Chapter 1
Protein Folding Simulations by Generalized-Ensemble Algorithms

Takao Yoda, Yuji Sugita, and Yuko Okamoto

Abstract In the protein folding problem, conventional simulations in physical statistical mechanical ensembles, such as the canonical ensemble with fixed temperature, face a great difficulty. This is because there exist a huge number of local-minimum-energy states in the system and the conventional simulations tend to get trapped in these states, giving wrong results. Generalized-ensemble algorithms are based on artificial unphysical ensembles and overcome the above difficulty by performing random walks in potential energy, volume, and other physical quantities or their corresponding conjugate parameters such as temperature, pressure, etc. The advantage of generalized-ensemble simulations lies in the fact that they not only avoid getting trapped in states of energy local minima but also allows the

T. Yoda
Nagahama Institute of Bio-Science and Technology, Tamura, Nagahama, Shiga 526-0829, Japan

Y. Sugita
RIKEN Theoretical Molecular Science Laboratory, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

RIKEN Quantitative Biology Center, 7-1-26 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

RIKEN Advanced Science Institute for Computational Science, 7-1-26 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

Y. Okamoto (✉)
Department of Physics, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602, Japan

Structural Biology Research Center, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602, Japan

Center for Computational Science, Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan

Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
e-mail: okamoto@phys.nagoya-u.ac.jp

K. Han et al. (eds.), *Protein Conformational Dynamics*, Advances in Experimental Medicine and Biology 805, DOI 10.1007/978-3-319-02970-2_1,

calculations of physical quantities as functions of temperature or other parameters from a single simulation run. In this article we review the generalized-ensemble algorithms. Four examples, multicanonical algorithm, replica-exchange method, replica-exchange multicanonical algorithm, and multicanonical replica-exchange method, are described in detail. Examples of their applications to the protein folding problem are presented.

Keywords Generalized-ensemble algorithm • Multicanonical algorithm • Replica-exchange molecular dynamics • Replica-exchange multicanonical algorithm • Multicanonical replica-exchange method • Protein folding

1.1 Introduction

In order to study the protein folding problem, molecular simulation methods such as Monte Carlo (MC) and molecular dynamics (MD) are often used. However, conventional canonical simulations at physically relevant temperatures tend to get trapped in states of energy-local-minima, giving wrong results. A class of simulation methods, which are referred to as the *generalized-ensemble algorithms*, overcome this difficulty (for reviews see, e.g., Refs. [1–5]). In the generalized-ensemble algorithm, each state is weighted by an artificial, non-Boltzmann probability weight factor so that random walks in potential energy, volume, and other physical quantities or their corresponding conjugate parameters such as temperature, pressure, etc. may be realized. The random walks allow the simulation to escape from any energy barrier and to sample much wider conformational space than by conventional methods.

One of effective generalized-ensemble algorithms for molecular simulations is the multicanonical algorithm (MUCA) [6, 7], which was first applied to the protein folding problem in Ref. [8]. In this method, the weight factor is defined to be inversely proportional to the density of states and a free random walk in potential energy space is realized. Another effective generalized-ensemble algorithm is the *replica-exchange method* (REM) [9] (the method is also referred to as *parallel tempering* [10]), which was first applied to the protein folding problem in Ref. [11]. In this method, a number of non-interacting copies (or, replicas) of the original system at different temperatures are simulated independently and exchanged with a specified transition probability. The details of molecular dynamics algorithm for REM, which is referred to as the *replica-exchange molecular dynamics* (REMD), have been worked out in Ref. [12], and this led to a wide application of REMD in the protein and other biomolecular systems. One is naturally led to combine MUCA and REM, and two methods, *replica-exchange multicanonical algorithm* (REMUCA) and *multicanonical replica-exchange method* (MUCAREM), have been developed [13–15]. MUCAREM can be considered to be a special case of the multidimensional (or, multivariable) extension of REM, which we refer to as the *multidimensional replica-exchange method* (MREM) [16]. MREM is now widely used and often referred to as *Hamiltonian replica-exchange method* [17].

