two specific configurations. Accordingly, the detailed free-energy landscape in the neighborhood of specific configurations is necessary. Recently, a new algorithm, which is a generalization of the multicanonical algorithm and is referred to as the multi-overlap algorithm [15], was proposed to focus on specific configurations and overcome potential energy barriers, where an overlap of a configuration is a measure of structural similarity to a reference configuration. In the multi-overlap ensemble, the probability distribution is expressed by the product of the density of states and a non-Boltzmann weight factor, which we refer to as the multi-overlap weight factor, and we have a flat probability distribution in the overlap space. Consequently, the multi-overlap ensemble realizes a random walk in the overlap space and efficiently samples the conformational space, and we can obtain the detailed free-energy landscape in the neighborhood of specific configurations.

A Monte Carlo (MC) version of this algorithm was proposed in Ref. [15]. In general, for linear molecules such as proteins, MC simulations mostly use the dihedral angles, not Cartesian coordinates, in order to maintain the covalent geometry of such linear molecules. A small update of a single dihedral angle can then result in a large motion of the linear molecule, and the trial MC step will be almost always rejected. Therefore, in many particle systems such as proteins in solution, MC algorithm would sample inefficiently the conformational space, and it is difficult to estimate correctly the free-energy landscape.

To avoid such problems in the MC simulations, the molecular dynamics (MD) algorithm is often employed. For instance, the MD version of multicanonical algorithm was developed in Refs. [13],[14]. In this thesis we propose an MD version of the multi-overlap method. This multi-overlap MD method realizes a random walk in the dihedral-angle distance space and does not get trapped in local-minimum states. Furthermore, this method is able to sample efficiently the conformational space so as to focus on specific configurations. Accordingly, we obtain the detailed free-energy landscape among the specific configurations. We apply this multi-overlap MD method to Met-enkephalin in vaccum and check the effectiveness of the method by comparing the results with those of the conventional canonical MD method and the multicanonical MD method. Moreover, from the detailed free-energy landscape obtained from the results of the multi-overlap MD simulation, we predict a transition pathway between two specific configurations of Met-enkephalin.

In Chapter 2 we present the conventional canonical and multicanonical MD algorithms and propose the multi-overlap MD algorithms. We also introduce a jackknife method which can give accurate expectation values and readily estimate error bars. Furthermore, we explain the potential energy functions for biomolecules in this Chapter. In Chapter 3 we compare the results of the three simulations, namely, the conventional canonical, multicanonical, and multi-overlap MD simulations, and demonstrate the effectiveness of the multi-overlap MD algorithms. Chapter 4 is devoted to conclusions.

Bibliography

[1] C. B. Anfinsen, E. Haber, M. Sela and F. H. White, Proc. Natl. Acad. Sci. USA **47**, 1309 (1961).

[2] C. Levinthal, J. Chim. Phys. **65**, 44 (1968).

[3] D. B. Wetlaufer: Proc. Natl. Acad. Sci. USA **70**, 697 (1973).

[4] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[5] W. G. Hoover, A. J. C. Ladd, and B. Moran, Phys. Rev. Lett. **48**, 1818 (1982).

[6] D. J. Evans, J. Chem. Phys. **78**, 3297 (1983).

[7] S. Nosé, Mol. Phys. **52**, 255 (1984).

[8] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[9] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

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

[11] B. A. Berg and T. Neuhaus, Phys. Lett. **B267**, 249 (1991).

[12] B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).

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

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

[15] B. A. Berg, H. Noguchi and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).

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

[17] Y. Okamoto, Recent Res, Devel. in Pure & Applied Chem. **2**, 1 (1998).

[18] U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999).

Chapter 2

Simulation Methods

Satoru G. Itoh and Yuko Okamoto, “Multi-overlap molecular dynamics methods for biomolecular systems,” Chemical Physics Letters **400**, 308-313 (2004).

2.1 Introduction

In order to understand the protein folding or function, we must obtain the free-energy landscape of the protein system. We need simulation methods to sample efficiently the conformational space. Generalized-ensemble algorithms [1] are very powerful tools to have efficient sampling in the conformational space and very useful tools to understand the protein folding problem. In the following we discuss about several simulation methods. In Sec. 2.2 we explain the canonical-ensemble MD method [2]-[6] and we clarify problems of this method. Generalized-ensemble algorithms have been proposed to solve these problems. In Sec. 2.3 we present the multicanonical-ensemble MD method [8],[9], which is one of the generalized-ensemble algorithms. In Sec. 2.4 we propose a new generalized-ensemble algorithm, which we refer to as the multi-overlap MD algorithm. This method makes it possible to find transition states among any specific reference configurations. We also introduce the jackknife methods [26],[27] to estimate simulation errors in Sec. 2.5. In Sec. 2.6 we give details of the potential energy functions for biomolecules [29], which we employ in this thesis.

2.2 Canonical-Ensemble Algorithms

In this section we explain the conventional canonical-ensemble algorithm [2]-[6],[10]. The canonical ensemble is based on a system that keeps the temperature, volume, and number of particles fixed. In Sec. 2.2.1 we present on the canonical MD methods with the Gaussian thermostat [2],[3]. This method realizes a constant temperature by restricting the momentum vectors so that the total kinetic energy is a constant. In Sec. 2.2.2 we outline the simulated annealing method [7], which is a simple generalization of canonical-ensemble algorithms.

2.2.1 Canonical-ensemble MD algorithms

In the canonical ensemble at a constant temperature T_0, the probability distribution P_c of the potential energy E is represented by the product of the density of states $n(E)$ and the Boltzmann weight factor W_c:

$$\begin{aligned} P_c(E;T_0) &= n(E)W_c(E;T_0) \\ &= n(E)e^{-\beta_0 E} \ , \end{aligned} \tag{2.1}$$

where β_0 is given by $\beta_0 = 1/k_{\rm B}T_0$ ($k_{\rm B}$ is the Boltzmann constant). In Fig. 2.1, we show a probability distribution in Eq. (2.1). Canonical-ensemble algorithms [2]-[6],[10] at a constant temperature reproduce the probability distribution in Eq. (2.1) in computer simulations. Among various canonical-ensemble algorithms, we employed in this thesis canonical-ensemble MD method with the Gaussian thermostat [2],[3], which we refer to as the Gaussian constraint method. The equations of motion in this method are given by

$$\begin{aligned} \dot{\boldsymbol{q}}_i &= \frac{d\boldsymbol{q}_i}{dt} = \frac{\boldsymbol{p}_i}{m_i} \ , \\ \dot{\boldsymbol{p}}_i &= \boldsymbol{F}_i - \zeta_c \boldsymbol{p}_i \ , \end{aligned} \tag{2.2}$$

where m_i, $\boldsymbol{q}_i$, and $\boldsymbol{p}_i$ are the mass, coordinate vector, and momentum vector of atom i. The force $\boldsymbol{F}_i$ acting on atom i is given by

$$\boldsymbol{F}_i = -\frac{\partial E}{\partial \boldsymbol{q}_i} \ . \tag{2.3}$$

The coefficient ζ_c is chosen so as to guarantee that the total kinetic energy is constant:

$$\zeta_c = \frac{\sum\limits_i \boldsymbol{F}_i \cdot \dot{\boldsymbol{q}}_i}{2\sum\limits_i \dfrac{\boldsymbol{p}_i^2}{2m_i}} \ . \tag{2.4}$$

In Fig. 2.2, we show shapes of the probability distribution in the canonical ensemble at a high temperature and a low temperature. As shown in Fig. 2.2(b), the probability distributions in the canonical ensemble at low temperatures are very sharp, whereas those at high temperatures in Fig. 2.2(a) have wide shapes. In other words, fluctuations of the potential energy at low temperatures are small, and the simulation cannot overcome energy barriers for making transition from one state to another. Therefore, it is difficult for the conventional canonical-ensemble methods to sample the configurational space at low temperatures. In complex systems such as proteins, especially, these methods sample inefficiently the conformational space at low temperatures. This is because the complex systems such as proteins have many local-minimum states, and the usual

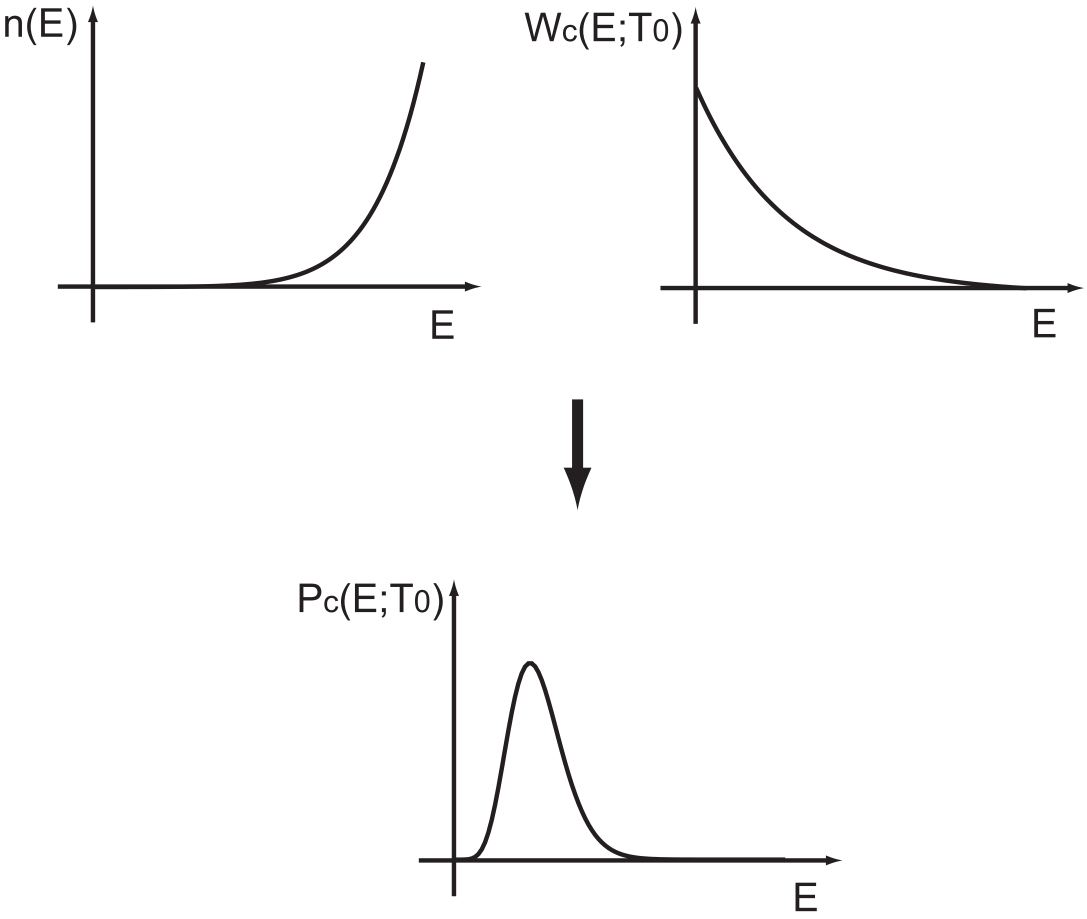

Figure 2.1: Probability distribution $P_c(E;T_0)$ of the potential energy E is represented by the product of the density of states $n(E)$ and the Boltzmann weight factor $W_c(E;T_0)$.

canonical-ensemble simulations tend to get trapped in the local-minimum states. Accordingly, we cannot obtain accurately the free-energy landscape of the complex systems.

2.2.2 Simulated annealing

Simulated annealing [7] is based on the process of annealing in which liquids freeze or metals recrystallize. In the annealing process, the system, which is initially at high temperature and disordered, is slowly cooled so that it is always approximately in thermal equilibrium. As cooling proceeds, the system becomes more ordered and approaches the global-minimum-energy state. However, if the initial temperature of the system is too low

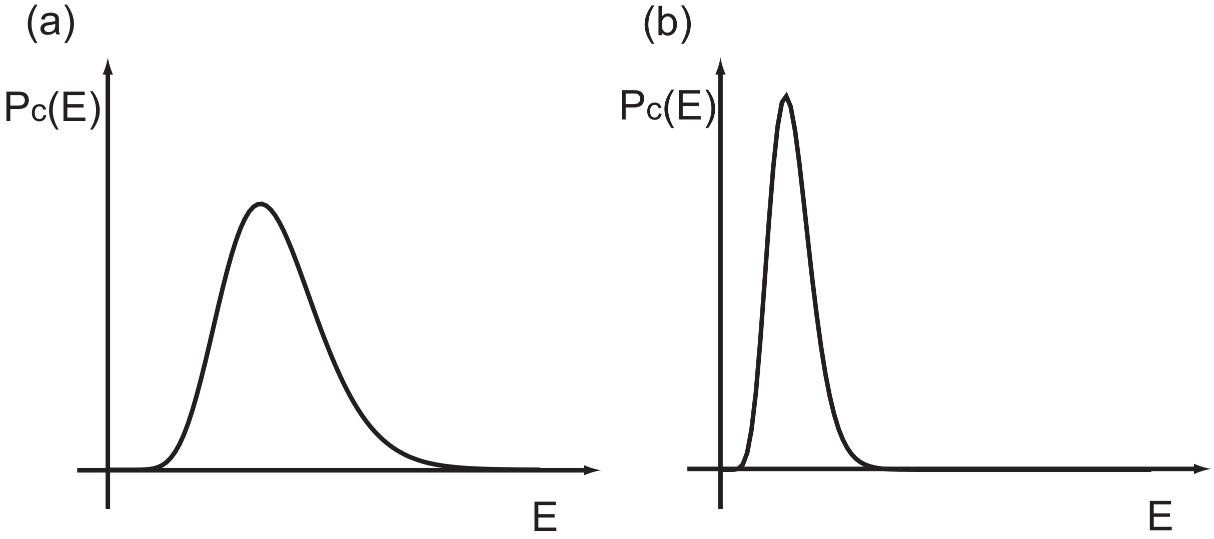

Figure 2.2: Probability distribution P_c of the potential energy E at (a) a high temperature and (b) a low temperature in the canonical ensemble.

or the process of annealing is not sufficiently slow, then the system may get trapped in a local-minimum-energy state.

In this thesis, we performed simulated annealing as follows. The initial temperature T_i of a protein system is set sufficiently high so that the protein is in a random-coil state. First, we carry out a canonical MD simulation at this temperature until the system achieves thermal equilibrium. Secondly, we slightly lower the temperature of the system and also perform the canonical MD simulation at that temperature until the system achieves equilibrium. This process is repeated until the temperature of the system reaches the final temperature T_f at which the protein folds into the native structure.

2.3 Multicanonical-Ensemble Algorithms

In this section we explain the multicanonical-ensemble algorithms [11],[12]. Multicanonical-ensemble simulations realize a free random walk in the potential energy space and efficiently sample the conformational space, as will be described in Sec. 2.3.1. In Sec. 2.3.2 we show the equations of motion for the multicanonical ensemble. In Sec. 2.3.3 we describe the reweighting techniques [13],[14]. We can calculate any physical quantities as functions of temperature by these techniques.

2.3.1 Free random walk in energy space

In the canonical-ensemble simulation methods at low temperatures, it is difficult to overcome energy barriers. To surmount this difficulty and sample the conformational space efficiently, generalized-ensemble algorithms have been proposed (for a review, see Ref. [1]). The multicanonical-ensemble MD method [8],[9] is one of commonly used generalized-ensemble algorithms. In the multicanonical ensemble [11],[12], each state is weighted by a non-Boltzmann weight factor W_{muca}, which we refer to as the multicanonical weight factor, instead of the Boltzmann weight factor W_c. The probability distribution P_{muca} of the potential energy E is defined to be uniform:

$$\begin{aligned} P_{muca}(E) &= n(E)W_{muca}(E) \\ &= \text{constant} . \end{aligned} \tag{2.5}$$

The multicanonical weight factor W_{muca} is given by

$$W_{muca}(E) = e^{-\beta_0 E_{muca}(E;T_0)} , \tag{2.6}$$

where E_{muca} is the 'multicanonical potential energy' and is defined so that the probability distribution in Eq. (2.5) becomes flat, and we have chosen an arbitrary reference temperature $T_0 = 1/k_{\mathrm{B}}\beta_0$. Note that by definition in Eq. (2.5) the multicanonical algorithm is independent of temperature. Here, we just introduce T_0 in order to write $W_{muca}(E)$ in a Boltzmann-weight-like factor. Eq. (2.5) implies that a free random walk in the potential energy space is realized. As a results, multicanonical simulations can overcome any energy barriers and sample efficiently the conformational space. In Fig. 2.3 we show a probability distribution in the multicanonical ensemble. That in the conventional canonical ensemble is also shown for comparison.

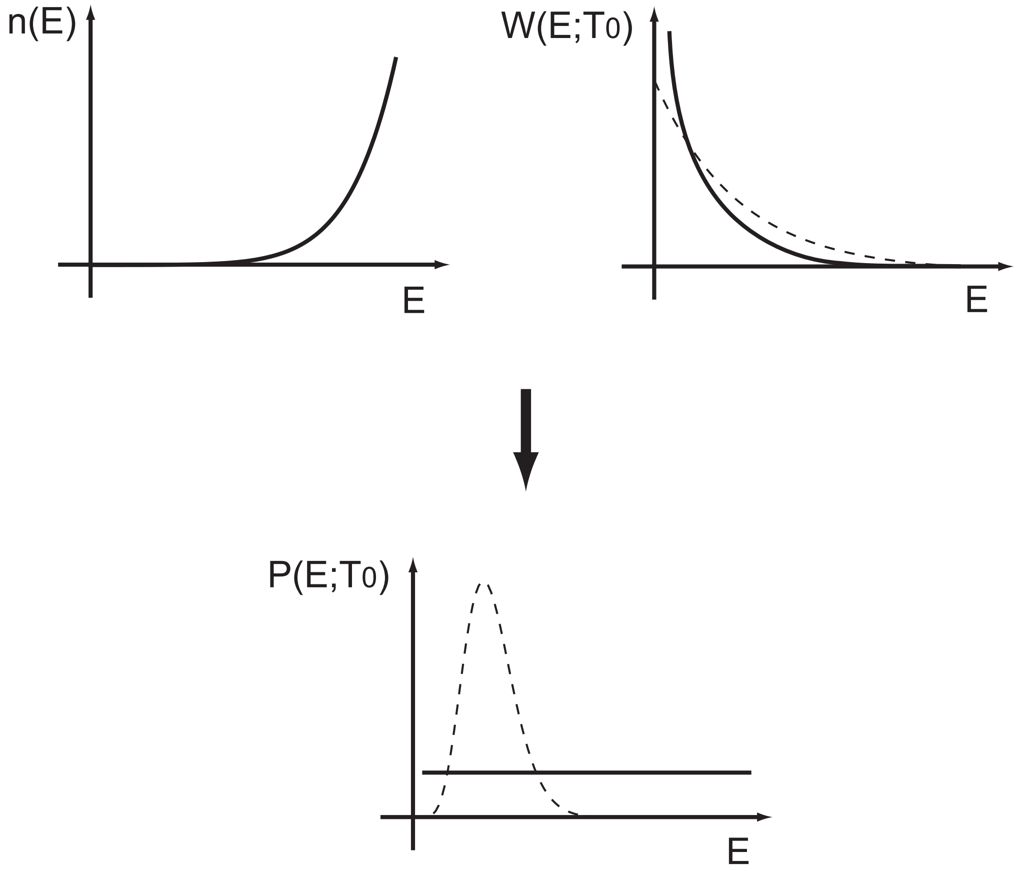

Figure 2.3: Probability distribution $P_{muca}(E)$ of the potential energy E is represented by the product of the density of states $n(E)$ and the multicanonical weight factor $W_{muca}(E)$. The solid line shows a case of the multicanonical ensemble. The dashed line shows a case of the canonical ensemble.

2.3.2 Equations of motion

The equations of motion for the multicanonical MD methods with the Gaussian thermostat are given by

$$\begin{aligned} \dot{\boldsymbol{q}}_i &= \frac{\boldsymbol{p}_i}{m_i} \ , \\ \dot{\boldsymbol{p}}_i &= \boldsymbol{F}_i^{muca} - \zeta_{muca}\boldsymbol{p}_i \ . \end{aligned} \tag{2.7}$$

The ‘force’ $\boldsymbol{F}_i^{muca}$ acting on atom i is defined by

$$\boldsymbol{F}_i^{muca} \quad = \quad -\frac{\partial E_{muca}(E;T_0)}{\partial \boldsymbol{q}_i}$$

$$= \frac{\partial E_{muca}(E;T_0)}{\partial E}\boldsymbol{F}_i \ . \tag{2.8}$$

The coefficient ζ_{muca} is calculated from

$$\begin{aligned}\zeta_{muca} &= \frac{\sum_i \boldsymbol{F}_i^{muca} \cdot \dot{\boldsymbol{q}}_i}{2\sum_i \frac{\boldsymbol{p}_i^2}{2m_i}} \\ &= \frac{\frac{\partial E_{muca}(E;T_0)}{\partial E}\sum_i \boldsymbol{F}_i \cdot \dot{\boldsymbol{q}}_i}{2\sum_i \frac{\boldsymbol{p}_i^2}{2m_i}} \ . \end{aligned} \tag{2.9}$$

The multicanonical potential energy $E_{muca}(E;T_0)$ is not *a priori* known and we must obtain its good estimate to flatten the probability distribution of the potential energy. From Eq. (2.5), incidentally, the multicanonical weight factor $W_{muca}(E;T_0)$ can be written as follows:

$$W_{muca}(E) = \frac{1}{n(E)} \ . \tag{2.10}$$

From Eq. (2.6), therefore, the multicanonical potential energy $E_{muca}(E;T_0)$ is given by

$$\begin{aligned} E_{muca}(E;T_0) &= k_{\mathrm{B}}T_0 \ \ln n(E) \\ &= T_0 S(E) \ , \end{aligned} \tag{2.11}$$

where $S(E)$ is the entropy in the microcanonical ensemble. One way to obtain the estimate of the multicanonical potential energy is to use the following relation:

$$\frac{\partial E_{muca}(E;T_0)}{\partial E} = \frac{T_0}{T(E)} \ , \tag{2.12}$$

where the following thermodynamic relation gives the definition of the temperature $T(E)$:

$$\left.\frac{\partial S(E)}{\partial E}\right|_{E=E_{ave}(T)} = \frac{1}{T(E)} \ , \tag{2.13}$$

with

$$E_{ave}(T) = \langle E \rangle_T \ . \tag{2.14}$$

Namely, $T(E_{ave})$ is the inverse function of Eq. (2.14). Accordingly, the multicanonical potential energy $E_{muca}(E;T_0)$ can be obtained by integrating Eq. (2.12) [15]-[17]:

$$E_{muca}(E;T_0) = T_0 \int_{E_{low}}^{E} \frac{dE'}{T(E')} \ , \tag{2.15}$$

where E_{low} is an arbitrary value close to the lower limit of the potential energy range of interest. Moreover, the equations of motion in Eq. (2.7) can be rewritten as follows:

$$\begin{aligned} \dot{\boldsymbol{q}}_i &= \frac{\boldsymbol{p}_i}{m_i} \ , \\ \dot{\boldsymbol{p}}_i &= \tfrac{T_0}{T(E)}\boldsymbol{F}_i - \zeta_{muca}\boldsymbol{p}_i \ , \end{aligned} \tag{2.16}$$

and

$$\zeta_{muca} = \frac{\dfrac{T_0}{T(E)}\sum_i \boldsymbol{F}_i \cdot \dot{\boldsymbol{q}}_i}{2\sum_i \dfrac{\boldsymbol{p}_i^2}{2m_i}} \ . \tag{2.17}$$

Multicanonical-ensemble MD simulations are performed by solving numerically these equations of motion. These simulations realize free random walks and sample efficiently the conformational space. By such efficient sampling in the configurational space, the multicanonical-ensemble MD simulations are able to give an accurate free-energy landscape in comparison with conventional canonical simulations.

2.3.3 Reweighting techniques

By using the obtained histogram $N_{muca}(E)$ of the potential energy from the results of a multicanonical MD simulation, the expectation value of a physical quantity A at any temperature T is calculated from

$$\langle A \rangle_T = \frac{\sum_E A(E)n(E)e^{-\beta E}}{\sum_E n(E)e^{-\beta E}} \ , \tag{2.18}$$

where the best estimate of the density of states is given by the single-histogram reweighting techniques [13],[14]:

$$\begin{aligned} n(E) &= \frac{N_{muca}(E)}{W_{muca}(E)} \\ &= N_{muca}(E)e^{\beta_0 E_{muca}(E;T_0)} \ . \end{aligned} \tag{2.19}$$

Because the multicanonical-ensemble MD simulations can sample the potential energy space widely, we are able to estimate correctly the density of states over a wide range in the potential energy space.

2.4 Multi-Overlap Algorithm

In this section we explain the formulation of the multi-overlap MD algorithm. The dihedral-angle distance [18] is defined as a reaction coordinate in Sec. 2.4.1. In Sec. 2.4.2 we introduce a non-Boltzmann weight factor, which we refer to as the multi-overlap weight factor. The multi-overlap weight factor realizes a constant probability distribution in a multi-dimensional dihedral-angle-distance space. In Sec. 2.4.3 we present the equations of motion in the multi-overlap ensemble. The equations of motion in the multi-overlap ensemble is constructed by adding the derivative of a dimensionless free energy to the equations of motion in the usual canonical ensemble. We discuss details of the updating procedure of the dimensionless free energy in Sec. 2.4.4. In Sec. 2.4.5 we explain the reweighting techniques [13],[14]. Utilizing the reweighting techniques, we can calculate appropriate physical quantities and obtain the free-energy landscape at any temperature.

2.4.1 Definition of dihedral-angle distance

In order to explore transition states among any reference configurations, we would like to perform a simulation which focuses on the reference configurations and does not get trapped in local-minimum states. While a free random walk in the potential energy space is realized in the multicanonical MD method, we would like to perform a random walk in some reaction coordinate so that the reference configuration can be efficiently sampled. In the multi-overlap algorithm [18], the overlap is introduced as this reaction coordinate. The overlap O with respect to a reference configuration is defined as follows [18],[19]:

$$O = 1 - d \ , \tag{2.20}$$

where d is the dihedral-angle distance given by

$$d = \frac{1}{n\pi} \sum_i d_a(v_i, v_i^0) \ . \tag{2.21}$$

Here, n is the total number of dihedral angles, v_i is the dihedral angle i, and v_i^0 is the dihedral angle i of the reference configuration. The distance $d_a(v_i, v_i^0)$ between two dihedral angles is defined by

$$d_a(v_i, v_i^0) = \min(|v_i - v_i^0|, 2\pi - |v_i - v_i^0|) \ . \tag{2.22}$$

The dihedral-angle distance d in Eq. (2.21) takes on a value in the range $0 \leq d \leq 1$. From Eq. (2.20), correspondingly, $0 \leq O \leq 1$. In particular, if we consider a system at infinite temperature ($T_0 = \infty$), the average values of the dihedral-angle distance d and the overlap O are $\frac{1}{2}$. This is because the distance $d_a(v_i, v_i^0)$ in Eq. (2.22) will have a uniform distribution in the range between 0 and π at $T_0 = \infty$. Furthermore, if $d = 0$ ($O = 1$), all dihedral angles are coincident with those of the reference configuration. The dihedral-angle distance (the overlap) is thus an indicator of how similar the conformation is to the reference conformation. As one can see in Eq. (2.20), the dihedral-angle distance d is equivalent to the overlap O. Hereafter, we employ the dihedral-angle distance d as the reaction coordinate in the multi-overlap algorithm.

2.4.2 Constant probability distribution in dihedral-angle distance space

We want the simulation to realize a random walk in a multi-dimensional dihedral-angle-distance space. In other words, the simulation needs to have a constant probability distribution with the dihedral-angle distance reaction coordinates. As discussed in Sec. 2.2, the probability distribution P_c is not constant and takes a much smaller value in high-energy region for the Boltzmann weight factor. Consequently, canonical-ensemble simulations will get trapped in local-minimum states at a low temperature. In the multi-overlap ensemble at a constant temperature T_0, on the other hand, the probability distribution is determined by the following non-Boltzmann weight factor, which we refer to as the multi-overlap weight factor:

$$W_{muov}(d; E) = e^{-\beta_0 E_{muov}(d;E)} , \tag{2.23}$$

where $E_{muov}(d; E)$ is the 'multi-overlap potential energy' defined by

$$E_{muov}(d; E) = E - k_{\rm B} T_0 f(d) . \tag{2.24}$$

The function $f(d)$ is the dimensionless free energy at dihedral-angle distance d.

The generalization to the multi-dimensional dihedral-angle distance space is straightforward, and the multi-overlap weight factor is given by

$$W_{muov}(d_1, \cdots, d_N; E) = e^{-\beta_0 E_{muov}(d_1,\cdots,d_N;E)} , \tag{2.25}$$

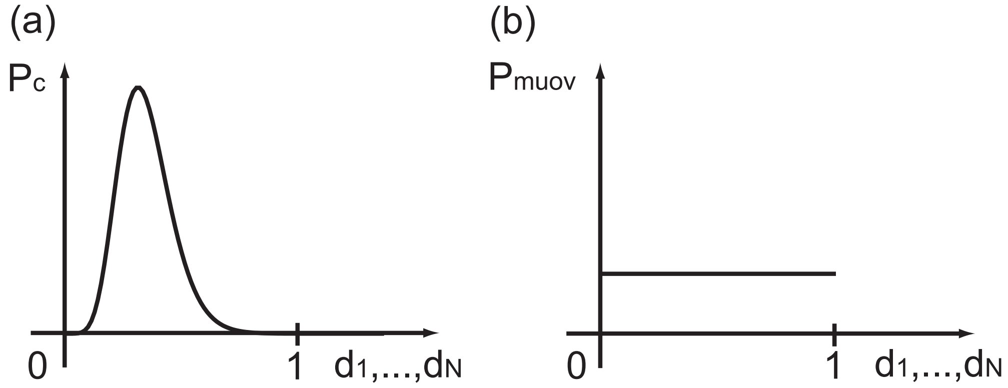

Figure 2.4: (a) probability distribution $P_c(d_1, \cdots, d_N)$ in canonical ensemble and (b) $P_{muov}(d_1, \cdots, d_N)$ in multi-overlap ensemble. $P_c(d_1, \cdots, d_N)$ has a bell-like shape but $P_{muov}(d_1, \cdots, d_N)$ has a uniform distribution.

and

$$E_{muov}(d_1, \cdots, d_N; E) = E - k_{\rm B}T_0 f(d_1, \cdots, d_N) \ , \tag{2.26}$$

where N is the number of the reference configurations and d_i is the dihedral-angle distance of reference configuration i ($i = 1, \cdots, N$). The function $f(d_1, \cdots, d_N)$ is the dimensionless free energy with the fixed values of dihedral-angle distances $d_1, \cdots, d_N$. The dimensionless free energy $f(d_1, \cdots, d_N)$ is defined so that the probability distribution P_{muov} is flat:

$$\begin{aligned} P_{muov}(d_1, \cdots, d_N) &= \int dE \ n(d_1, \cdots, d_N; E) W_{muov}(d_1, \cdots, d_N; E) \\ &= \int dE \ n(d_1, \cdots, d_N; E) e^{-\beta_0 E + f(d_1, \cdots, d_N)} \\ &\equiv \text{constant} \ , \end{aligned} \tag{2.27}$$

where $n(d_1, \cdots, d_N; E)$ is the density of states. In Fig. 2.4 we show probability distributions in canonical and multi-overlap ensemble with dihedral-angle distance axes. Thus, we are able to perform simulations, which realize a random walk in the multi-dimensional dihedral-angle distance space.

In this thesis we use only the two-dimensional version of these methods. Namely, $N = 2$ in Eqs. (2.25), (2.26), and (2.27). we can then perform a simulation which is focused on two specific reference configuration. Accordingly, we can explore a transition state between the two reference configurations. We will deal with the two-dimensional version of these methods hereafter.

2.4.3 Equations of motion in multi-overlap MD simulations

The equations of motion with Gaussian thermostat for the canonical MD simulations and the multicanonical MD simulations are described in Eq. (2.2) and Eq. (2.7), respectively. Correspondingly, the multi-overlap MD simulation is carried out by solving the following modified equations of motion with Gaussian thermostat:

$$\begin{aligned} \dot{\boldsymbol{q}}_i &= \frac{d\boldsymbol{q}_i}{dt} = \frac{\boldsymbol{p}_i}{m_i} \ , \\ \dot{\boldsymbol{p}}_i &= \boldsymbol{F}_i^{muov} - \zeta_{muov}\boldsymbol{p}_i \ . \end{aligned} \tag{2.28}$$

The 'force' $\boldsymbol{F}_i^{muov}$ acting on atom i is calculated from (see Eq. (2.26))

$$\begin{aligned} \boldsymbol{F}_i^{muov} &= -\frac{\partial E_{muov}}{\partial \boldsymbol{q}_i} \\ &= \boldsymbol{F}_i + k_{\rm B}T_0 \frac{\partial f(d_1,d_2)}{\partial \boldsymbol{q}_i} \ . \end{aligned} \tag{2.29}$$

The coefficient ζ_{muov} is defined by

$$\zeta_{muov} = \frac{\sum\limits_i \boldsymbol{F}_i^{muov} \cdot \dot{\boldsymbol{q}}_i}{2\sum\limits_i \frac{\boldsymbol{p}_i^2}{2m_i}} \ . \tag{2.30}$$

2.4.4 Determination of the dimensionless free energy

The dimensionless free energy $f(d_1,d_2)$ in Eq. (2.26) is not *a priori* known and we must obtain its good estimate by iterations of short simulations. Several methods [20]-[25] to determine the dimensionless free energy $f(d_1,d_2)$ exist and we determine it by the following process [22]. We update the dimensionless free energy $f(d_1,d_2)$ at each MD step of a short multi-overlap MD simulation, and we iterate this procedure. Suppose that we have $f = f^{(l)}(d_1,d_2 : k-1)$ at the $(k-1)$th MD step of the lth iteration of the short multi-overlap MD simulation, and that the configuration at the kth MD step has the value $d_1 = c_1$ and $d_2 = c_2$. We then update the dimensionless free energy by

$$f^{(l)}(d_1 = c_1, d_2 = c_2; k) = f^{(l)}(d_1 = c_1, d_2 = c_2; k-1) - a^{(l)} \ , \tag{2.31}$$

where $a^{(l)}$ is an appropriately chosen positive constant. The lth iteration of the multi-overlap MD simulation with the updating procedure of Eq. (2.31) is continued until the probability distribution $P_{muov}(d_1, d_2)$ in Eq. (2.27) becomes reasonably flat with fluctuations of order $a^{(l)}$. For the $(l+1)$th iteration, we make the value of the constant a smaller i.e., $a^{(l+1)} \leq a^{(l)}$, and repeat the updating procedure of Eq. (2.31) with l replaced by $l+1$. The initial value can be set as follows:

$$f^{(1)}(d_1, d_2; 0) = 0 \ . \tag{2.32}$$

The iteration is terminated when the probability distribution $P_{muov}(d_1, d_2)$ becomes satisfactorily flat. After the dimensionless free energy $f(d_1, d_2)$ is determined, we make a long production multi-overlap MD simulation of Eqs. (2.28) and (2.29) with this $f(d_1, d_2)$.

2.4.5 Reweighting techniques

Results of the multi-overlap production run can be analyzed by the reweighting techniques. Suppose that we have determined the dimensionless free energy $f(d_1, d_2)$ at a constant temperature T_0 and that we have made a production run at this temperature. The expectation value of a physical quantity A at any temperature T is calculated from

$$< A >_T = \frac{\sum\limits_{d_1,d_2,E} A(d_1, d_2; E) n(d_1, d_2; E) e^{-\beta E}}{\sum\limits_{d_1,d_2,E} n(d_1, d_2; E) e^{-\beta E}} \ , \tag{2.33}$$

where the best estimate of the density of states is given by the single-histogram reweighting techniques [13],[14]:

$$n(d_1, d_2; E) = \frac{N_{muov}(d_1, d_2; E)}{W_{muov}(d_1, d_2; E)} \ , \tag{2.34}$$

and $N_{muov}(d_1, d_2; E)$ is the histogram of the probability distribution that was obtained by the multi-overlap production run. By substituting Eqs. (2.25), (2.26), and (2.34) into Eq. (2.33), we have

$$< A >_T = \frac{\sum\limits_{d_1,d_2,E} A(d_1, d_2; E) N_{muov}(d_1, d_2; E) e^{\beta_0 E - f(d_1,d_2) - \beta E}}{\sum\limits_{d_1,d_2,E} N_{muov}(d_1, d_2; E) e^{\beta_0 E - f(d_1,d_2) - \beta E}} \ . \tag{2.35}$$

We can also calculate the free energy (or, the potential of mean force) with appropriate reaction coordinates. For example the free energy $F(\xi_1, \xi_2; T)$ with reaction coordinates ξ_1, ξ_2 at temperature T is defined by

$$F(\xi_1, \xi_2; T) = -k_{\rm B}T \ln P_c(\xi_1, \xi_2; T) \ , \tag{2.36}$$

where $P_c(\xi_1, \xi_2; T)$ is the reweighted canonical probability distribution of ξ_1 and ξ_2 and given by (see Eq. (2.35))

$$P_c(\xi_1, \xi_2; T) = \frac{\sum\limits_{d_1, d_2, E} N_{muov}(\xi_1, \xi_2; d_1, d_2; E) e^{\beta_0 E - f(d_1, d_2) - \beta E}}{\sum\limits_{\xi_1, \xi_2, d_1, d_2, E} N_{muov}(\xi_1, \xi_2; d_1, d_2; E) e^{\beta_0 E - f(d_1, d_2) - \beta E}} \ , \tag{2.37}$$

and $N_{muov}(\xi_1, \xi_2; d_1, d_2; E)$ is the histogram of the probability distribution that was obtained from the multi-overlap production run.

2.5 Jackknife Methods

We introduce the jackknife methods [26]-[28] to correct errors of results from computer simulations. Assume that we have random variables $\{x_i\}$ $(i = 1, \cdots, N)$. The mean value $\bar{x}$ is defined by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \ . \tag{2.38}$$

We consider an arbitrary function f of x. When f is a non-linear function, we have, in general,

$$\langle f(x) \rangle \neq \langle f(\bar{x}) \rangle \overset{N \to \infty}{\longrightarrow} \langle f(\hat{x}) \rangle \ , \tag{2.39}$$

where $\langle \cdots \rangle$ stands for expectation values and we write

$$\hat{x} \equiv \langle x \rangle \ . \tag{2.40}$$

Note that we have

$$\hat{x} = \lim_{N \to \infty} \bar{x} \ . \tag{2.41}$$

Thus, any non-linear function of random variables $\{x_i\}$ have bias. Typically, the function $f(x)$ and the function $\bar{f} \equiv f(\bar{x})$ have the bias of order 1 and the bias of order $\frac{1}{N}$, respectively. Namely, we have

$$\begin{aligned} \text{bias}(f) &\equiv \hat{f} - \langle f \rangle \\ &= O(1) \ , \end{aligned} \tag{2.42}$$

and

$$\begin{aligned} \text{bias}(\bar{f}) &\equiv \hat{f} - \left\langle \bar{f} \right\rangle \\ &= \frac{a_1}{N} + \frac{a_2}{N^2} + O\left(\frac{1}{N^3}\right) \ , \end{aligned} \tag{2.43}$$

where the function $\hat{f}$ is defined by

$$\begin{aligned} \hat{f} &\equiv \langle f(\hat{x}) \rangle \\ &= f(\hat{x}) \ . \end{aligned} \tag{2.44}$$

Therefore, the variance $\sigma^2(\bar{f})$, which is defined by

$$\sigma^2(\bar{f}) \equiv \left\langle (\bar{f} - \hat{f})^2 \right\rangle \ , \tag{2.45}$$

cannot be calculated from the standard equation for error bars:

$$\begin{aligned} \sigma^2(\bar{f}) &= \frac{1}{N}\sigma^2(f) \\ &= \frac{1}{N(N-1)} \sum_{i=1}^{N} (f_i - \bar{f})^2 \ . \end{aligned} \tag{2.46}$$

This is because $\langle f_i \rangle$ $(= \langle f(x_i) \rangle = \langle f \rangle)$ is not a valid estimator of $\hat{f}$ (see Eq. (2.42)). The jackknife methods reduce such a bias and provide well-founded error bar analysis.