In this article, we describe the generalized-ensemble algorithms mentioned above. Namely, we review the four methods: MUCA, REM, REUMCA, and MUCAREM. Examples of the results in which these methods were applied to the protein folding problem are then presented.

1.2 Methods

1.2.1 Multicanonical Algorithm

Let us consider a system of N atoms of mass m_k ($k = 1, \ldots, N$) with their coordinate vectors and momentum vectors denoted by $q = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ and $p = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)$, respectively. The Hamiltonian $H(q,p)$ of the system is the sum of the kinetic energy $K(p)$ and the potential energy $E(q)$:

$$H\,(q, p) = K(p) + E(q), \tag{1.1}$$

where

$$K(p) = \sum_{k=1}^{N} \frac{{\boldsymbol{p}_k}^2}{2m_k}. \tag{1.2}$$

In the canonical ensemble at temperature T each state $x \equiv (q,p)$ with the Hamiltonian $H(q,p)$ is weighted by the Boltzmann factor:

$$W_{\mathrm{B}}\,(x; T) = \exp\left(-\beta H\,(q, p)\right), \tag{1.3}$$

where the inverse temperature β is defined by $\beta = 1/k_{\mathrm{B}}T$ (k_{B} is the Boltzmann constant). The average kinetic energy at temperature T is then given by

$$\left\langle K(p)\right\rangle_T = \left\langle \sum_{k=1}^{N} \frac{{\boldsymbol{p}_k}^2}{2m_k} \right\rangle_T = \frac{3}{2} N k_{\mathrm{B}} T. \tag{1.4}$$

Because the coordinates q and momenta p are decoupled in Eq. (1.1), we can suppress the kinetic energy part and can write the Boltzmann factor as

$$W_{\mathrm{B}}\,(x; T) = W_{\mathrm{B}}\,(E; T) = \exp\left(-\beta E\right). \tag{1.5}$$

The canonical probability distribution of potential energy $P_{\mathrm{NVT}}(E;T)$ is then given by the product of the density of states $n(E)$ and the Boltzmann weight factor $W_{\mathrm{B}}(E;T)$:

$$P_{\mathrm{NVT}}\,(E; T) \propto n(E) W_{\mathrm{B}}\,(E; T). \tag{1.6}$$

Because $n(E)$ is a rapidly increasing function and the Boltzmann factor decreases exponentially, the canonical ensemble yields a bell-shaped distribution of potential energy which has a maximum around the average energy at temperature T. The conventional MC or MD simulations at constant temperature are expected to yield $P_{\mathrm{NVT}}(E;T)$. A MC simulation based on the Metropolis algorithm [18] is performed with the following transition probability from a state x of potential energy E to a state x' of potential energy E':

$$w\left(x \to x'\right) = \min\left(1, \frac{W_{\mathrm{B}}\left(E';T\right)}{W_{\mathrm{B}}\left(E;T\right)}\right) = \min\left(1, \exp\left(-\beta \Delta E\right)\right), \tag{1.7}$$

where

$$\Delta E = E' - E. \tag{1.8}$$

A MD simulation, on the other hand, is based on the following Newton equations of motion:

$$\dot{\boldsymbol{q}}_k = \frac{\boldsymbol{p}_k}{m_k}, \tag{1.9}$$

$$\dot{\boldsymbol{p}}_k = -\frac{\partial E}{\partial \boldsymbol{q}_k} = \boldsymbol{f}_k, \tag{1.10}$$

where $\boldsymbol{f}_k$ is the force acting on the k-th atom ($k = 1, \ldots, N$). This set of equations actually yield the microcanonical ensemble, however, and we have to add a thermostat in order to obtain the canonical ensemble at temperature T. Here, we just follow Nosé's prescription [19, 20], and we have

$$\dot{\boldsymbol{q}}_k = \frac{\boldsymbol{p}_k}{m_k}, \tag{1.11}$$

$$\dot{\boldsymbol{p}}_k = -\frac{\partial E}{\partial \boldsymbol{q}_k} - \frac{\dot{s}}{s}\boldsymbol{p}_k = \boldsymbol{f}_k - \frac{\dot{s}}{s}\boldsymbol{p}_k, \tag{1.12}$$

$$\dot{s} = s\frac{P_s}{Q}, \tag{1.13}$$

$$\dot{P}_s = \sum_{k=1}^{N} \frac{{\boldsymbol{p}_k}^2}{m_k} - 3Nk_{\mathrm{B}}T = 3Nk_{\mathrm{B}}\left(T(t) - T\right), \tag{1.14}$$

where s is Nosé's scaling parameter, P_s is its conjugate momentum, Q is its mass, and the "instantaneous temperature" $T(t)$ is defined by

$$T(t) = \frac{1}{3Nk_{\mathrm{B}}} \sum_{k=1}^{N} \frac{\boldsymbol{p}_k(t)^2}{m_k}. \tag{1.15}$$

However, in practice, it is very difficult to obtain accurate canonical distributions of complex systems at low temperatures by conventional MC or MD simulation methods. This is because simulations at low temperatures tend to get trapped in one or a few of local-minimum-energy states. This difficulty is overcome by, for instance, the generalized-ensemble algorithms, which greatly enhance conformational sampling.

In the multicanonical ensemble [6, 7], on the other hand, each state is weighted by a non-Boltzmann weight factor $W_{\mathrm{MUCA}}(E)$ (which we refer to as the *multicanonical weight factor*) so that a uniform potential energy distribution $P_{\mathrm{MUCA}}(E)$ is obtained:

$$P_{\mathrm{MUCA}}(E) \propto n(E) W_{\mathrm{MUCA}}(E) \equiv \text{const.} \tag{1.16}$$

The flat distribution implies that a free random walk in the potential energy space is realized in this ensemble. This allows the simulation to escape from any local minimum-energy states and to sample the configurational space much more widely than the conventional canonical MC or MD methods.

The definition in Eq. (1.16) implies that the multicanonical weight factor is inversely proportional to the density of states, and we can write it as follows:

$$W_{\mathrm{MUCA}}(E) \equiv \exp\left[-\beta_0 E_{\mathrm{MUCA}}\left(E; T_0\right)\right] = \frac{1}{n(E)}, \tag{1.17}$$

where we have chosen an arbitrary reference temperature, $T_0 = 1/k_{\mathrm{B}}\beta_0$, and the "*multicanonical potential energy*" is defined by

$$E_{\mathrm{MUCA}}\left(E; T_0\right) \equiv k_{\mathrm{B}} T_0 \ln n(E) = T_0 S(E). \tag{1.18}$$

Here, $S(E)$ is the entropy in the microcanonical ensemble. Because the density of states of the system is usually unknown, the multicanonical weight factor has to be determined numerically by iterations of short preliminary runs [6, 7].

A multicanonical MC simulation is performed, for instance, with the usual Metropolis criterion [18]: The transition probability of state x with potential energy E to state x' with potential energy E' is given by

$$\begin{aligned} w\left(x \rightarrow x'\right) &= \min\left(1, \frac{W_{\mathrm{MUCA}}\left(E'\right)}{W_{\mathrm{MUCA}}(E)}\right) = \min\left(1, \frac{n(E)}{n\left(E'\right)}\right) \\ &= \min\left(1, \exp\left(-\beta_0 \Delta E_{\mathrm{MUCA}}\right)\right), \end{aligned} \tag{1.19}$$

where

$$\Delta E_{\mathrm{MUCA}} = E_{\mathrm{MUCA}}\left(E'; T_0\right) - E_{\mathrm{MUCA}}\left(E; T_0\right). \tag{1.20}$$

The MD algorithm in the multicanonical ensemble also naturally follows from Eq. (1.17), in which the regular constant temperature MD simulation (with $T = T_0$) is performed by replacing E by E_{MUCA} in Eq. (1.12) [21, 22]:

$$\dot{\boldsymbol{p}}_k = -\frac{\partial E_{\mathrm{MUCA}}(E; T_0)}{\partial \boldsymbol{q}_k} - \frac{\dot{s}}{s}\boldsymbol{p}_k = \frac{\partial E_{\mathrm{MUCA}}(E; T_0)}{\partial E}\boldsymbol{f}_k - \frac{\dot{s}}{s}\boldsymbol{p}_k. \tag{1.21}$$