In jackknife methods we use jackknife estimators f_i^J and $\bar{f}^J$:

$$f_i^J = f(x_i^J) \ , \tag{2.47}$$

and

$$\bar{f}^J = \frac{1}{N} \sum_{i=1}^{N} f_i^J \ , \tag{2.48}$$

where

$$x_i^J = \frac{1}{N-1}\sum_{k\neq i} x_k \ . \tag{2.49}$$

From Eq. (2.43) the bias of $\bar{f^J}$ is given by

$$\begin{aligned} \mathrm{bias}(\bar{f^J}) &\equiv \hat{f} - \left\langle \bar{f^J} \right\rangle \\ &= \frac{a_1}{N-1} + \frac{a_2}{(N-1)^2} + O\left(\frac{1}{N^3}\right) \ . \end{aligned} \tag{2.50}$$

Subsequently, we introduce bias-corrected estimators f_i^c and $\bar{f^c}$:

$$f_i^c = N\bar{f} - (N-1)f_i^J \ , \tag{2.51}$$

and

$$\bar{f^c} = \frac{1}{N}\sum_{i=1}^{N} f_i^c \ . \tag{2.52}$$

From Eqs. (2.43) and (2.50), the bias of $\bar{f^c}$ is given by

$$\begin{aligned} \mathrm{bias}(\bar{f^c}) &\equiv \hat{f} - \left\langle \bar{f^c} \right\rangle \\ &= -\frac{a_2}{N(N-1)} + O\left(\frac{1}{N^3}\right) \ . \end{aligned} \tag{2.53}$$

Accordingly, utilizing the bias-corrected estimators, we can reduce bias of an arbitrary function. Furthermore, the variance $\sigma^2(\bar{f^c})$ can be calculated from the standard equation:

$$\begin{aligned} \sigma^2(\bar{f^c}) &= \frac{1}{N}\sigma^2(f^c) \\ &= \frac{1}{N(N-1)}\sum_{i=1}^{N}(f_i^c - \bar{f^c})^2 \ . \end{aligned} \tag{2.54}$$

This variance is rewritten with the jackknife estimators f_i^J, $\bar{f^J}$ as follows:

$$\begin{aligned} \sigma^2(\bar{f^c}) &= \frac{1}{N(N-1)}\sum_{i=1}^{N}(f_i^c - \bar{f^c})^2 \\ &= \frac{1}{N(N-1)}\sum_{i=1}^{N}\Big((N-1)f_i^J - (N-1)\bar{f^J}\Big)^2 \\ &= \frac{N-1}{N}\sum_{i=1}^{N}(f_i^J - \bar{f^J})^2 \ . \end{aligned} \tag{2.55}$$

Thus, by using the jackknife methods, we can calculate readily the variance and the error bar of the an arbitrary function with respect to the mean of random variables in Eq. (2.49).

In summary, we use $\bar{f}^c$ in Eq. (2.52) as the estimator for $\hat{f} = f(\hat{x})$ and $\sigma(\bar{f}^c)$ in Eqs. (2.54) or (2.55) as the estimator for the error bar of the measurement of $\hat{f}$. In Appendix B we explain concretely the application of the jackknife methods to the reweighting techniques.

2.6 Potential Energy Function

Potential energy function which we adopted in this thesis is an all-atom model (CHARMM param22) [29]. This potential energy function E_{tot} is expressed as follows:

$$E_{tot} = E_{bond} + E_{angle} + E_{dih} + E_{imp} + E_{UB} + E_{LJ} + E_C \ . \tag{2.56}$$

Namely, it consists of the bond-stretching potential energy term E_{bond}, the bond-angle-bending potential energy term E_{angle}, the torsion potential energy term E_{dih}, the improper torsion potential energy term E_{imp}, the Urey-Bradley term E_{UB}, the Lennard-Jones 12-6 term E_{LJ}, and the electrostatic term E_C. In the following we explain each term one by one.

The first five potential energy terms characterize geometries of biomolecules. The bond-stretching potential energy term E_{bond} is defined by

$$E_{bond} = \sum_{bonds} K_b(b - b_0)^2 \ , \tag{2.57}$$

where K_b is the bond force constant, b is the bond length, and b_0 is the natural bond length, and $\sum\limits_{bonds}$ stands for the summation over all covalent bonds in biomolecules. This term is needed to keep the covalent bond of biomolecules and represents the vibration two atoms that are covalently bound. In Fig.2.5 we illustrate a vibration of two atoms connected by a covalent bond.

The bond-angle-bending potential energy term E_{angle} is given by

$$E_{angle} = \sum_{angle} K_\theta(\theta - \theta_0)^2 \ , \tag{2.58}$$

where K_θ is the angle force constant, θ is the bond angle, θ_0 is the natural bond angle, and $\sum\limits_{angle}$ stands for the summation over all bond angles in biomolecules. This term controls bendings of bond angles. Fig.2.6 illustrates the bending of a bond angle.

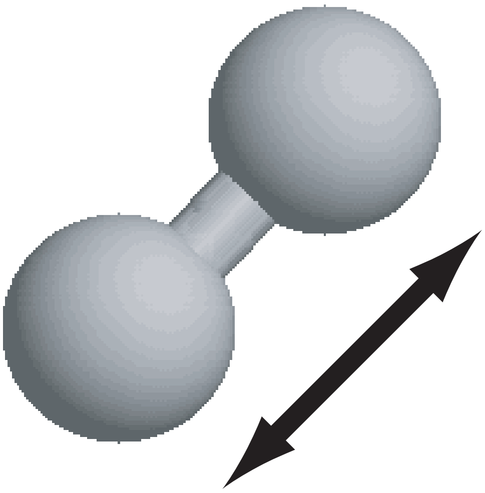

Figure 2.5: Illustration that represents the bond-stretching potential energy term. The figure was created with RasMol [30]

In Fig. 2.7 we describe the model of the Urey-Bradley term E_{UB}. The Urey-Bradley term E_{UB} is defined by

$$E_{UB} = \sum_{UB} K_{UB}(S - S_0)^2 \ , \tag{2.59}$$

where K_{UB} is the Urey-Bradley force constant, S is the Urey-Bradley 1,3-distance, S_0 represents the equilibrium value, and $\sum\limits_{UB}$ stands for the summation over all pairs of atoms in 1,3 configurations. The purpose of this term is to account for steric interactions between non-bonded atoms.

The torsion potential energy term E_{dih} is calculated from

$$E_{dih} = \sum_{dihedrals} K_{\chi} \left(1 + \cos(n\chi - \delta)\right) \ , \tag{2.60}$$

where K_{χ} is the dihedral angle force constant, χ is the dihedral angle, and $\sum\limits_{dihedrals}$ stands for the summation over all dihedral angles in biomolecules. Fig.2.8 is an illustration that corresponds to this potential energy term.

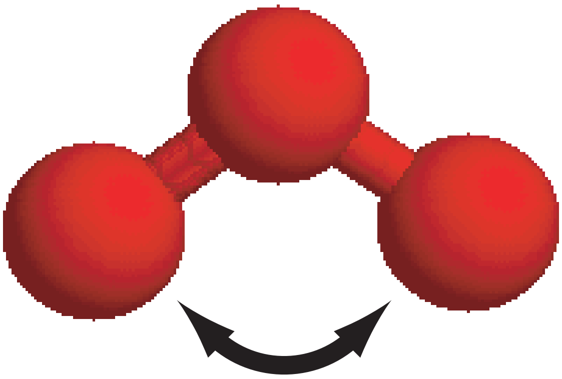

Figure 2.6: Illustration that represents the bond-angle-bending potential energy term.

The improper torsion potential energy term E_{imp} is given by

$$E_{imp} = \sum_{impropers} K_{imp}(\phi - \phi_0)^2 \ , \tag{2.61}$$

where K_{imp} is the improper torsion force constant, ϕ is the improper torsion angle, with the subscript zero representing the equilibrium value, and $\sum\limits_{impropers}$ stands for the summation over all improper torsion angles in biomolecules. The improper torsion potential energy term has been designed both to maintain chirality about a tetrahedral extended heavy atom (e.g., an α carbon), and to maintain planarity about certain planar atoms (such as a carbonyl carbon). In Fig.2.9 we show a perpendicular motion for a plane constructed by three atoms excepted for a center atom.

The Lennard-Jones 12-6 term E_{LJ} is defined by

$$E_{LJ} = \sum_{nonbonds} \epsilon_{ij} \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] \ , \tag{2.62}$$

where ϵ_{ij} is the Lennard-Jones 12-6 well depth, r_{ij} is the distance between atoms i and j, $R_{min_{ij}}$ is the value of r_{ij} where the Lennard-Jones 12-6 potential energy becomes zero, and $\sum\limits_{nonbonds}$ stands for the summation over all non-bond atom-pairs. The Lennard-Jones parameters between pairs of different atoms are obtained from the Lorentz-Berthelodt

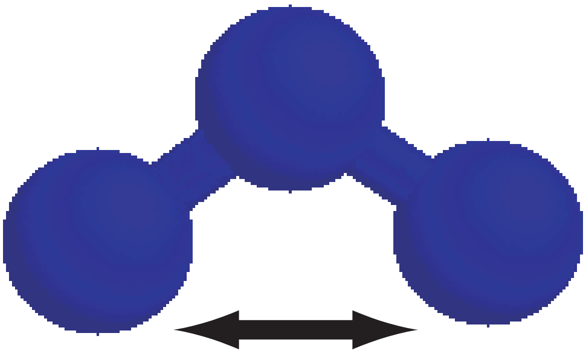

Figure 2.7: Illustration that represents the Urey-Bradley term.

combination rules, in which ϵ_{ij} values are based on the geometric mean of ϵ_{ii} and ϵ_{jj} and $R_{min_{ij}}$ values are based on the arithmetic mean between $R_{min_{ii}}$ and $R_{min_{jj}}$.

The electrostatic term E_C is the Coulombic term that is given by

$$E_C = \sum_{nonbonds} \frac{q_i q_j}{\epsilon r_{ij}} \ , \tag{2.63}$$

where q_i is the partial atomic charge and ϵ is the effective dielectric constant. The value of ϵ is 1 for vacuum and about 80 for the bulk water environment. Note that when we include explicit water molecules in the simulation, we also take the value $\epsilon = 1$.

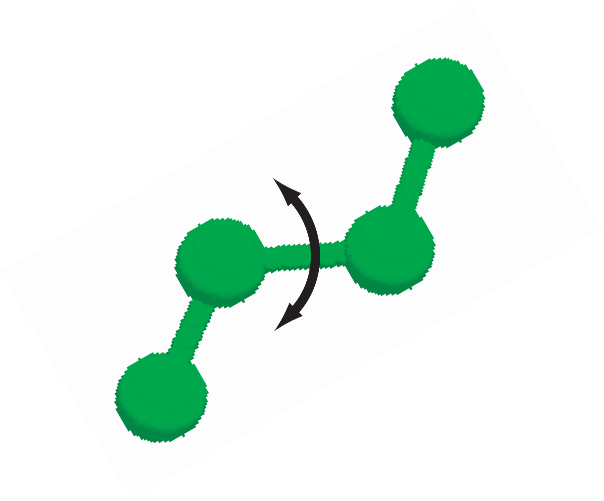

Figure 2.8: Illustration that represents the torsion potential energy term.

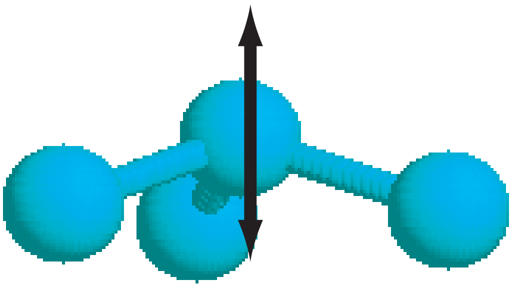

Figure 2.9: Illustration that represents the improper torsion potential energy term.

Bibliography

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

[2] W. G. Hoover, A. J. C. Ladd, and B. Moran, Phys. Rev. Lett. **48**, 1818 (1982).

[3] D. J. Evans, J. Chem. Phys. **78**, 3297 (1983).

[4] S. Nosé, Mol. Phys. **52**, 255 (1984).

[5] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[6] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[7] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).

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

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

[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[11] B. A. Berg and T. Neuhaus, Phys. Lett. **B267**, 249 (1991).

[12] B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).

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

[14] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1658 (1989).

[15] U. H. E. Hansmann, Phys. Rev. E **56**, 6200 (1997).

[16] Y. Sugita and Y. Okamoto in: T. Schlick and H.H. Gan (Eds.) Lecture Notes in Computatinal Science and Engineering, Springer-Verlag, Berlin, 2002, pp. 304-332; cond-mat/0102296.

[17] T. Terada, Y. Matsuo, and A. Kidera, J. Chem. Phys. **118**, 4306 (2003).

[18] B. A. Berg, H. Noguchi and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).

[19] U. H. E. Hansmann, M. Masuya and Y. Okamoto, Proc. Natl. Acad. Sci. USA **94**, 10652 (1997).

[20] B. A. Berg and T. Celik, Phys. Rev. Lett. **69**, 2292 (1992).

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

[22] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[23] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329**, 261 (2000).

[24] A. Mitsutake, Y. Sugita and Y. Okamoto, J. Chem. Phys. **118**, 6664 (2003).

[25] A. Mitsutake, Y. Sugita and Y. Okamoto, J. Chem. Phys. **118**, 6676 (2003).

[26] M. H. Quenouille, Biometrika **43**, 353 (1956).

[27] R. G. Miller, Biometrika **61**, 1 (1974).

[28] B. A. Berg "Markov Chain Monte Carlo Simulations and Their Statistical Analysis,", (World Scientific, Singapore, 2004).

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

[30] R. A. Sayle and E. J. Milner-White, Trends Biochem. Sci. **20**, 374 (1995).

Chapter 3

Theoretical Studies of Transition States by Multi-Overlap Molecular Dynamics Methods

Satoru G. Itoh and Yuko Okamoto, “Multi-overlap molecular dynamics methods for biomolecular systems,” Chemical Physics Letters **400**, 308-313 (2004).

Satoru G. Itoh and Yuko Okamoto, “Theoretical studies of transition states by the multi-overlap molecular dynamics methods,” in preparation.

3.1 Introduction

We proposed the multi-overlap MD methods to sample efficiently the conformational space and explore the transition states among specific reference configurations in Sec. 2.4. In this Chapter, we apply the multi-overlap MD method to a penta-peptide Met-enkephalin in vacuum with two reference configurations and check the effectiveness of this method by comparing the results with those of the conventional canonical MD method [1]-[5] and the multicanonical MD method [6],[7]. Moreover, from the detailed free-energy landscape obtained from the results of the multi-overlap MD simulation, we identify a transition pathway between two specific configurations of Met-enkephalin. The details of the condition of various simulation methods are given in Section 3.2. We present the results of the application of these methods to Met-enkephalin in Section 3.3.

3.2 Computational Details

Met-enkephalin is one of the simplest peptides and has the amino-acid sequence Tyr-Gly-Gly-Phe-Met. This peptide is often adopted as a test system in biomolecular simulations. Therefore, we also adopted Met-enkephalin in vacuum as a test system of the multi-overlap MD method. In our simulations the N-terminus and the C-terminus were blocked with the acetyl group and the N-methyl group, respectively. This is because we wanted the total charge of the Met-enkephalin system to be neutral. Accordingly, the total number of atoms of Met-enkephalin in our simulations is 84. The force field that we adopted is the CHARMM param 22 parameter set [8] (see Sec. 2.6). Our multi-overlap MD simulations were performed by implementing the method in the CHARMM macromolecular mechanics program [9]. The main part of implementation is shown as follows. We introduced the Gaussian constraint method (Gaussian thermostat) [1],[2] to the CHARMM macromolecular mechanics program. The corresponding equations of motion were implemented. Namely, we used Eq. (2.2) for the canonical MD simulations, Eq. (2.7) for the multicanonical MD simulations, and Eq. (2.28) for the multi-overlap MD simulations. The time step was taken to be 0.5 fs and leap-frog algorithm [10] was employed for the numerical integration (see Appendix A).

We consider two local-minimum-energy states of Met-enkephalin as reference configurations. These configurations were obtained by the simulated annealing MD method [11], which was explained in Sec. 2.2.2. During the simulated annealing run, the temperature was decreased linearly from 1000 K to 100 K with an increment of 50 K, and the canonical MD simulations were performed for 500 ps at each temperature (9.5 ns in total). This simulated annealing MD run was repeated 10 times with different initial random numbers. The obtained final conformations were further minimized by the conjugate gradient method, and two conformations were identified as the reference configurations from the backbone hydrogen-bond patterns. In Fig. 3.2 we show these reference configurations of Met-enkephalin. Reference configuration 1 (RC1) has a β-turn structure with two backbone hydrogen bonds between Gly-2 and Met-5, and reference configuration 2 (RC2) has a γ-turn structure with two backbone hydrogen bonds between Gly-2 and Phe-4. Reference configuration 1 also has a hydrogen bond between hydrogen bond acceptor CO of Gly-2 and hydrogen bond donor NH of Phe-4. We remark that with ECEPP/2 energy function [12]-[14] RC1 corresponds to the global-minimum state and RC2 corresponds to a local-minimum state [15].

The backbone dihedral angles are of three types: the rotation angle about the $\mathrm{N}-\mathrm{C}^{\alpha}$ bond of the backbone (ϕ), that about the $\mathrm{C}^{\alpha}-\mathrm{C}$ bond (ψ), and that about the peptide bond $\mathrm{C}-\mathrm{N}$ (ω). In Fig. 3.1 we illustrate the definitions for these dihedral angles. Our multi-overlap MD simulation was performed using the all-atom model, but we used only ϕ and ψ angles in the definition of the dihedral-angle distances in Eq. (2.21). This is because the dihedral angles of the backbone ω have almost the fixed value of 180° for the peptide bond $\mathrm{C}-\mathrm{N}$. Furthermore, by using only the backbone dihedral angles (and not side-chain dihedral angles) as the elements of the dihedral-angle distances, we focused on the backbone structures of Met-enkephalin. In Eq. (2.21), consequently, the number n of the elements of the dihedral-angle distances is 10 because Met-enkephalin has five pairs of ϕ and ψ. In Table 3.1 we list the dihedral angles ϕ, ψ of the two reference configurations in Fig. 3.2.

Our multi-overlap MD simulation was carried out at $T_0 = 300$ K. We first have to determine the multi-overlap weight factor $W_{muov}(d_1, d_2; E)$ in Eq. (2.25), or the dimensionless free energy $f(d_1, d_2)$ in Eq. (2.26), to get a flat probability distribution in the two-dimensional dihedral-angle distance space (d_1, d_2). For that purpose we used the procedure in Sec. 2.4.4. We first set $f^{(1)}(d_1, d_2) = 0$ according to Eq. (2.32). We then performed the multi-overlap MD simulation of Eq. (2.28) for 14 ns. The dimensionless free energy $f^{(1)}(d_1, d_2)$ was updated by Eq. (2.31) at each MD step with $a^{(1)} = 0.0001$. For this calculation, the dihedral-angle distances (d_1, d_2) were discretized with a bin size of 0.01. This 14 ns MD simulation was sufficient to obtain an optimal multi-overlap weight factor, and we did not further iterate the process. Finally, the multi-overlap MD production run was then performed with this weight factor for 24 ns after equilibration of 1 ns. Because the multi-overlap MD simulations perform a random walk in the configurational space, the results will not depend on the initial conformation. For the initial conformation of the multi-overlap MD simulation production run, we thus simply adopted one of the final conformations obtained by the above simulated annealing runs. In Fig. 3.3 we show this initial conformation and list their backbone dihedral angles in Table 3.2.

For the purpose of comparisons, we also performed a usual canonical MD simulation and a multicanonical MD simulation for 24 ns at $T_0 = 300$ K. We already explained the canonical and multicanonical MD methods in Secs. 2.2 and 2.3. These MD simulations were also performed by implementing the corresponding equations of motion in the CHARMM macromolecular mechanics program. The multicanonical weight factor, or equivalently the multicanonical potential energy $E_{muca}(E; T_0)$, was determined from Eq. (2.15). Namely, we obtained $\langle E \rangle_T$ from the canonical MD simulations at 19 temperatures ranging from 100 K to 1000 K with an equal interval of 50 K. We then obtained 19 values of $T(E)$ as the inverse function of $\langle E \rangle_T$. We then numerically integrated Eq. (2.15) by the trapezoidal rule to obtain $E_{muca}(E; T_0)$. The initial conformation for both the canonical production run and the multicanonical production run was the same as that for the multi-overlap production run.