From Eq. (1.18) this equation can be rewritten as

$$\dot{\boldsymbol{p}}_k = \frac{T_0}{T(E)}\boldsymbol{f}_k - \frac{\dot{s}}{s}\boldsymbol{p}_k. \tag{1.22}$$

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

$$\left.\frac{\partial S(E)}{\partial E}\right|_{E=E_a} = \frac{1}{T(E_a)}, \tag{1.23}$$

with

$$E_a = {<}E{>}_{T(E_a)}. \tag{1.24}$$

If the exact multicanonical weight factor $W_{\mathrm{MUCA}}(E)$ is known, one can calculate the ensemble averages of any physical quantity A at any temperature T $(= 1/k_{\mathrm{B}}\beta)$ as follows:

$${<}A{>}_T = \frac{\sum_E A(E) P_{\mathrm{NVT}}(E; T)}{\sum_E P_{\mathrm{NVT}}(E; T)} = \frac{\sum_E A(E) n(E) \exp(-\beta E)}{\sum_E n(E) \exp(-\beta E)}, \tag{1.25}$$

where the density of states is given by (see Eq. (1.17))

$$n(E) = \frac{1}{W_{\mathrm{MUCA}}(E)}. \tag{1.26}$$

The summation instead of integration is used in Eq. (1.25), because we often discretize the potential energy E with step size ε ($E = E_{\mathrm{i}}$; i = 1, 2, ...). Here, the explicit form of the physical quantity A should be known as a function of potential energy E. For instance, $A(E) = E$ gives the average potential energy ${<}E{>}_T$ as a function of temperature, and $A(E) = \beta^2 (E - {<}E{>}_{\mathrm{T}})^2$ gives specific heat.

In general, the multicanonical weight factor $W_{\mathrm{MUCA}}(E)$, or the density of states $n(E)$, is not *a priori* known, and one needs its estimator for a numerical simulation.

This estimator is usually obtained from iterations of short trial multicanonical simulations. However, the iterative process can be non-trivial and very tedious for complex systems.

In practice, it is impossible to obtain the ideal multicanonical weight factor with completely uniform potential energy distribution. The question is when to stop the iteration for the weight factor determination. Our criterion for a satisfactory weight factor is that as long as we do get a random walk in potential energy space, the probability distribution $P_{\mathrm{MUCA}}(E)$ does not have to be completely flat with a tolerance of, say, an order of magnitude deviation. In such a case, we usually perform with this weight factor a multicanonical simulation with high statistics (production run) in order to get even better estimate of the density of states. Let $N_{\mathrm{MUCA}}(E)$ be the histogram of potential energy distribution $P_{\mathrm{MUCA}}(E)$ obtained by this production run. The best estimate of the density of states can then be given by the single-histogram reweighting techniques [23] as follows (see the proportionality relation in Eq. (1.16)):

$$n(E) = \frac{N_{\mathrm{MUCA}}(E)}{W_{\mathrm{MUCA}}(E)}. \tag{1.27}$$

By substituting this quantity into Eq. (1.25), one can calculate ensemble averages of physical quantity $A(E)$ as a function of temperature. Moreover, ensemble averages of any physical quantity A (including those that cannot be expressed as functions of potential energy) at any temperature T $(=1/k_{\mathrm{B}}\beta)$ can now be obtained as long as one stores the "trajectory" of configurations (and A) from the production run. Namely, we have

$$<A>_T = \frac{\sum_{k=1}^{n_0} A\left(x(k)\right) W_{\mathrm{MUCA}}^{-1}\left(E\left(x(k)\right)\right) \exp\left[-\beta E\left(x(k)\right)\right]}{\sum_{k=1}^{n_0} W_{\mathrm{MUCA}}^{-1}\left(E\left(x(k)\right)\right) \exp\left[-\beta E\left(x(k)\right)\right]}, \tag{1.28}$$

where $x(k)$ is the configuration at the k-th MC (or MD) step and n_0 is the total number of configurations stored. Note that when A is a function of E, Eq. (1.28) reduces to Eq. (1.25) where the density of states is given by Eq. (1.27).