3.3 Results and Discussion

We developed the multi-overlap MD method to realize a random walk in the dihedral-angle distance space and focus on the specific reference configurations (see Sec. 2.4). In this

Table 3.1: Backbone dihedral angles ϕ and ψ for reference configuration 1 and reference configuration 2.

Reference configuration 1			Reference configuration 2		
Residue	Type	Angle	Residue	Type	Angle
1	ϕ_1	$-100.1°$	1	ϕ_1	$-136.0°$
1	ψ_1	$136.2°$	1	ψ_1	$139.3°$
2	ϕ_2	$-149.2°$	2	ϕ_2	$-163.8°$
2	ψ_2	$56.6°$	2	ψ_2	$68.8°$
3	ϕ_3	$76.4°$	3	ϕ_3	$88.7°$
3	ψ_3	$-78.2°$	3	ψ_3	$-61.0°$
4	ϕ_4	$-87.9°$	4	ϕ_4	$-108.3°$
4	ψ_4	$-37.5°$	4	ψ_4	$-179.7°$
5	ϕ_5	$-79.8°$	5	ϕ_5	$-92.2°$
5	ψ_5	$138.9°$	5	ψ_5	$146.1°$

section we present the results of the multi-overlap MD simulation of Met-enkephalin in vacuum. Furthermore, we compare the results of the usual canonical, multicanonical, and multi-overlap MD simulations. The various time series are given in Sec. 3.3.1. In Sec. 3.3.2 we show the raw data of the probability distributions and discuss the effectiveness of the multi-overlap MD method. The physical quantities can be calculated by the reweighting techniques [16],[17]. In Sec. 3.3.3 the physical quantities, which were obtained from the usual canonical and multi-overlap MD simulations are compared with those from the multicanonical MD simulation. In the last Section we describe the detailed free-energy landscape calculated from the multi-overlap MD simulation and identify conformations in the transition state between RC1 and RC2.

3.3.1 Time series of simulations

We first examine time series of various quantities from the usual canonical, multicanonical, and multi-overlap MD simulations. Figs. 3.4, 3.5, and 3.6 show the time series of the dihedral-angle distances with respect to each of the two reference configurations. When $d_1 = 0$, the values of the backbone dihedral angles are completely coincident with those

Table 3.2: Backbone dihedral angles ϕ and ψ for the initial conformation.

Initial conformation		
Residue	Type	Angle
1	ϕ_1	$-107.3°$
1	ψ_1	$149.5°$
2	ϕ_2	$-156.7°$
2	ψ_2	$64.6°$
3	ϕ_3	$68.3°$
3	ψ_3	$-89.2°$
4	ϕ_4	$-89.3°$
4	ψ_4	$-13.4°$
5	ϕ_5	$-77.7°$
5	ψ_5	$110.3°$

of reference configuration 1 and it turned out that $d_2 = 0.159$. Conversely, when $d_2 = 0$, we have $d_1 = 0.159$. In the usual canonical MD simulation at $T_0 = 300$ K (see Fig. 3.4), the configuration transited from a RC1-like state to a RC2-like state near 5 ns, and did not transit back from the RC2-like state to the RC1-like state. In other words, the canonical MD simulation got trapped in the RC2-like local-minimum state. Thus, the usual canonical MD simulation does not sample efficiently the conformational space, and we cannot calculate accurate free-energy landscape. On the one hand, the multicanonical MD simulation did not get trapped in the local-minimum states, as we can see Fig. 3.5. Both d_1 and d_2 we observe random walks both in d_1 space and in d_2 space; both dihedral-angle distances often visited small values as well as large values beyond 0.5, which is the average value at $T_0 = \infty$. Therefore, we had efficient sampling in the conformational space in the multicanonical MD simulation. When we look into Fig. 3.5(a) more carefully, however, we find that the multicanonical MD simulation did not sample around the RC1-like state very much (d_1 values did not take very small values). Accordingly, we may not obtain accurate free-energy landscape near RC1 from the results of the multicanonical MD simulation. Finally, as one can see in Fig. 3.6, the multi-overlap MD simulation did not get trapped in the local-minimum states, either. Although the ranges of the dihedral-angle distances that were covered are less in the multi-overlap MD simulation than in the multicanonical MD simulation (reflecting the fact that the latter explores a wide conformational space than the former), the multi-overlap simulation indeed visited both RC1 state and RC2 state. We observe transitions between RC1 state and RC2 state several times in the Figure. Thus, the multi-overlap MD simulation can realize a random walk in the two-dimensional dihedral-angle distance space and yet focus on the two reference configurations RC1 and RC2.

In Figs. 3.7, 3.8, and 3.9 we show the time series of the root-mean-square distance (RMSD) of the backbone of Met-enkephalin with respect to each of the two reference configurations from the canonical, multicanonical, and multi-overlap MD simulations, respectively. The RMSD r_i with respect to reference configuration i is defined by

$$r_i = \min \left[\sqrt{\frac{1}{N} \sum_j (\boldsymbol{q}_j - \boldsymbol{q}_j^{(i)})^2} \right] \ , \qquad (3.1)$$

where N is the number of atoms, $\{\boldsymbol{q}_j^{(i)}\}$ are the coordinates of reference configuration i, and the minimization is over the rigid translations and rigid rotations of the coordinates of the configuration $\{\boldsymbol{q}_j\}$ with respect to the center of geometry. The behavior of the three simulations in Fig. 3.7, Fig. 3.8, and Fig. 3.9 is the same as in Fig. 3.4, Fig. 3.5, and Fig. 3.6, respectively; there are strong correlations between the dihedral-angle distance d_1 (d_2) and the RMSD r_1 (r_2). By employing the RMSD as the reaction coordinates, however, the boundary between RC1-like state and RC2-like state is more clarified. Incidentally, when $r_1 = 0$ ($r_2 = 0$), we have $r_2 = 1.52$ ($r_1 = 1.52$).

Figs. 3.10, 3.11, and 3.12 show the time series of the potential energy of the three simulations. The multicanonical MD simulation covers widely the potential energy space, as we can see in Fig. 3.11. The time series of the potential energy of the multi-overlap MD simulation, however, is not much different from that of the canonical MD simulation (compare Figs. 3.10 and 3.12). This is because the multi-overlap algorithm is based on the Boltzmann weight factor at temperature T_0 as far as energy dependence is concerned (see Eqs. 2.25 and 2.26), while the multicanonical algorithm is independent of temperature. The multi-overlap MD method aims at a random walk in the dihedral-angle distance space, not in the potential energy space.

3.3.2 Raw probability distributions of configurations

We discuss the probability distributions of configuration from the three simulations, the usual canonical, multicanonical, and multi-overlap MD simulations. In Figs. 3.13, 3.14, and 3.15 we show the raw data of the histograms with respect to the two dihedral-angle-distance coordinates. The bin size of the two-dimensional histograms is 0.01×0.01. These histograms represent the probability distributions in the two-dimensional dihedral-angle-distance space for the three ensembles. From Figs. 3.13 and 3.14, it is obvious that the probability distributions of the usual canonical and multicanonical MD simulations are biased towards RC2; there are pronounced peaks near $(d_1, d_2) = (0.159, 0.0)$. In other words, as previously stated, the usual canonical and multicanonical MD simulations did not sample efficiently the RC1-like states (near $(d_1, d_2) = (0.0, 0.159)$). In Fig. 3.15, on the other hand, we confirm that the multi-overlap MD simulation has a rather flat probability distribution in the two-dimensional dihedral-angle-distance space containing both RC1 state and RC2 state (see Eq. (2.27)).

In Figs. 3.16, 3.17, and 3.18 we show the raw data of the histograms with respect to the two RMSD coordinates. The bin size of the two-dimensional histograms is 0.1 Å × 0.1 Å. In this case, the histograms were taken every 100 MD steps (50 fs). Therefore, these histograms are rugged in comparison with those with the dihedral-angle-distance coordinates in Figs. 3.13, 3.14, and 3.15, where the data were taken every MD step (0.5 fs). The two peaks that correspond to RC1 and RC2 states are disconnected in the case of the canonical MD simulation (see Fig 3.16), and they are connected in both the multicanonical MD simulation and the multi-overlap MD simulation (see Figs. 3.17 and 3.18). However, while the multicanonical MD simulation has to visit a region with large r_1 and r_2 (high energy region) in order to have transitions between RC1 and RC2, the multi-overlap MD simulation can connect both states within a region with small r_1 and r_2. The characteristics of the probability distributions in Fig. 3.16, Fig. 3.17, and Fig. 3.18 are essentially the same as in Fig. 3.13, Fig. 3.14, and Fig. 3.15, respectively. Namely, in Figs. 3.16 and 3.17 the probability distributions are biased distribution towards RC2, and that from the multi-overlap MD simulation in Fig. 3.18 has finite contributions in both RC1 state and RC2 state. In the multi-overlap ensemble, however, the probability distribution is not needed to become flat in the RMSD space (compare Figs. 3.14 and 3.17). This is because the multi-overlap ensemble is devised to obtain a flat probability distribution in the dihedral-angle-distance space and not in the RMSD space. Nevertheless, the multi-overlap MD simulation realized efficient sampling in the RMSD space between RC1 and RC2. Thus, the multi-overlap MD simulation is suitable to sample between the reference configurations in comparison with the other methods.

Fig. 3.19 shows the raw data of the probability distributions of the potential energy. The bin size of histograms is 1.0 kcal/mol. The probability distribution of the potential energy in the multicanonical MD simulation, as a matter of course, is flat. Thus, the multicanonical methods is suitable to sample the potential energy space, not the conformational space between the specific reference configurations. The probability distribution of the potential energy in the multi-overlap MD simulation is almost the same as that in the usual canonical MD simulation. The probability distribution in the multi-overlap MD simulation is, however, a little wider than in the usual canonical MD simulation. This is because the multi-overlap MD simulation has to sample a little higher-energy region in order to overcome the potential energy barrier between RC1-like state and RC2-like state.

3.3.3 Physical quantities calculated by the reweighting techniques

We now examine the physical quantities calculated from the results of the three simulations, the usual canonical, multicanonical, multi-overlap MD simulation, by the reweighting techniques. The reweighting techniques for the multicanonical MD method and the multi-overlap MD method were explained in Sec. 2.3.3 and Sec. 2.4.5, respectively. Those for the canonical MD method are essentially the same as for the multicanonical algorithm; in Eq. 2.19 we just replace the multicanonical weight factor $W_{muca}(E)$ by the Boltzmann weight factor $W_c(E; T_0)$.

In Fig. 3.20 we show the probability distributions and physical quantities calculated by the reweighting techniques. Here, the error bars were calculated by the jackknife method [18]-[20] (see Sec. 2.5 and Appendix B). The number of bins was taken to be 8. The results from the multicanonical MD simulation are shown as a reference in the Figure, because the multicanonical algorithm is well-known for giving accurate expectation values for a wide range of temperature [21]. As we can see Fig. 3.20(a), the probability distributions of the potential energy at $T = 300$ K calculated from the results of the usual canonical and multi-overlap MD simulations are in good agreement with those of the multicanonical MD simulation. Furthermore, the average potential energy as a function of temperature is also in agreement with that from the multicanonical MD simulation, although we see slight deviations beyond error bars below $T \approx 250$ K and above $T \approx 350$ K in the case of the canonical MD simulation. In Fig. 3.20(c), however, we see that the specific heat as a function of temperature calculated from the results of the canonical MD simulation does not coincide with those of the multicanonical MD simulation in the entire temperature range (the error bars do not overlap). This is because the usual canonical MD simulation got trapped in the local-minimum states and did not have enough sampling in the conformational space. The specific heat here is defined by

$$\begin{aligned} C_v &= \frac{1}{k_\mathrm{B}} \frac{d \langle E \rangle_T}{dT} \\ &= \beta^2 \left(\left\langle E^2 \right\rangle_T - \langle E \rangle_T^2 \right) . \end{aligned} \tag{3.2}$$

The specific heat is the derivative of the average potential energy, and it is more difficult to obtain accurate results than the average potential energy itself. In the case of the multi-overlap MD simulation, the results well coincide with those from the multicanonical MD simulation between about 250 K and 350 K. In the region under 250 K and above 350 K, however, we see deviations between the results of the two simulations. This sets a reliable range of temperature where accurate thermodynamic quantities can be calculated by the multi-overlap MD simulation. The reason for the deviations is that the multi-overlap algorithm samples conformations in the dihedral-angle distance space but not in the energy space. Accordingly, the multi-overlap simulation is difficult to give an accurate estimate of the density of states in Eq. (2.34) over a wide potential energy range. Thus, in the multi-overlap MD method, the expectation values calculated by the reweighting techniques in Eq. (2.35) are correct only in the neighborhood of the temperature at which simulations were performed.

3.3.4 Free-energy landscape and transition states

We now study the free-energy landscape that given information about the transition between the two states, RC1 and RC2. The free-energy landscape was calculated from Eq. (2.36) with appropriate reaction coordinates by the reweighting techniques. In Figs. 3.21, 3.22, and 3.23 we show the free-energy landscape at $T = 300$ K obtained from the three simulations with respect to the reaction coordinates of the two dihedral-angle distances. The free-energy landscape of the usual canonical MD simulation is inaccurate due to insufficient sampling in the conformational space as previously mentioned. The results from the multicanonical MD simulation have rugged surface but cover a wide region in the two-dimensional dihedral-angle distance space in comparison with those of the multi-overlap MD simulation (compare Figs. 3.22 and 3.23). This is because the multi-overlap method samples efficiently and selectively the conformational space between the two reference configurations. On the other hand, the multicanonical MD simulation makes widely sampling in the conformational space, but not focuses on specific reference configurations. Thus, the multi-overlap method is better in the sense that a detailed free-energy landscape in the neighborhood and between the two specific reference configurations can be obtained

In Figs. 3.24, 3.25, and 3.26 we show the free-energy landscape at $T = 300$ K calculated from the three simulations with respect to the two RMSD axes. Although the characteristics of these Figures are essentially the same as those in Figs. 3.21, 3.22, and 3.23, the saddle point between the two local-minimum states (RC1 and RC2 states) can be clearly identified. In Fig. 3.26 we labeled the local-minimum states (A_1, A_2, and B) and the transition state (C). In Fig. 3.27 we show representative conformations in the local-minimum states A_1, A_2, and B. The conformations in the local-minimum states A_1 and A_2 have the same backbone hydrogen bonds as in RC1. The local-minimum state B, which has the same as the backbone hydrogen bonds as in RC2, corresponds to the global-minimum free-energy state at $T = 300$ K. The free energy difference between the global-minimum state (B) and the local-minimum state (A_1) (or (A_2)) is about 3 kcal/mol.

The saddle point C in Fig. 3.26 corresponds to the transition state between the global-minimum state (B) and the local-minimum state (A_1) (or (A_2)). The free energy difference between B and C is about 6 kcal/mol and that between A_1 (or A_2) and C is about

Table 3.3: Free energy difference (kcal/mol) among the states.

	A_1	A_2	B	C
A_1	0	0	3	3
A_2	0	0	3	3
B	3	3	0	6
C	3	3	6	0

3 kcal/mol. Because $k_{\rm B}T \approx 0.6$ kcal/mol at $T = 300$ K, these barrier heights are rather high. This is why the usual canonical MD simulation got trapped in the vicinity of the global-minimum state B (RC2-lile state). In Table 3.3 we list the free energy difference among the states. Two representative conformations in C are shown in Fig. 3.28. These structures have a backbone hydrogen bond between CO of Gly-2 and NH of Phe-4. This hydrogen bond in C is common to both RC1 and RC2. The hydrogen bond between NH of Gly-2 and CO of Met-5 which exists in RC1 and that between NH of Gly-2 and CO of Phe-4 which exists in RC2 are missing in C. These structures are thus more extended than reference configurations 1 and 2. Accordingly, the conformations in C are very reasonable as intermediate structures between RC1 and RC2.

In Table 3.4 we list the backbone dihedral angles ϕ and ψ of the conformations in Figs. 3.27 and 3.28. From Tables 3.1 and 3.4 and Figs. 3.27 and 3.28, we can deduce the transition pathways from RC1 to RC2. Note that the major difference between RC1 and RC2 in Tables 3.1 is the value of ψ_4. The two hydrogen bonds (between Gly-2 and Met-5) in RC1 will be simultaneously broken by a large rotation of ψ_4, but this direct pathway is impossible because of high energy barriers. In the following we focus on the relation between the changes of backbone dihedral angle and the formation/breakage of backbone hydrogen bonds in order to elucidate a possible transition pathway from RC1 to RC2. The dihedral angle ϕ_5 first rotates while keeping the hydrogen bonds. This process corresponds to the conformational change from Fig. 3.27(a) to Fig. 3.27(b). The dihedral angles ϕ_2 and ϕ_5 then rotate and the hydrogen bond between NH of Gly-2 and

CO of Met-5 is broken (transition from Fig. 3.27(b) to Fig. 3.28(a)). From Fig. 3.28(a) and Fig. 3.28(b), we also see that the hydrogen bond between CO of Gly-2 and NH of Met-5 is brink of collapse. Finally, the dihedral angle ψ_4 rotates again and the hydrogen bond between NH of Gly-2 and CO of Phe-5 is formed (transition from Fig. 3.28(b) to Fig. 3.27(c)). In summary, we have the following transition pathway: A_1 (Fig. 3.27(a)) $\rightarrow$ A_2 (Fig. 3.27(b)) $\rightarrow$ C (Fig. 3.28(a)) $\rightarrow$ C (Fig. 3.28(b)) $\rightarrow$ B (Fig. 3.27(c)).

Table 3.4: Backbone dihedral angles ϕ and ψ for the structures in Figs. 3.27 and 3.28.

Conformation in Fig. 3.27(a)			Conformation in Fig. 3.27(b)			Conformation in Fig. 3.27(c)		
Residue	Type	Angle	Residue	Type	Angle	Residue	Type	Angle
1	ϕ_1	−157.3°	1	ϕ_1	−145.9°	1	ϕ_1	−131.2°
1	ψ_1	122.7°	1	ψ_1	122.0°	1	ψ_1	142.3°
2	ϕ_2	−130.4°	2	ϕ_2	−126.7°	2	ϕ_2	−171.6°
2	ψ_2	47.2°	2	ψ_2	59.1°	2	ψ_2	64.6°
3	ϕ_3	88.1°	3	ϕ_3	74.3°	3	ϕ_3	90.7°
3	ψ_3	−91.1°	3	ψ_3	−64.7°	3	ψ_3	−59.4°
4	ϕ_4	−95.0°	4	ϕ_4	−83.1°	4	ϕ_4	−118.3°
4	ψ_4	−34.5°	4	ψ_4	−69.9°	4	ψ_4	−167.9°
5	ϕ_5	−74.1°	5	ϕ_5	−136.2°	5	ϕ_5	−82.9°
5	ψ_5	135.9°	5	ψ_5	−169.3°	5	ψ_5	139.1°

Conformation in Fig. 3.28(a)			Conformation in Fig. 3.28(b)		
Residue	Type	Angle	Residue	Type	Angle
1	ϕ_1	−148.5°	1	ϕ_1	−147.5°
1	ψ_1	136.6°	1	ψ_1	163.7°
2	ϕ_2	−166.6°	2	ϕ_2	−174.9°
2	ψ_2	66.1°	2	ψ_2	64.7°
3	ϕ_3	86.1°	3	ϕ_3	73.3°
3	ψ_3	−66.1°	3	ψ_3	−62.5°
4	ϕ_4	−86.9°	4	ϕ_4	−86.0°
4	ψ_4	−68.2°	4	ψ_4	−72.9°
5	ϕ_5	−170.1°	5	ϕ_5	−164.9°
5	ψ_5	151.8°	5	ψ_5	177.2°

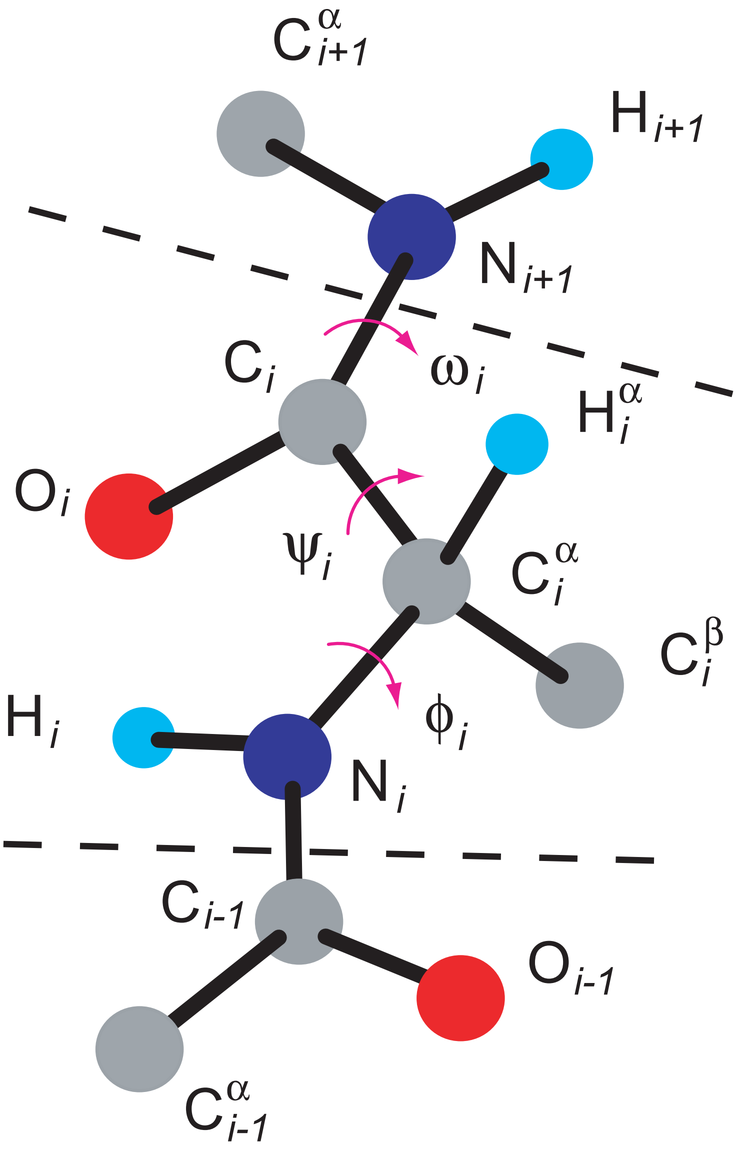

Figure 3.1: Dihedral angles for a polypeptide backbone. ϕ_i, ψ_i, and ω_i are dihedral angles defined by $C_{i-1}, N_i, C^{\alpha}_i, C_i$, $N_i, C^{\alpha}_i, C_i, N_{i+1}$, and $C^{\alpha}_i, C_i, N_{i+1}, C^{\alpha}_{i+1}$, respectively.