Equations (1.25) and (1.28) or any other equations which involve summations of exponential functions often encounter with numerical difficulties such as overflows. These can be overcome by using, for instance, the following equation [24, 25]: For $C = A + B$ (with $A > 0$ and $B > 0$) we have

$$\begin{aligned} \ln C &= \ln\left[\max\left(A, B\right)\left(1 + \frac{\min\left(A, B\right)}{\max\left(A, B\right)}\right)\right] \\ &= \max\left(\ln A, \ln B\right) + \ln\left\{1 + \exp\left[\min\left(\ln A, \ln B\right) - \max\left(\ln A, \ln B\right)\right]\right\}. \end{aligned} \tag{1.29}$$

1.2.2 Replica-Exchange Method

The *replica-exchange method* (REM) is another effective generalized-ensemble algorithm. The system for REM consists of M *non-interacting* copies (or, replicas) of the original system in the canonical ensemble at M different temperatures $T_m (m = 1, \ldots, M)$. We arrange the replicas so that there is always exactly one replica at each temperature. Then there exists a one-to-one correspondence between replicas and temperatures; the label $i(=1, \ldots, M)$ for replicas is a permutation of the label $m(=1, \ldots, M)$ for temperatures, and vice versa:

$$\begin{cases} i = i(m) \equiv f(m), \\ m = m(i) \equiv f^{-1}(i), \end{cases} \tag{1.30}$$

where $f(m)$ is a permutation function of m and $f^{-1}(i)$ is its inverse.

Let $X = \{x_1^{[i(1)]}, \ldots, x_M^{[i(M)]}\} = \{x_{m(1)}^{[1]}, \ldots, x_{m(M)}^{[M]}\}$ stand for a "state" in this generalized ensemble. Each "substate" $x_m^{[i]}$ is specified by the coordinates $q^{[i]}$ and momenta $p^{[i]}$ of N atoms in replica i at temperature T_m:

$$x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m. \tag{1.31}$$

Because the replicas are non-interacting, the weight factor for the state X in this generalized ensemble is given by the product of Boltzmann factors for each replica (or at each temperature):

$$\begin{aligned} W_{\text{REM}}(X) &= \prod_{i=1}^{M} \exp\left\{-\beta_{m(i)} H\left(q^{[i]}, p^{[i]}\right)\right\} = \prod_{m=1}^{M} \exp\left\{-\beta_m H\left(q^{[i(m)]}, p^{[i(m)]}\right)\right\} \\ &= \exp\left\{-\sum_{i=1}^{M} \beta_{m(i)} H\left(q^{[i]}, p^{[i]}\right)\right\} = \exp\left\{-\sum_{m=1}^{M} \beta_m H\left(q^{[i(m)]}, p^{[i(m)]}\right)\right\}, \end{aligned} \tag{1.32}$$

where $i(m)$ and $m(i)$ are the permutation functions in Eq. (1.30).

We now consider exchanging a pair of replicas in this ensemble. Suppose we exchange replicas i and j which are at temperatures T_m and T_n, respectively:

$$X = \left\{\ldots, x_m^{[i]}, \ldots, x_n^{[j]}, \ldots\right\} \rightarrow X' = \left\{\ldots, x_m^{[j]'}, \ldots, x_n^{[i]'}, \ldots\right\}. \tag{1.33}$$

The exchange of replicas can be written in more detail as

$$\begin{cases} x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m \rightarrow x_m^{[j]'} \equiv \left(q^{[j]}, p^{[j]'}\right)_m, \\ x_n^{[j]} \equiv \left(q^{[j]}, p^{[j]}\right)_n \rightarrow x_n^{[i]'} \equiv \left(q^{[i]}, p^{[i]'}\right)_n, \end{cases} \tag{1.34}$$

where the definitions for $p^{[i]'}$ and $p^{[j]'}$ will be given below.

In the original implementation of the *replica-exchange method* (REM) [9], Monte Carlo algorithm was used, and only the coordinates q (and the potential energy function $E(q)$) had to be taken into account. In molecular dynamics algorithm, on the other hand, we also have to deal with the momenta p. We proposed the following momentum assignment in Eq. (1.34) [12]:

$$\begin{cases} p^{[i]\prime} \equiv \sqrt{\frac{T_n}{T_m}}\, p^{[i]}, \\ p^{[j]\prime} \equiv \sqrt{\frac{T_m}{T_n}}\, p^{[j]}, \end{cases} \tag{1.35}$$

which we believe is the simplest and the most natural. This assignment means that we just rescale uniformly the velocities of all the atoms in the replicas by the square root of the ratio of the two temperatures so that the temperature condition in Eq. (1.4) may be satisfied immediately after replica exchange is accepted. We remark that similar momentum rescaling formulae for various constant-temperature algorithms have been worked out in Ref. [26].

The transition probability of this replica-exchange process is given by the usual Metropolis criterion:

$$w\left(X \to X'\right) \equiv w\left(x_m^{[i]} \middle| x_n^{[j]}\right) = \min\left(1, \frac{W_{\text{REM}}\left(X'\right)}{W_{\text{REM}}(X)}\right) = \min\left(1, \exp\left(-\varDelta\right)\right), \tag{1.36}$$

where in the second expression (i.e., $w(x_m^{[i]}|x_n^{[j]})$) we explicitly wrote the pair of replicas (and temperatures) to be exchanged. From Eqs. (1.1), (1.2), (1.32), and (1.35), we have

$$\varDelta = \beta_m\left(E\left(q^{[j]}\right) - E\left(q^{[i]}\right)\right) - \beta_n\left(E\left(q^{[j]}\right) - E\left(q^{[i]}\right)\right) \tag{1.37}$$

$$= \left(\beta_m - \beta_n\right)\left(E\left(q^{[j]}\right) - E\left(q^{[i]}\right)\right). \tag{1.38}$$

Note that after introducing the momentum rescaling in Eq. (1.35), we have the same Metropolis criterion for replica exchanges, i.e., Eqs. (1.36) and (1.38), for both MC and MD versions.

Without loss of generality we can assume that $T_1 < T_2 < \ldots < T_M$. The lowest temperature T_1 should be sufficiently low so that the simulation can explore the experimentally relevant temperature region, and the highest temperature T_M should be sufficiently high so that no trapping in an energy-local-minimum state occurs. A REM simulation is then realized by alternately performing the following two steps:

1. Each replica in canonical ensemble of the fixed temperature is simulated *simultaneously* and *independently* for a certain MC or MD steps.

2. A pair of replicas at neighboring temperatures, say, $x_m^{[i]}$ and $x_{m+1}^{[j]}$, are exchanged with the probability $w(x_m^{[i]}|x_{m+1}^{[j]})$ in Eq. (1.36).

A random walk in "temperature space" is realized for each replica, which in turn induces a random walk in potential energy space. This alleviates the problem of getting trapped in states of energy local minima.

After a long production run of a replica-exchange simulation, the canonical expectation value of a physical quantity A at temperature $T_m (m = 1, \ldots, M)$ can be calculated by the usual arithmetic mean:

$$\langle A \rangle_{T_m} = \frac{1}{n_m} \sum_{k=1}^{n_m} A\left(x_m(k)\right), \tag{1.39}$$

where $x_m(k)(k = 1, \ldots, n_m)$ are the configurations obtained at temperature T_m and n_m is the total number of measurements made at $T = T_m$. The expectation value at any intermediate temperature T $(= 1/k_{\rm B}\beta)$ can also be obtained from Eq. (1.25), where the density of states $n(E)$ in Eq. (1.25) is now given by the multiple-histogram reweighting techniques, or, the weighted histogram analysis method (WHAM) [27, 28] as follows. Let $N_m(E)$ and n_m be respectively the potential-energy histogram and the total number of samples obtained at temperature $T_m = 1/k_{\rm B}\beta_m (m = 1, \ldots, M)$. The best estimate of the density of states is then given by

$$n(E) = \frac{\displaystyle\sum_{m=1}^{M} N_m(E)}{\displaystyle\sum_{m=1}^{M} n_m \exp\left(f_m - \beta_m E\right)}, \tag{1.40}$$

where we have for each $m(=1, \ldots, M)$

$$\exp\left(-f_m\right) = \sum_{E} n(E) \exp\left(-\beta_m E\right). \tag{1.41}$$

Note that Eqs. (1.40) and (1.41) are solved self-consistently by iteration [27, 28] to obtain the density of states $n(E)$ and the dimensionless Helmholtz free energy f_m. Namely, we can set all the $f_m (m = 1, \ldots, M)$ to, e.g., zero initially. We then use Eq. (1.40) to obtain $n(E)$, which is substituted into Eq. (1.41) to obtain next values of f_m, and so on.