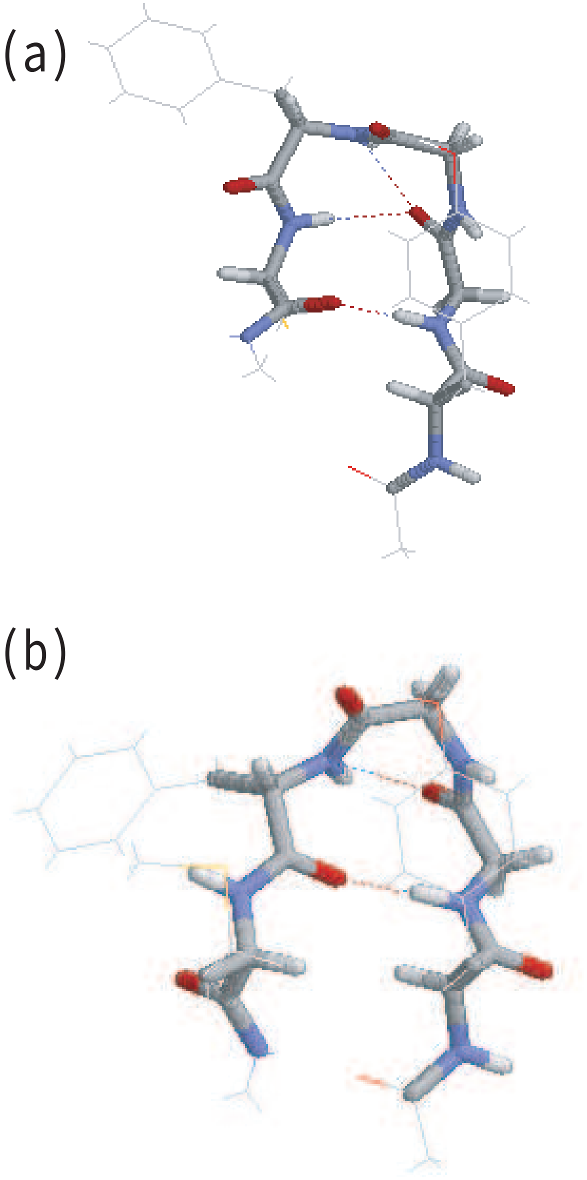

Figure 3.2: (a) Reference configuration 1 and (b) reference configuration 2. The dotted lines denote the hydrogen bonds. The N-terminus and the C-terminus are on the right-hand side and on the left-hand side, respectively. The figures were created with RasMol [22].

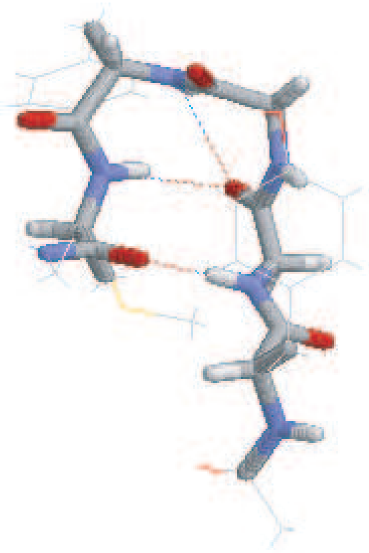

Figure 3.3: The common initial conformation of the usual canonical, multicanonical, multi-overlap MD simulations. See also the caption of Fig. 3.2.

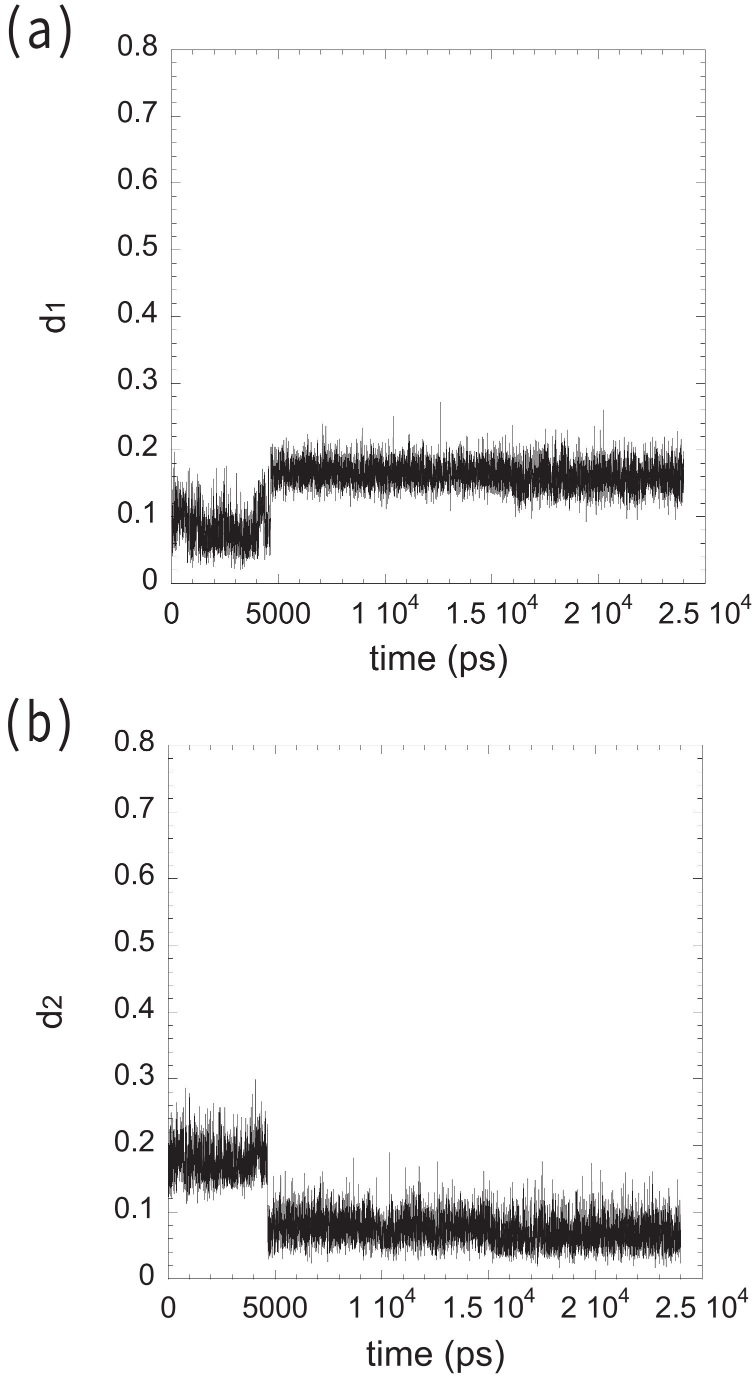

Figure 3.4: The time series of the dihedral-angle distances (a) d_1 and (b) d_2 in the usual canonical MD simulation at $T_0 = 300$ K.

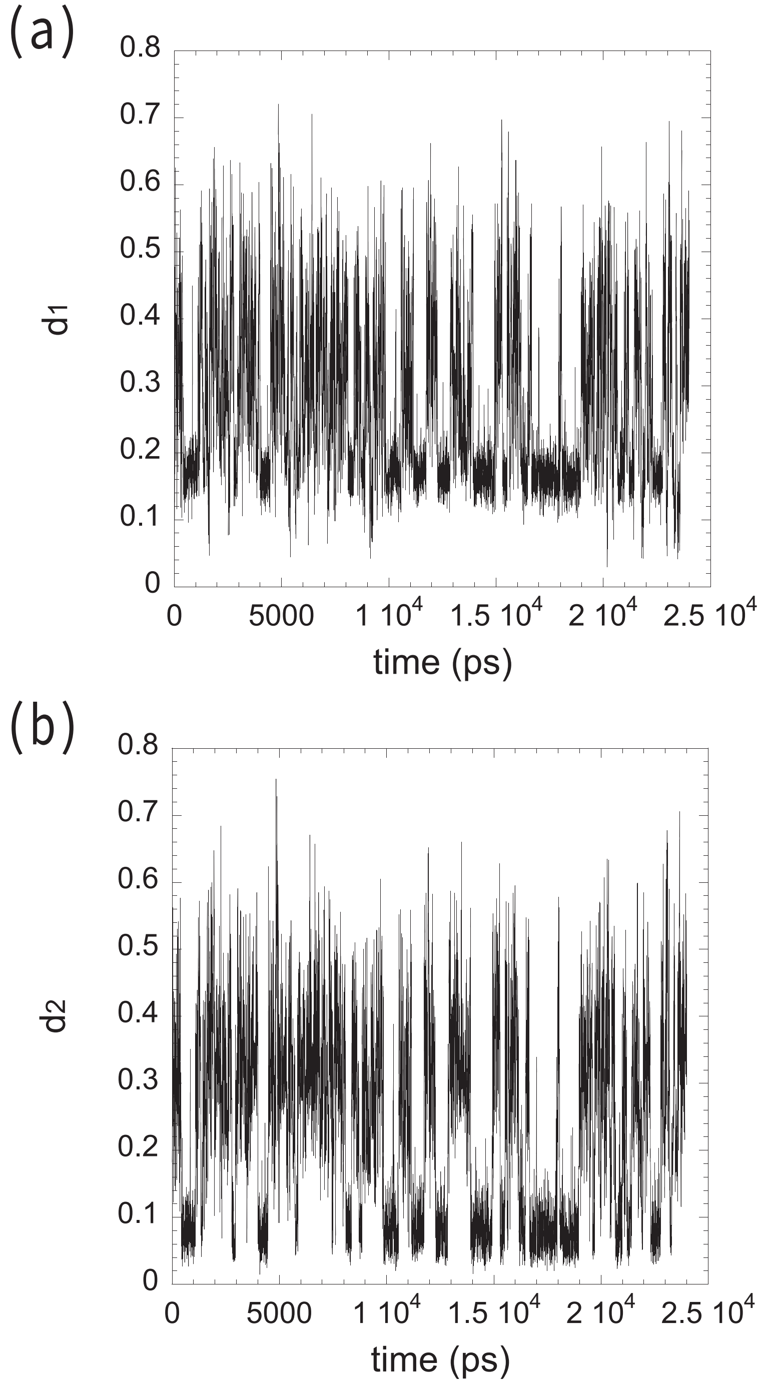

Figure 3.5: The time series of the dihedral-angle distances (a) d_1 and (b) d_2 in the multicanonical MD simulation at $T_0 = 300$ K.

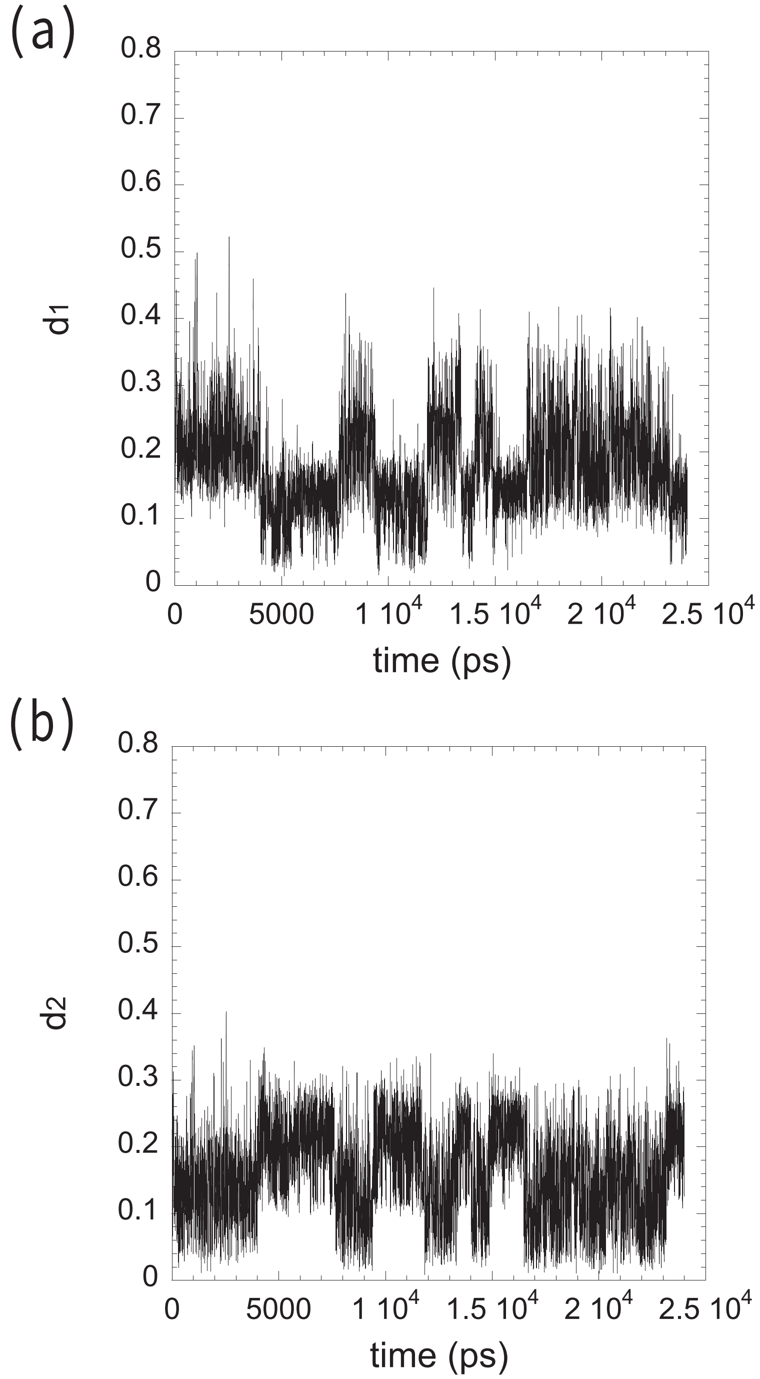

Figure 3.6: The time series of the dihedral-angle distances (a) d_1 and (b) d_2 in the multi-overlap MD simulation at $T_0 = 300$ K.

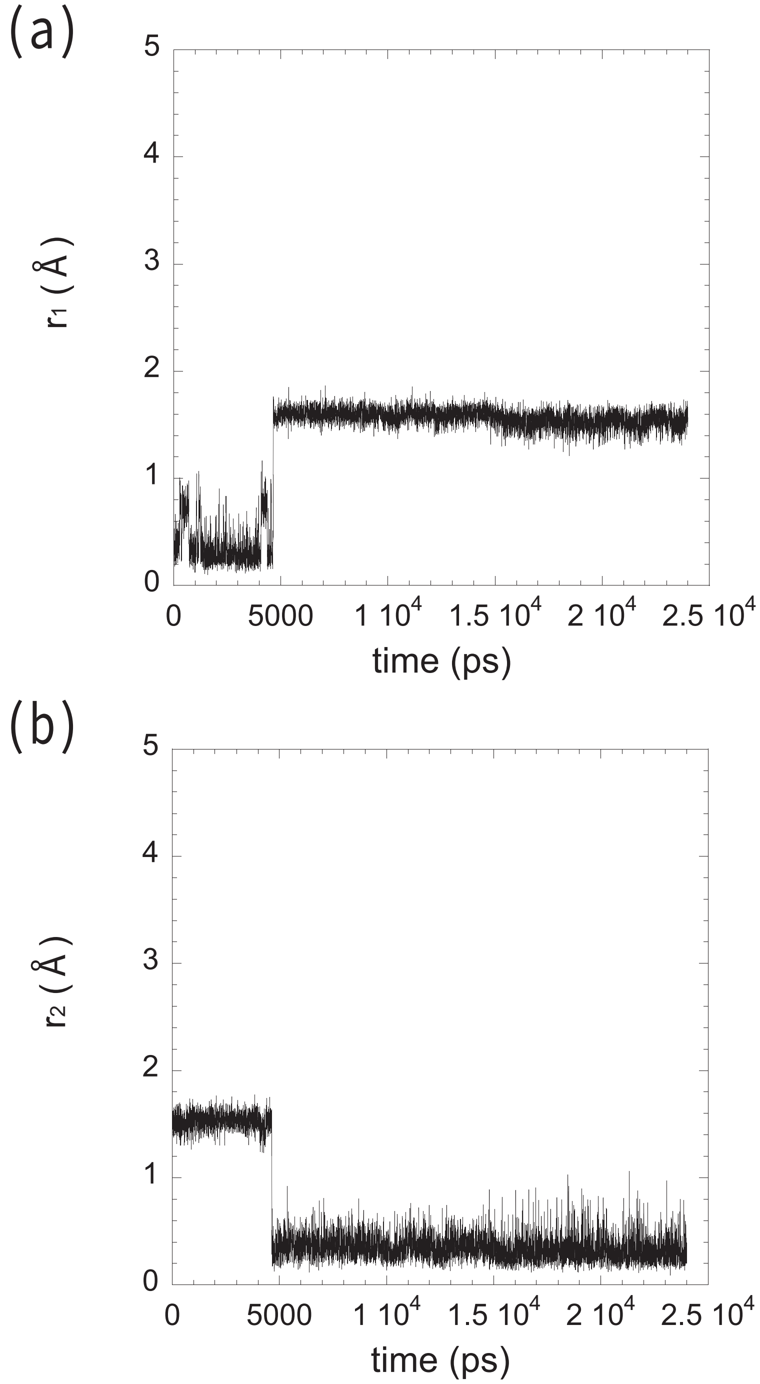

Figure 3.7: The time series of the RMSD (a) r_1 and (b) r_2 in the usual canonical MD simulation at $T_0 = 300$ K.

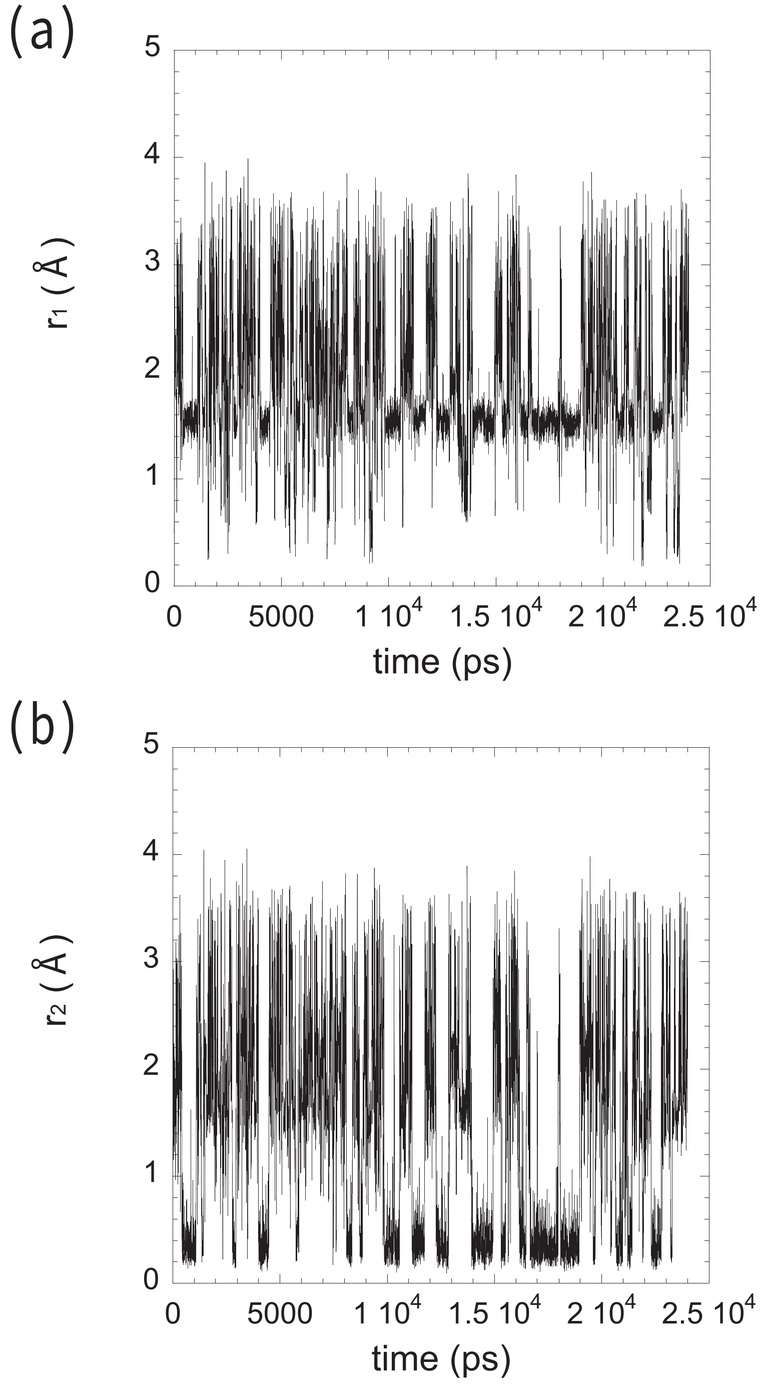

Figure 3.8: The time series of the RMSD (a) r_1 and (b) r_2 in the multicanonical MD simulation at $T_0 = 300$ K.

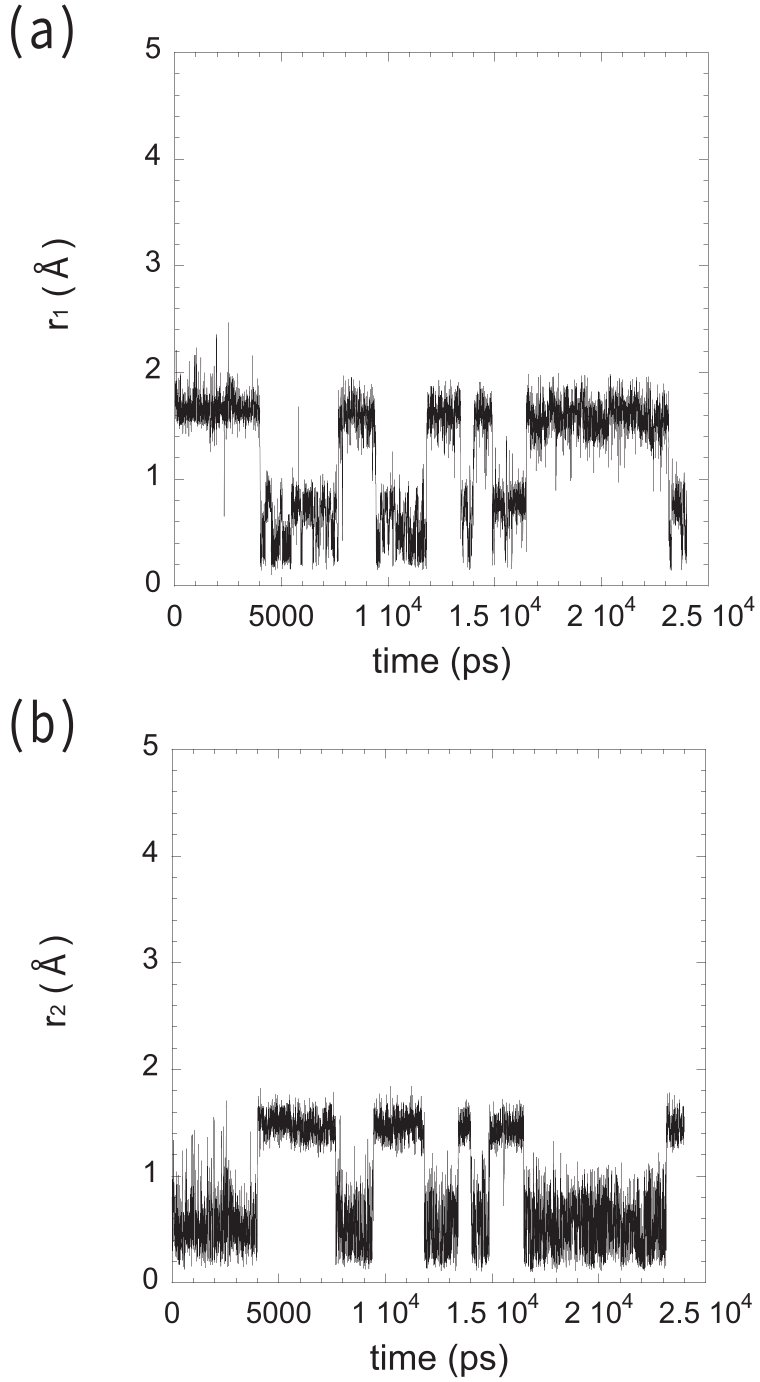

Figure 3.9: The time series of the RMSD (a) r_1 and (b) r_2 in the multi-overlap MD simulation at $T_0 = 300$ K.

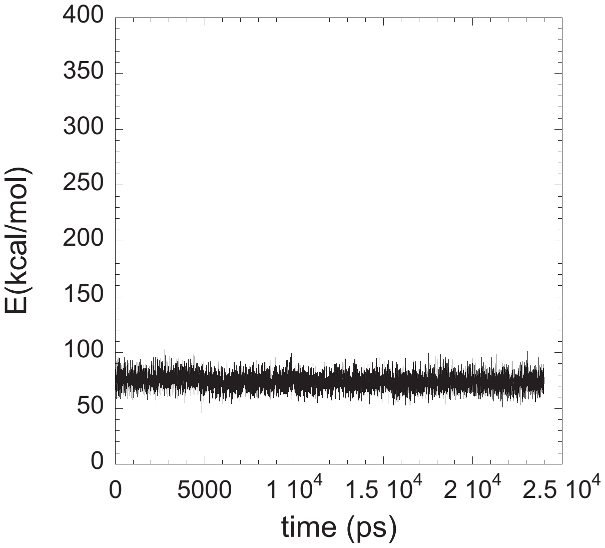

Figure 3.10: The time series of the potential energy E in the usual canonical MD simulation at $T_0 = 300$ K.

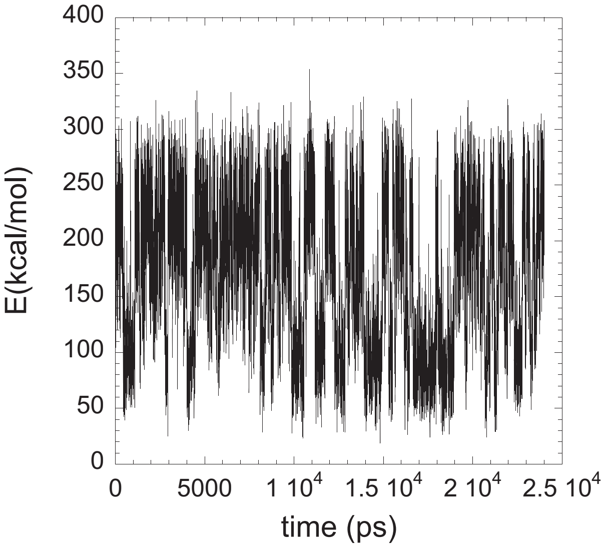

Figure 3.11: The time series of the potential energy E in the multicanonical MD simulation at $T_0 = 300$ K.

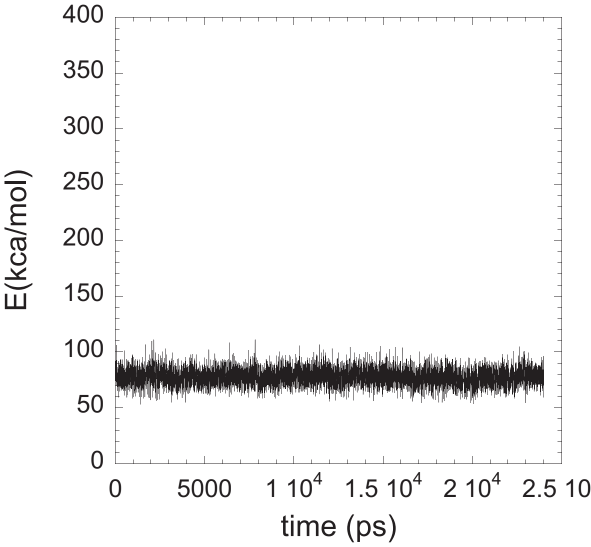

Figure 3.12: The time series of the potential energy E in the multi-overlap MD simulation at $T_0 = 300$ K.

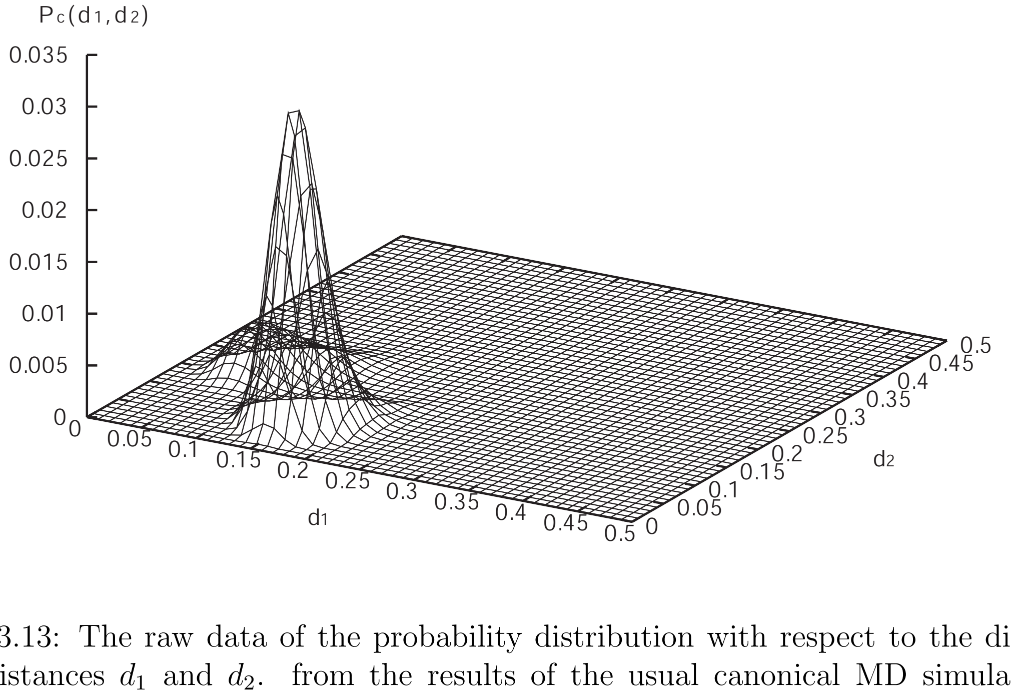

Figure 3.13: The raw data of the probability distribution with respect to the dihedral-angle distances d_1 and d_2. from the results of the usual canonical MD simulation at $T_0 = 300$ K.

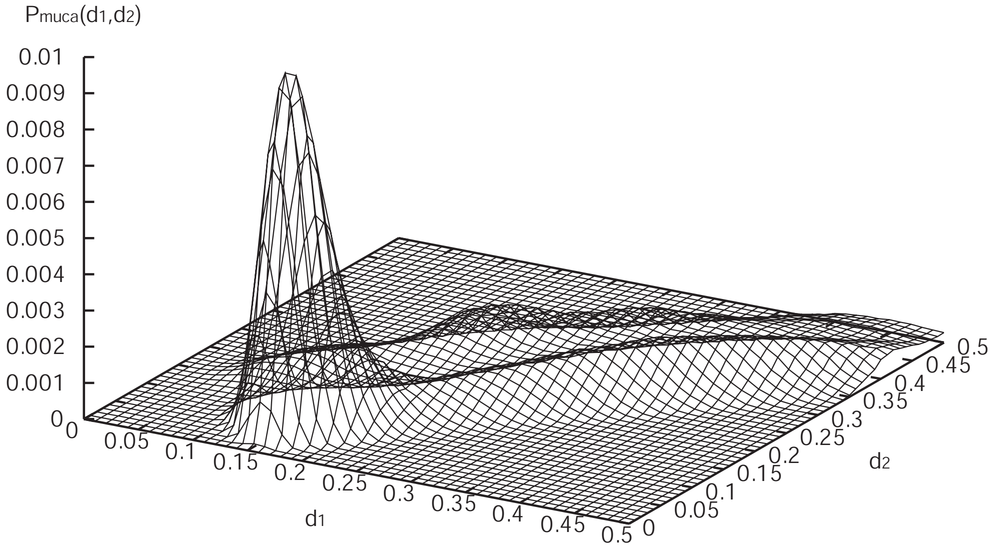

Figure 3.14: The raw data of the probability distribution with respect to the dihedral-angle distances d_1 and d_2. from the results of the multicanonical MD simulation at $T_0 = 300$ K.

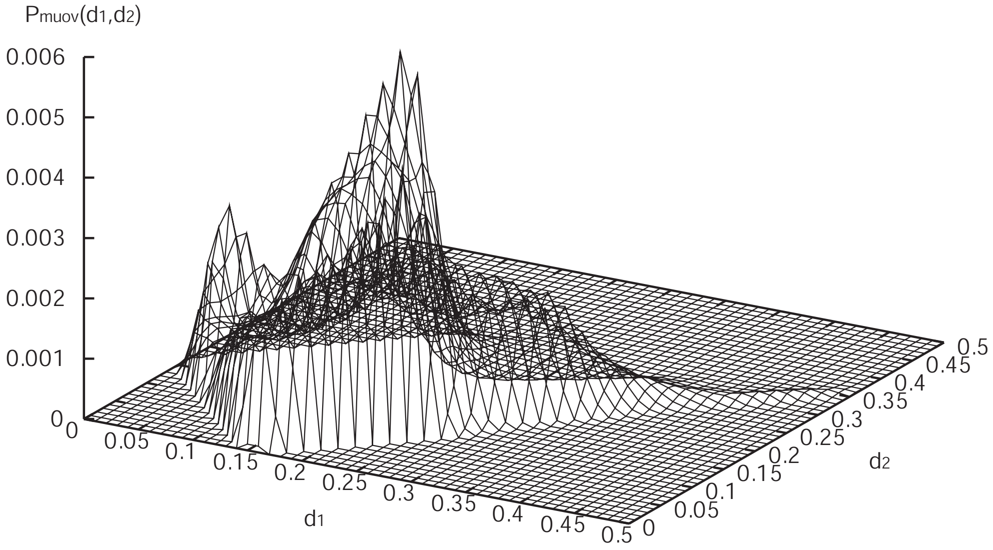

Figure 3.15: The raw data of the probability distribution with respect to the dihedral-angle distances d_1 and d_2. from the results of the multi-overlap MD simulation at $T_0 = 300$ K.

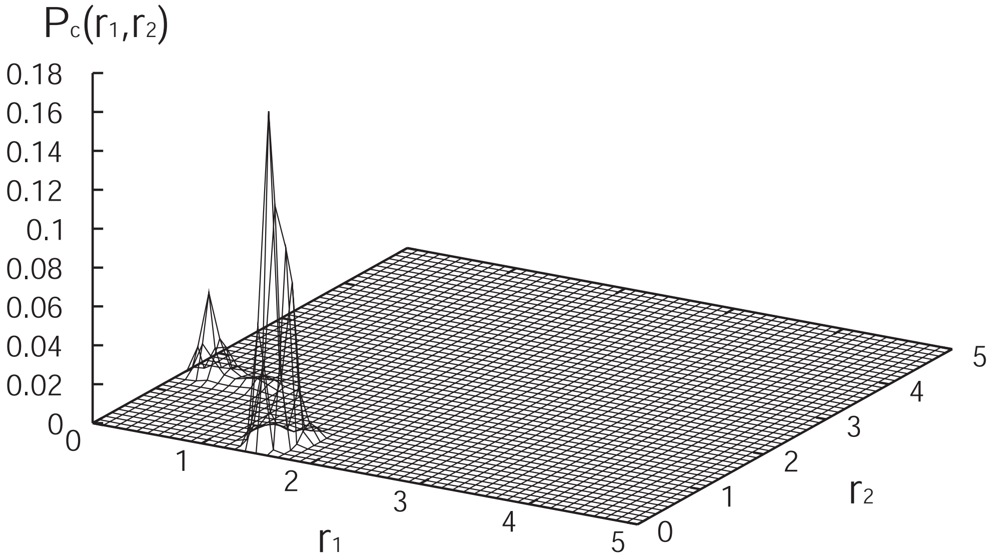

Figure 3.16: The raw data of the probability distribution with respect to the RMSD r_1 and r_2. from the results of the usual canonical MD simulation at $T_0 = 300$ K.

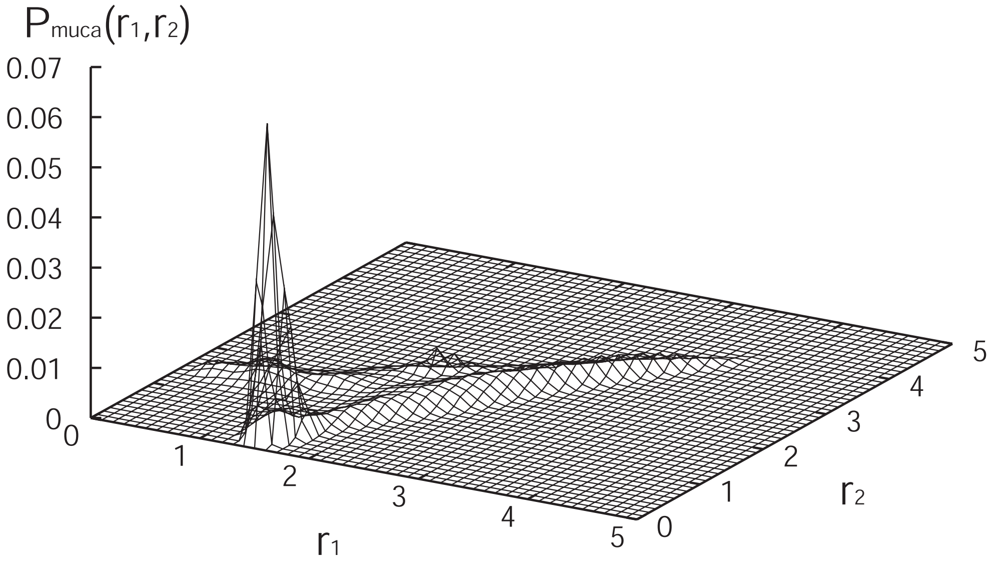

Figure 3.17: The raw data of the probability distribution with respect to the RMSD r_1 and r_2. from the results of the multicanonical MD simulation at $T_0 = 300$ K.