Moreover, ensemble averages of any physical quantity A (including those that cannot be expressed as functions of potential energy) at any temperature T $(= 1/k_{\rm B}\beta)$ can now be obtained from the "trajectory" of configurations of the production run. Namely, we first obtain $f_m (m = 1, \ldots, M)$ by solving Eqs. (1.40) and (1.41) self-consistently, and then we have [14]

$$<A>_T = \frac{\sum_{m=1}^{M}\sum_{k=1}^{n_m} A\left(x_m(k)\right) \frac{1}{\sum_{l=1}^{M} n_l \exp\left[f_l - \beta_l E\left(x_m(k)\right)\right]} \exp\left[-\beta E\left(x_m(k)\right)\right]}{\sum_{m=1}^{M}\sum_{k=1}^{n_m} \frac{1}{\sum_{l=1}^{M} n_l \exp\left[f_l - \beta_l E\left(x_m(k)\right)\right]} \exp\left[-\beta E\left(x_m(k)\right)\right]}, \tag{1.42}$$

where $x_m(k)(k = 1, \ldots, n_m)$ are the configurations obtained at temperature T_m.

1.2.3 Replica-Exchange Multicanonical Algorithm and Multicanonical Replica-Exchange Method

The *replica-exchange multicanonical algorithm* (REMUCA) [13–15] overcomes both the difficulties of MUCA (the multicanonical weight factor determination is non-trivial) and REM (a lot of replicas, or computation time, is required). In REMUCA we first perform a short REM simulation (with M replicas) to determine the multicanonical weight factor and then perform with this weight factor a regular multicanonical simulation with high statistics. The first step is accomplished by the multiple-histogram reweighting techniques. Let $N_m(E)$ and n_m be respectively the potential-energy histogram and the total number of samples obtained at temperature T_m $(= 1/k_B\beta_m)$ of the REM run. The density of states $n(E)$ is then given by solving Eqs. (1.40) and (1.41) self-consistently by iteration.

Once the estimate of the density of states is obtained, the multicanonical weight factor can be directly determined from Eq. (1.17) (see also Eq. (1.18)). Actually, the density of states $n(E)$ and the multicanonical potential energy, $E_{\mathrm{MUCA}}(E;T_0)$, thus determined are only reliable in the following range:

$$E_1 \leq E \leq E_M, \tag{1.43}$$

where

$$\begin{cases} E_1 = <E>_{T_1}, \\ E_M = <E>_{T_M}, \end{cases} \tag{1.44}$$

and T_1 and T_M are respectively the lowest and the highest temperatures used in the REM run. Outside this range we extrapolate the multicanonical potential energy linearly [13]:

$$\mathcal{E}^{\{0\}}_{\mathrm{MUCA}}(E) \equiv \begin{cases} \left.\frac{\partial E_{\mathrm{MUCA}}(E;T_0)}{\partial E}\right|_{E=E_1} (E-E_1) + E_{\mathrm{MUCA}}(E_1;T_0), & \text{for } E < E_1, \\ E_{\mathrm{MUCA}}\left(E;T_0\right), & \text{for } E_1 \le E \le E_M, \\ \left.\frac{\partial E_{\mathrm{MUCA}}(E;T_0)}{\partial E}\right|_{E=E_M} (E-E_M) + E_{\mathrm{MUCA}}\left(E_M;T_0\right), & \text{for } E > E_M. \end{cases} \tag{1.45}$$

The multicanonical MC and MD runs are then performed respectively with the Metropolis criterion of Eq. (1.19) and with the modified Newton equation in Eq. (1.21), in which $\mathcal{E}^{\{0\}}_{\mathrm{MUCA}}(E)$ in Eq. (1.45) is substituted into $E_{\mathrm{MUCA}}(E;T_0)$. We expect to obtain a flat potential energy distribution in the range of Eq. (1.43). Finally, the results are analyzed by the single-histogram reweighting techniques as described in Eq. (1.27) (and Eq. (1.25)).