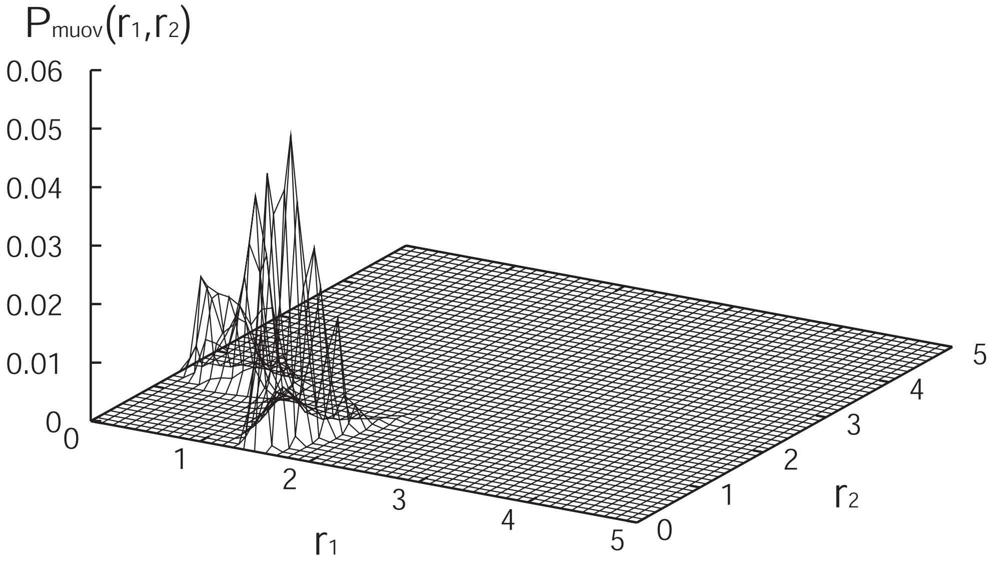

Figure 3.18: The raw data of the probability distribution with respect to the RMSD r_1 and r_2. from the results of the multi-overlap MD simulation at $T_0 = 300$ K.

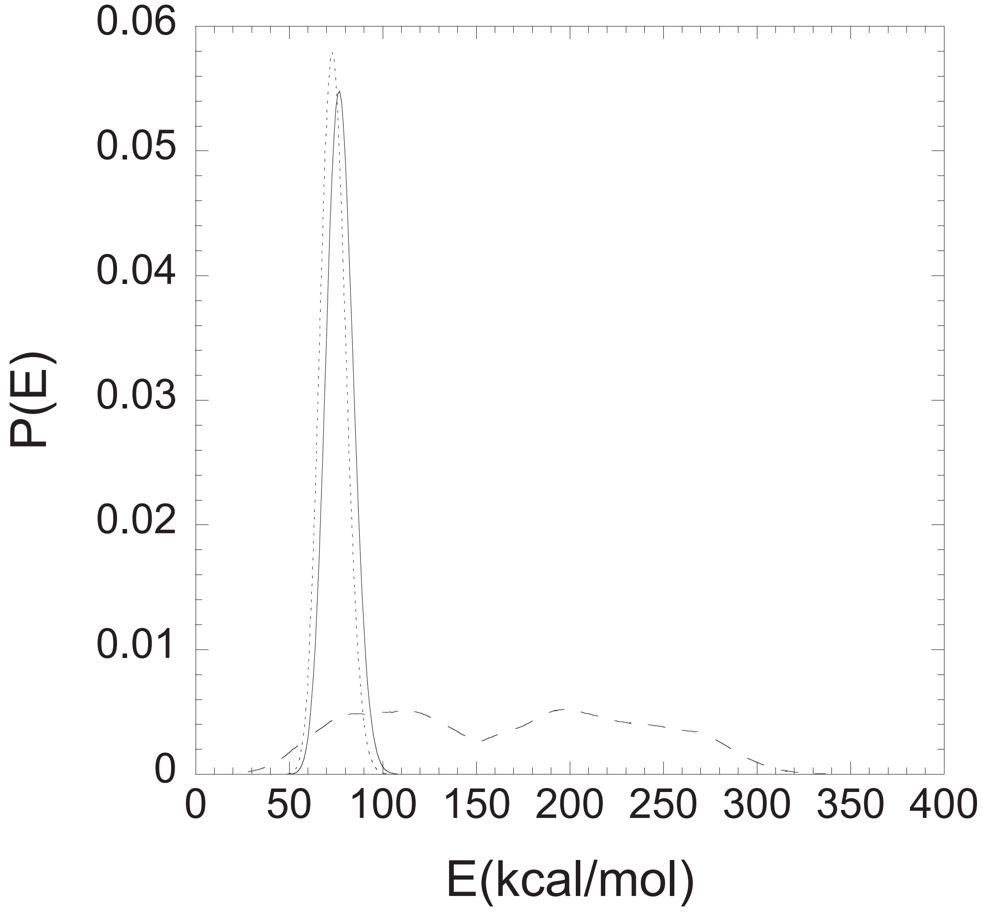

Figure 3.19: The raw data of the probability distribution of the potential energy E. The dotted line, the dashed line, and the solid line show the results from the usual canonical MD simulation at $T_0 = 300$ K, the results from the multicanonical MD simulation, and the results from the multi-overlap MD simulation at $T_0 = 300$ K, respectively.

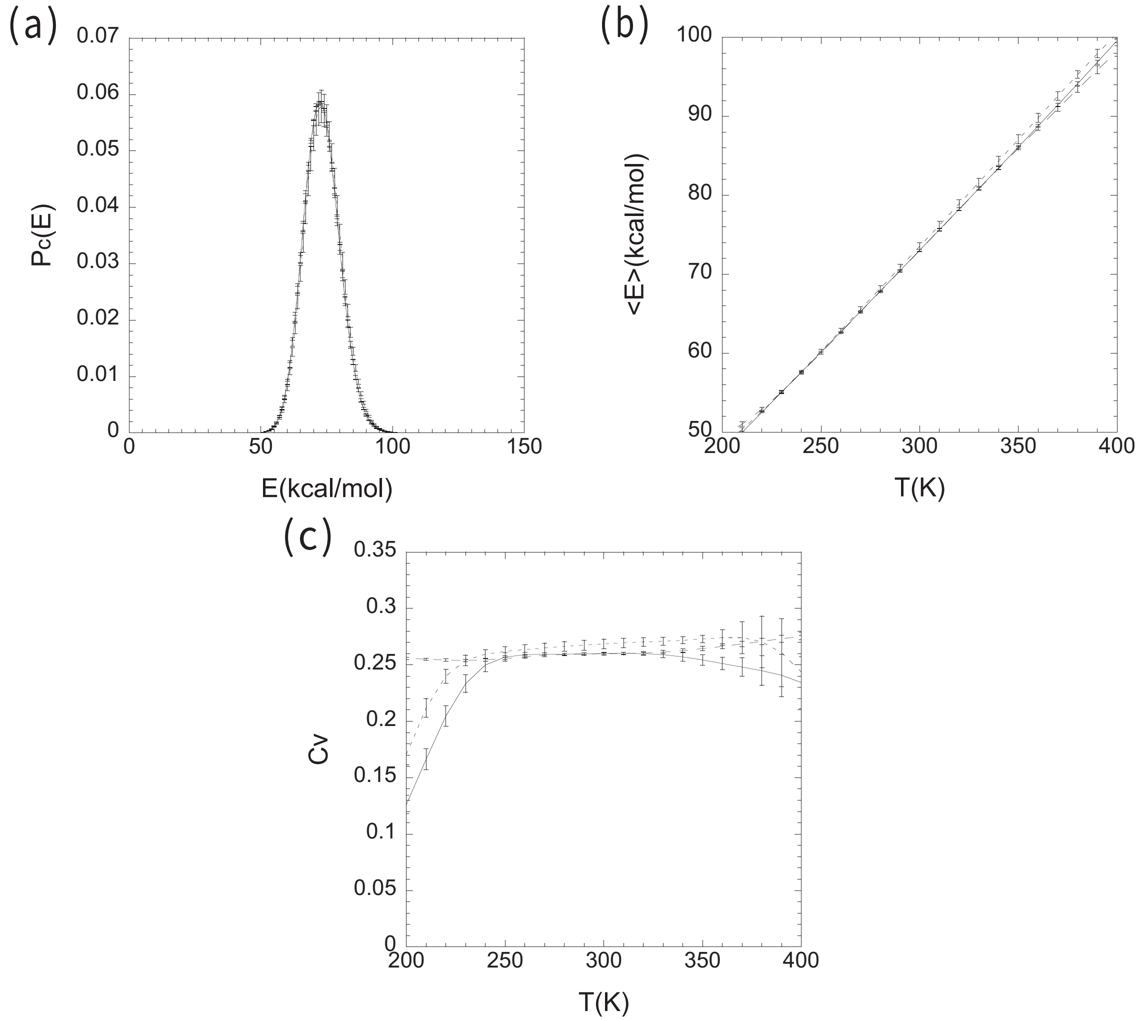

Figure 3.20: (a) Probability distributions of the potential energy at $T = 300$ K, (b) average potential energy as a function of temperature, and (c) specific heat as a function of temperature. These results were calculated from the usual canonical MD simulation (dotted line), the multicanonical MD simulation (dashed line), and the multi-overlap MD simulation (solid line) by the reweighting techniques.

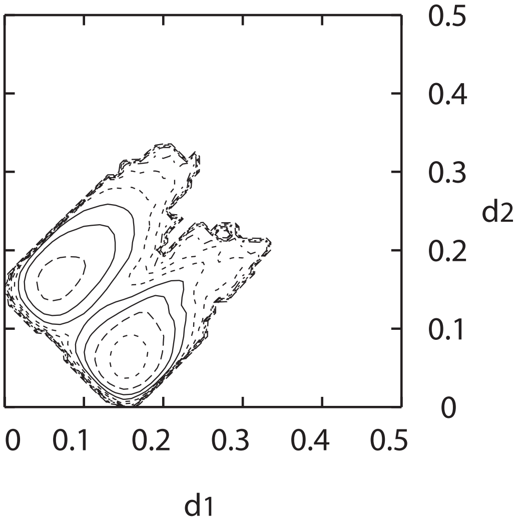

Figure 3.21: The free-energy landscape obtained from the usual canonical MD simulation at $T_0 = 300$ K with the dihedral-angle distance axes d_1, d_2. Contour lines are drawn every 1 kcal/mol.

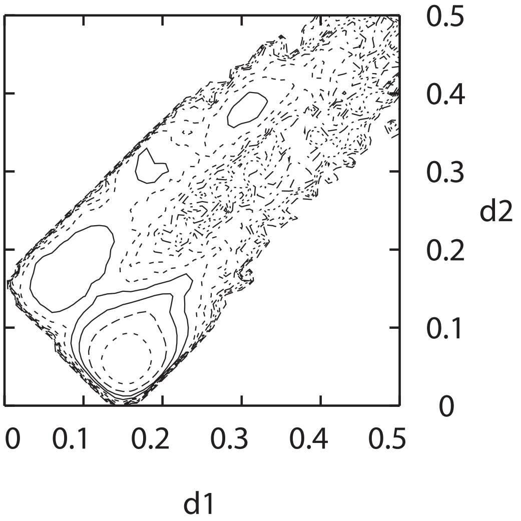

Figure 3.22: The free-energy landscape obtained from the multicanonical MD simulation at $T_0 = 300$ K with the dihedral-angle distance axes d_1, d_2. Contour lines are drawn every 1 kcal/mol.

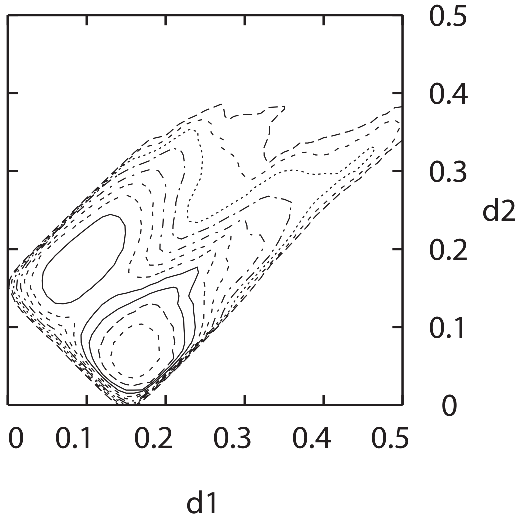

Figure 3.23: The free-energy landscape obtained from the multi-overlap MD simulation at $T_0 = 300$ K with the dihedral-angle distance axes d_1, d_2. Contour lines are drawn every 1 kcal/mol.

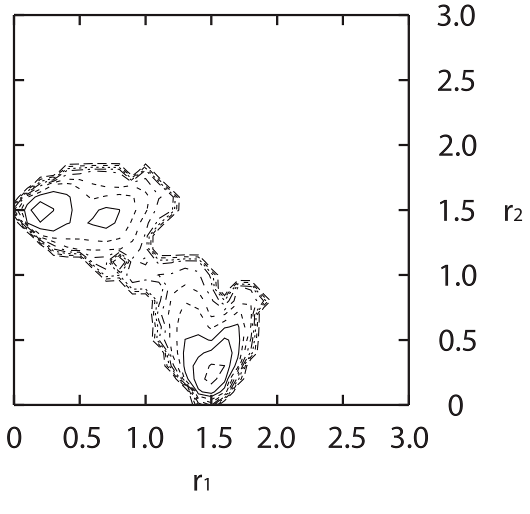

Figure 3.24: The free-energy landscape obtained from the usual canonical MD simulation at $T_0 = 300$ K with the RMSD axes r_1, r_2. Contour lines are drawn every 1 kcal/mol.

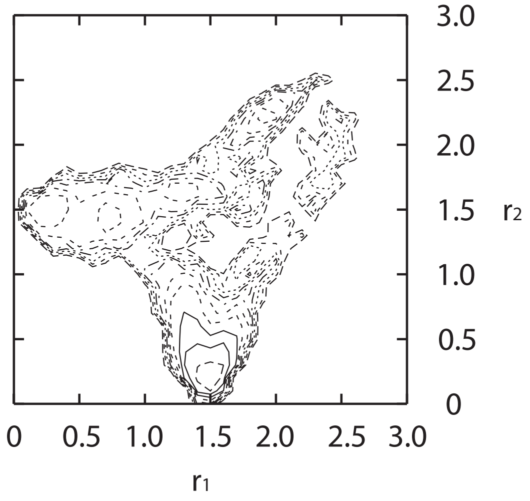

Figure 3.25: The free-energy landscape obtained from the multicanonical MD simulation at $T_0 = 300$ K with the RMSD axes r_1, r_2. Contour lines are drawn every 1 kcal/mol.

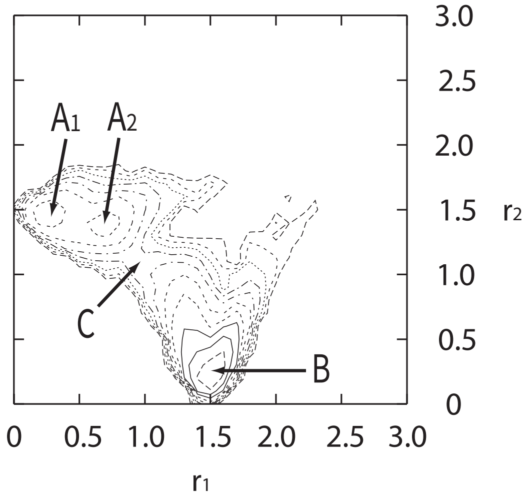

Figure 3.26: The free-energy landscape obtained from the multi-overlap MD simulation at $T_0 = 300$ K with the RMSD axes r_1, r_2. Contour lines are drawn every 1 kcal/mol. The labels A_1, A_2, and B locate the local-minimum states. The label C stands for the saddle point between A_1 (or A_2) and B.

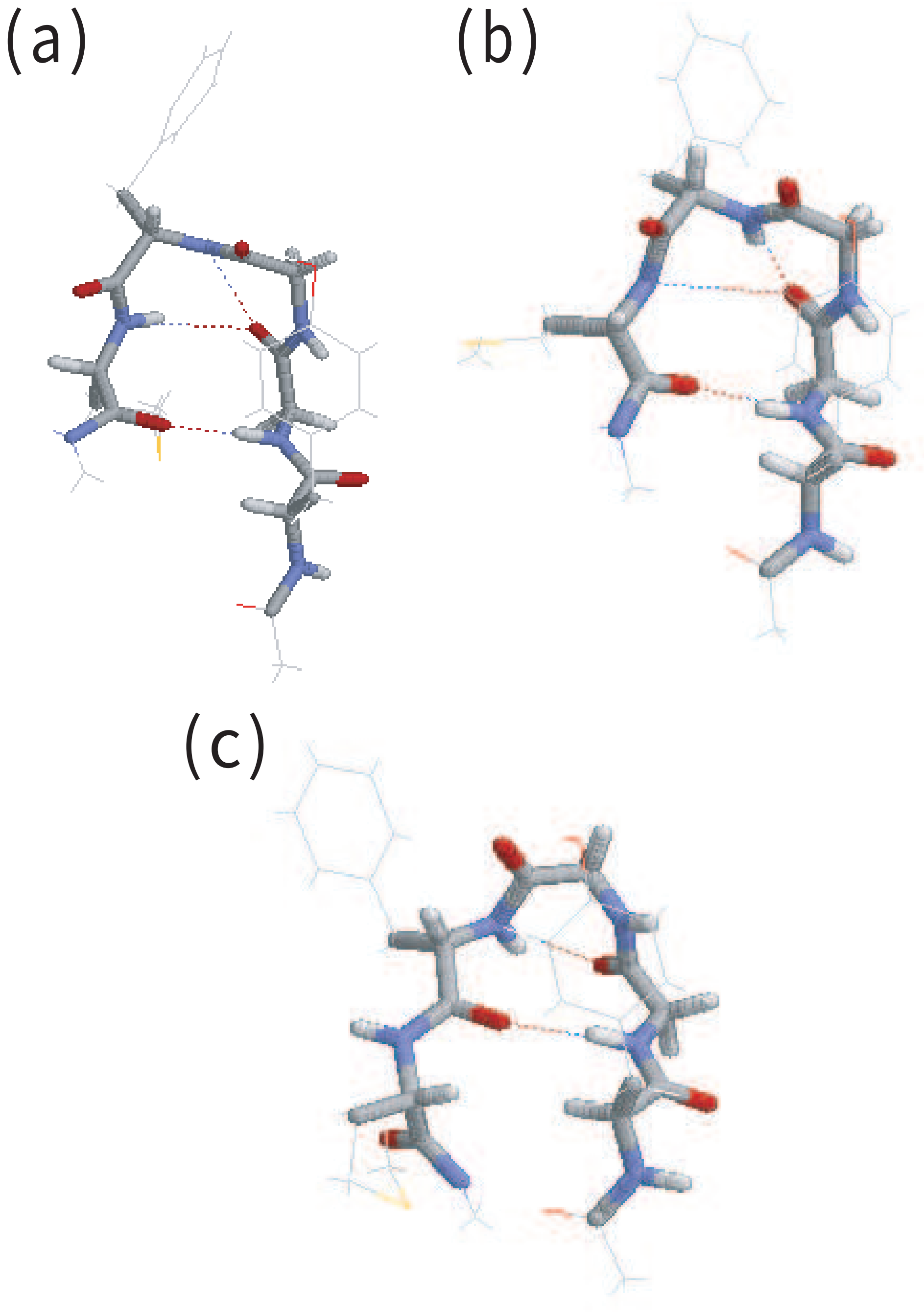

Figure 3.27: (a) The structure in A_1, (b) A_2, and (c) B in Fig 3.26. See also the caption of Fig. 3.2.

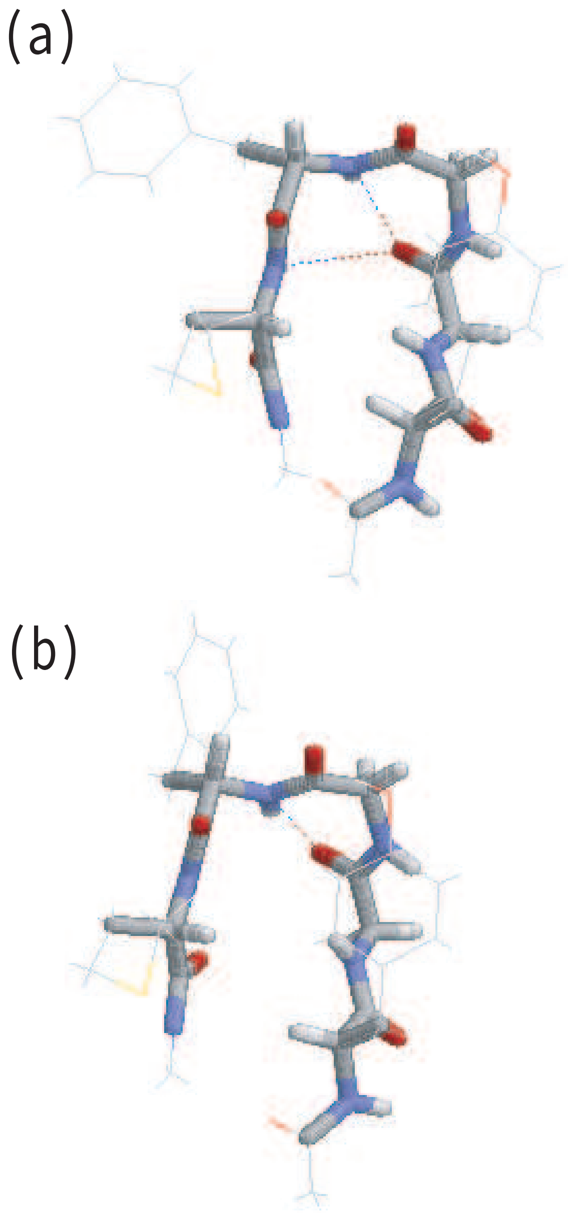

Figure 3.28: Two conformations in the saddle point C in Fig 3.26. See also the caption of Fig. 3.2.

Bibliography

[1] W. G. Hoover, A. J. C. Ladd, and B. Moran, Phys. Rev. Lett. **48**, 1818 (1982).

[2] D. J. Evans, J. Chem. Phys. **78**, 3297 (1983).

[3] S. Nosé, Mol. Phys. **52**, 255 (1984).

[4] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[5] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

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

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

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

[9] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, J. Comp. Chem. **4**, 187 (1983).

[10] D. Brown and J. H. R. Clarke, Mol. Phys. **51**, 1243 (1984).

[11] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).

[12] F. A. Momany, R. F. McGuire, A. W. Burgess and H. A. Scheraga, J. Phys. Chem. **79**, 2361 (1975).

[13] G. Némethy, M. S. Pottle and H. A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).

[14] M. J. Sippl, G. Némethy and H. A. Scheraga, J. Phys. Chem. **88**, 6231 (1984).

[15] A. Mitsutake, U. H. E. Hansmann and Y. Okamoto, J. Mol. Graph. Model. **16**, 226 (1998).

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

[17] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1658 (1989).

[18] M. H. Quenouille, Biometrika **43**, 353 (1956).

[19] R. G. Miller, Biometrika **61**, 1 (1974).

[20] B. A. Berg "Markov Chain Monte Carlo Simulations and Their Statistical Analysis,", (World Scientific, Singapore, 2004).

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

[22] R. A. Sayle and E. J. Milner-White, Trends Biochem. Sci. **20**, 374 (1995).

Chapter 4

Conclusions

In order to understand protein folding and function, it is important that a detailed free-energy landscape of the protein system is obtained. Given the detailed free-energy landscape, we are able to find the folding pathway and the stability of any structures of the protein. However, complex systems such as proteins have many energy local-minimum states, and it is difficult to get an accurate free-energy landscape by the usual canonical-ensemble simulations [1]-[6]. This is because the usual canonical-ensemble simulations will get trapped in states of energy local minima and cannot efficiently sample the conformational space. Furthermore, the conventional generalized-ensemble algorithms [7]-[11], which aim at achieving a wide range sampling in the conformational space, are not used to focus on any specific configurations. Accordingly, the conventional generalized-ensemble simulations cannot estimate the detailed free-energy landscape around the specific configurations and specify the transition states among the specific configurations.

In this thesis, we have proposed an MD version of multi-overlap algorithm [12], which we refer to as the multi-overlap MD algorithm, because it is difficult for the MC version of this method to have efficient sampling in many-particle systems such as proteins in solution. We also generalized this method to a multi-dimensional version. This method is useful to efficiently sample conformations around any reference configurations, while conventional simulation methods are hard to have sufficient sampling in the conformational space around any reference configurations.

We applied the usual canonical, multicanonical, and multi-overlap MD methods to a penta-peptide system of Met-enkephalin in vacuum and showed the effectiveness of the multi-overlap MD method over the canonical and multicanonical MD methods. The multi-overlap MD simulation was performed so that it will realize a random walk between two reference configurations. The canonical MD simulation got trapped in the vicinity of one of the two reference configurations. The multicanonical MD simulation, on the other hand, did not get trapped in states of energy local minima, but it sampled widely only around one of two reference configurations. Finally, the multi-overlap MD simulation did sample the configurational space around both reference configurations. Therefore, we could obtain the detailed free-energy landscape between two reference configurations from the results of the multi-overlap MD simulation. From the free-energy landscape we identified the transition state and deduced the transition pathway between the two local-minimum states. Thus, the multi-overlap MD method is a very powerful tool for studying the free-energy landscape and transition state between two specific configurations.

Some of possible future applications of the present method are as follows. Firstly, we used only the dihedral angles of the backbone as the elements of the dihedral-angle distances. If dihedral angles of side-chains are also included, we will be able to investigate the effects of the side chain conformations on protein folding. Secondly, we studied a peptide in vacuum. We can easily apply the method to a protein in solution, which is a more realistic system. Thirdly, we presented the case with two reference configurations. Because we generalized this method to a multi-dimensional version, it is straightforward to deal with more than two reference configurations

Bibliography

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[2] W. G. Hoover, A. J. C. Ladd, and B. Moran, Phys. Rev. Lett. **48**, 1818 (1982).

[3] D. J. Evans, J. Chem. Phys. **78**, 3297 (1983).

[4] S. Nosé, Mol. Phys. **52**, 255 (1984).

[5] S. Nosé, J. Chem. Phys. **81**, 511 (1984).

[6] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[7] B. A. Berg and T. Neuhaus, Phys. Lett. **B267**, 249 (1991).

[8] B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).

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

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

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

[12] B. A. Berg, H. Noguchi and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).

Appendix A

Leap-Frog Algorithm with the Gaussian Constraint Method

In this thesis we employ the leap-frog algorithm for the numerical integration of the three simulations, namely, the usual canonical, multicanonical, and multi-overlap MD simulations. In this Appendix we explain the leap-frog algorithm in Eq. (2.2) which is the equations of motion for canonical MD simulations with the Gaussian constraint method.

From a Taylor expansion, coordinate vectors $\boldsymbol{q}_i$ at time $t+\Delta t$ and $t-\Delta t$ are given by

$$\boldsymbol{q}_i(t+\Delta t) = \boldsymbol{q}_i(t) + \dot{\boldsymbol{q}}_i(t)\Delta t + \frac{1}{2}\ddot{\boldsymbol{q}}_i(t)(\Delta t)^2 + O\left((\Delta t)^3\right) , \tag{A.1}$$

and

$$\boldsymbol{q}_i(t-\Delta t) = \boldsymbol{q}_i(t) - \dot{\boldsymbol{q}}_i(t)\Delta t + \frac{1}{2}\ddot{\boldsymbol{q}}_i(t)(\Delta t)^2 + O\left((\Delta t)^3\right) . \tag{A.2}$$

From Eqs. (A.1) and (A.2) we obtain

$$\boldsymbol{q}_i(t+\Delta t) = 2\boldsymbol{q}_i(t) - \boldsymbol{q}_i(t-\Delta t) + \ddot{\boldsymbol{q}}_i(t)(\Delta t)^2 + O\left((\Delta t)^4\right) . \tag{A.3}$$

Velocity vectors at time t are obtained from the basic definition of derivative with an error of the order of $(\Delta t)^2$:

$$\dot{\boldsymbol{q}}_i(t) = \frac{\boldsymbol{q}_i(t+\Delta t) - \boldsymbol{q}_i(t-\Delta t)}{2\Delta t} . \tag{A.4}$$

From Eqs. (A.3) and (A.4), we get the following equation at half time step:

$$\dot{\boldsymbol{q}}_i(t+\frac{\Delta t}{2}) = \dot{\boldsymbol{q}}_i(t-\frac{\Delta t}{2}) + \ddot{\boldsymbol{q}}_i(t)\Delta t + O\left((\Delta t)^2\right) . \tag{A.5}$$

From Eqs. (A.3), (A.4), and (A.5), the coordinate vectors at time $t + \Delta t$ are obtained from

$$\boldsymbol{q}_i(t+\Delta t) = \boldsymbol{q}_i(t) + \dot{\boldsymbol{q}}_i(t+\frac{\Delta t}{2})\Delta t + O(\Delta t^3) \ . \tag{A.6}$$

In the leap-frog algorithm, we compute the coordinate and velocity vectors from Eqs. (A.6) and (A.5). Note that the time that gives the velocity vectors is shifted by $\frac{\Delta t}{2}$ from that for the coordinate vectors (and the force). The velocity at the latter time can be simply obtained from

$$\dot{\boldsymbol{q}}_i(t) = \frac{\dot{\boldsymbol{q}}_i(t+\frac{\Delta t}{2}) + \dot{\boldsymbol{q}}_i(t-\frac{\Delta t}{2})}{2} \ . \tag{A.7}$$

For the Gaussian constraint method in Eq. (2.2), this leap-frog algorithm is implemented as follows. Firstly, we make an unconstrained half step using Eq. (A.5):

$$\dot{\boldsymbol{q}}'_i(t) = \dot{\boldsymbol{q}}_i(t-\frac{\Delta t}{2}) + \frac{1}{2m_i}\boldsymbol{F}_i(t)\Delta t \ . \tag{A.8}$$

Secondly, calculate a scaling factor

$$\chi = \left(\frac{T_0}{\mathcal{T}}\right)^{\frac{1}{2}} \ , \tag{A.9}$$

where T_0 is the desired temperature of canonical ensemble and $\mathcal{T}$ is calculated from the unconstrained velocity vectors $\dot{\boldsymbol{q}}'_i(t)$ as follows:

$$\mathcal{T} = \frac{\sum_{i=1}^{N} \frac{m_i}{2}(\dot{\boldsymbol{q}}'_i)^2}{\frac{3}{2}Nk_{\rm B}} \ . \tag{A.10}$$

Finally, we complete the full step using

$$\dot{\boldsymbol{q}}_i(t+\frac{\Delta t}{2}) = (2\chi - 1)\dot{\boldsymbol{q}}_i(t-\frac{\Delta t}{2}) + \frac{\chi}{m_i}\boldsymbol{F}_i(t)\Delta t \ . \tag{A.11}$$

This equation can be derived by substituting Eqs. (2.2) and (A.7) into Eq. (A.5). We remark that $\chi^{-1} = 1 + \frac{1}{2}\zeta_c\Delta t$.

The leap-frog algorithm in Eq. (2.2) with the Gaussian constraint method is performed by these procedure (Eqs. (A.6), (A.8), and (A.11)). We can also carry out the multicanonical or multi-overlap MD simulations by replacing the force vectors $\boldsymbol{F}_i$ with $\boldsymbol{F}_i^{muca}$ or $\boldsymbol{F}_i^{muov}$.

Appendix B

Estimation of Simulation Errors in Reweighting Techniques

By using the reweighting techniques, we can calculate expectation values of physical quantities from the results of multicanonical or multi-overlap simulations. In Sec. 2.5 we introduced the jackknife methods which reduce bias of simulation data and allow us to estimate the correct error bars. In this Appendix we explain how to use the jackknife methods with the reweighting techniques.

We consider the case of multicanonical simulations in Eqs. (2.18) and (2.19). We first divide the entire MD simulation time into N bins with an equal interval. For each bin i of simulation time, we take a histogram of the potential energy, $N_{muca_i}(E)$ (see Fig. B.1). The binned histogram $N_{muca_i}(E)$ then corresponds to a random variable $\{x_i\}$ in Sec. 2.5. The number of samples, N, in Sec. 2.5 is now generalized to be the number of bins. we can write

$$\begin{aligned} N_{muca}(E) &= N \times \overline{N_{muca}(E)} \\ &= N \times \left(\frac{1}{N} \sum_{i=1}^{N} N_{muca_i}(E) \right) \ , \end{aligned} \tag{B.1}$$

we consider the expectation value of a the physical quantity A, $\langle A \rangle_T$, in Eqs. (2.18) and (2.19). This expectation value $\langle A\left(N_{muca}(E)\right)\rangle_T$, or equivalently, $\left\langle A\left(\overline{N_{muca}(E)}\right)\right\rangle_T$, corresponds to the arbitrary function f in Sec. 2.5, and we would like to obtain the quantities corresponding to $f(\hat{x})$ and $\sigma^2(\bar{f})$. For that purpose, we define the jackknife estimators for $\langle A \rangle_T$ (see f_i^J, $\bar{f}^J$, and x_i^J in Eqs. (2.47), (2.48), and (2.49)):

$$\left\langle A^J_{muca_i} \right\rangle_T = \left\langle A\left(N^J_{muca_i}(E)\right)\right\rangle_T$$

$$= \frac{\sum_E A(E) N^J_{muca_i}(E) e^{\beta_0 E_{muca}(E;T_0) - \beta E}}{\sum_E N^J_{muca_i}(E) e^{\beta_0 E_{muca}(E;T_0) - \beta E}} \ , \tag{B.2}$$

and

$$\overline{\langle A^J_{muca} \rangle_T} = \frac{1}{N} \sum_{i=1}^{N} \left\langle A^J_{muca_i} \right\rangle_T \ , \tag{B.3}$$

where

$$N^J_{muca_i}(E) = \frac{1}{N-1} \sum_{k \neq i} N_{muca_k}(E) \ . \tag{B.4}$$

From these jackknife estimators $\left\langle A^J_{muca_i} \right\rangle_T$ and $\overline{\langle A^J_{muca} \rangle_T}$, we can calculate the bias-corrected estimators as follows (see Eqs. (2.51) and (2.52)):

$$\left\langle A^c_{muca_i} \right\rangle_T = N \left\langle A\left(N_{muca}(E)\right)\right\rangle_T - (N-1) \left\langle A^J_{muca_i} \right\rangle_T \ , \tag{B.5}$$

and

$$\overline{\langle A^c_{muca} \rangle_T} = \frac{1}{N} \sum_{i=1}^{N} \left\langle A^c_{muca_i} \right\rangle_T \ . \tag{B.6}$$

Eq. (B.6) gives the bias-corrected estimator for the expectation value of the physical quantity $\left\langle A\left(N_{muca}(E)\right)\right\rangle_T$. Likewise, we can readily calculate the error bars as the square root of the following variance (see Eqs. (2.54) and (2.55)):

$$\sigma^2 \left(\overline{\langle A^c_{muca} \rangle_T} \right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\left\langle A^c_{muca_i} \right\rangle_T - \overline{\langle A^c_{muca} \rangle_T} \right)^2 \ , \tag{B.7}$$

or

$$\sigma^2 \left(\overline{\langle A^c_{muca} \rangle_T} \right) = \frac{N-1}{N} \sum_{i=1}^{N} \left(\left\langle A^J_{muca_i} \right\rangle_T - \overline{\langle A^J_{muca} \rangle_T} \right)^2 \ . \tag{B.8}$$

The case for the multi-overlap simulations essentially follows the same set of equations. The jackknife estimator for $\langle A \rangle_T$ are given by

$$\begin{aligned} \left\langle A^J_{muov_i} \right\rangle_T &= \left\langle A\left(N^J_{muov_i}(d_1, d_2; E) \right) \right\rangle_T \\ &= \frac{\sum_{d_1,d_2,E} A(d_1,d_2;E) N^J_{muov_i}(d_1,d_2;E) e^{\beta_0 E - f(d_1,d_2) - \beta E}}{\sum_{d_1,d_2,E} N^J_{muov_i}(d_1,d_2;E) e^{\beta_0 E - f(d_1,d_2) - \beta E}} \ , \end{aligned} \tag{B.9}$$

and

$$\overline{\langle A^J_{muov} \rangle_T} = \frac{1}{N} \sum_{i=1}^{N} \left\langle A^J_{muov_i} \right\rangle_T \ . \tag{B.10}$$

where

$$N^J_{muov_i}(d_1, d_2; E) = \frac{1}{N-1} \sum_{k \neq i} N_{muov_k}(d_1, d_2; E) \ . \tag{B.11}$$

We can calculate the bias-corrected estimators from

$$\left\langle A^c_{muov_i} \right\rangle_T = N \left\langle A \left(N_{muov}(E) \right) \right\rangle_T - (N-1) \left\langle A^J_{muov_i} \right\rangle_T \ , \tag{B.12}$$

and

$$\overline{\langle A^c_{muov} \rangle_T} = \frac{1}{N} \sum_{i=1}^{N} \left\langle A^c_{muov_i} \right\rangle_T \ . \tag{B.13}$$

The variance of $\left\langle A \left(N_{muov}(E) \right) \right\rangle_T$ is also given by

$$\sigma^2 \left(\overline{\langle A^c_{muov} \rangle_T} \right) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\left\langle A^c_{muov_i} \right\rangle_T - \overline{\langle A^c_{muov} \rangle_T} \right)^2 \ , \tag{B.14}$$

or

$$\sigma^2 \left(\overline{\langle A^c_{muov} \rangle_T} \right) = \frac{N-1}{N} \sum_{i=1}^{N} \left(\left\langle A^J_{muov_i} \right\rangle_T - \overline{\langle A^J_{muov} \rangle_T} \right)^2 \ . \tag{B.15}$$

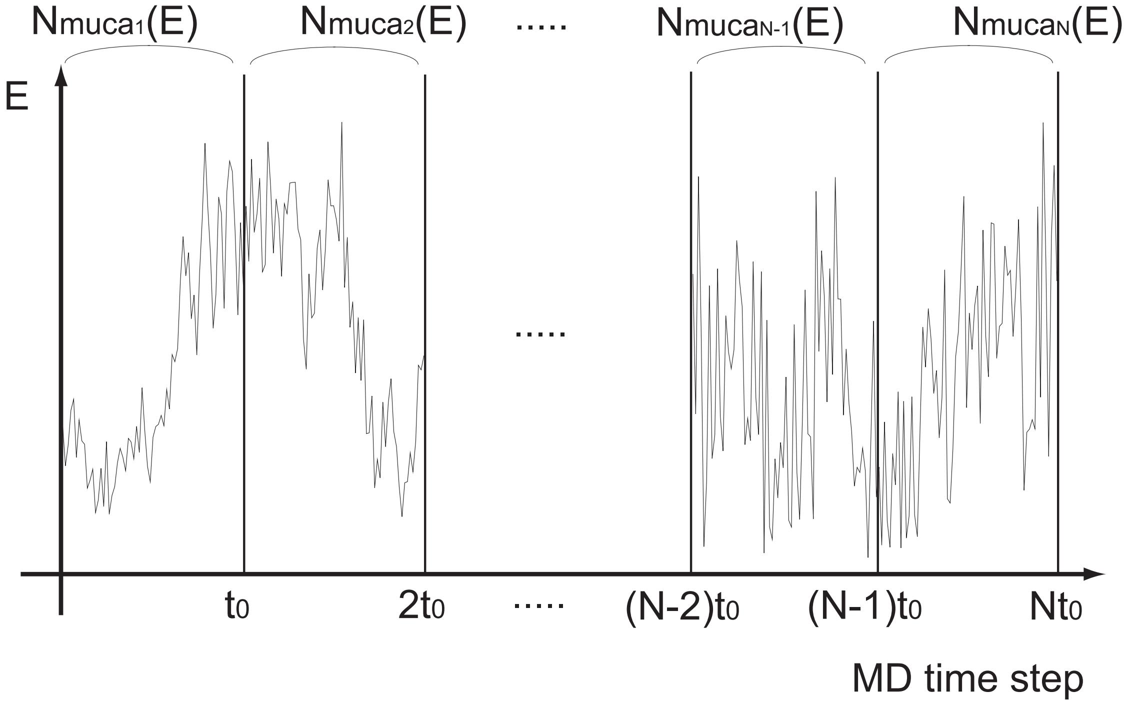

Figure B.1: The definition of the binned histogram $N_{muca_i}(E)$ $(i = 1, \cdots, N)$. The total MD time step is divided into N equal intervals of duration t_0. The histogram $N_{muca_i}(E)$ is taken for the time interval between $(i-1)t_0$ and it_0.