Some remarks are now in order. From Eqs. (1.18), (1.23), (1.24), and (1.44), Eq. (1.45) becomes

$$\mathcal{E}^{\{0\}}_{\mathrm{MUCA}}(E) \equiv \begin{cases} \frac{T_0}{T_1}\left(E-E_1\right) + T_0 S\left(E_1\right) = \frac{T_0}{T_1} E + \text{const}, & \text{for } E < E_1, \\ T_0 S(E), & \text{for } E_1 \le E \le E_M, \\ \frac{T_0}{T_M}\left(E-E_M\right) + T_0 S\left(E_M\right) = \frac{T_0}{T_M} E + \text{const}, & \text{for } E > E_M. \end{cases} \tag{1.46}$$

The Newton equation in Eq. (1.21) is then written as (see Eqs. (1.22), (1.23), and (1.24))

$$\dot{\boldsymbol{p}}_k = \begin{cases} \frac{T_0}{T_1} \boldsymbol{f}_k - \frac{\dot{s}}{s} \boldsymbol{p}_k, & \text{for } E < E_1, \\ \frac{T_0}{T(E)} \boldsymbol{f}_k - \frac{\dot{s}}{s} \boldsymbol{p}_k, & \text{for } E_1 \le E \le E_M, \\ \frac{T_0}{T_M} \boldsymbol{f}_k - \frac{\dot{s}}{s} \boldsymbol{p}_k, & \text{for } E > E_M. \end{cases} \tag{1.47}$$

Because only the product of inverse temperature β and potential energy E enters in the Boltzmann factor (see Eq. (1.5)), a rescaling of the potential energy (or force) by a constant, say α, can be considered as the rescaling of the temperature by $1/\alpha$ [21]. Hence, our choice of $\mathcal{E}^{\{0\}}_{\mathrm{MUCA}}(E)$ in Eq. (1.45) results in a canonical simulation at $T = T_1$ for $E < E_1$, a multicanonical simulation for $E_1 \le E \le E_M$, and a canonical simulation at $T = T_M$ for $E > E_M$. Note also that the above arguments are independent of the value of T_0, and we will get the same results, regardless of its value.

For Monte Carlo method, the above statement follows directly from the following equation. Namely, our choice of the multicanonical potential energy in Eq. (1.45) gives (by substituting Eq. (1.46) into Eq. (1.17))

$$W_{\text{MUCA}}(E) \equiv \exp\left[-\beta_0 \mathcal{E}_{\text{MUCA}}^{\{0\}}(E)\right] = \begin{cases} \exp\left(-\beta_1 E + \text{const}\right), & \text{for } E < E_1, \\ \frac{1}{n(E)}, & \text{for } E_1 \le E \le E_M, \\ \exp\left(-\beta_M E + \text{const}\right), & \text{for } E > E_M. \end{cases} \tag{1.48}$$

We now present the *multicanonical replica-exchange method* (MUCAREM) [13–15]. In MUCAREM the production run is a REM simulation with a few replicas not in the canonical ensemble but in the multicanonical ensemble, i.e., different replicas perform MUCA simulations with different energy ranges. While MUCA simulations are usually based on local updates, a replica-exchange process can be considered to be a global update, and global updates enhance the sampling further.

Let $\mathcal{M}$ be the number of replicas for a MUCAREM simulation. Here, each replica is in one-to-one correspondence not with temperature but with multicanonical weight factors of different energy range. Note that because multicanonical simulations cover much wider energy ranges than regular canonical simulations, the number of required replicas for the production run of MUCAREM is much less than that for the regular REM ($\mathcal{M} \ll M$). The weight factor for this generalized ensemble is now given by (see Eq. (1.32))

$$W_{\text{MUCAREM}}(X) = \prod_{i=1}^{\mathcal{M}} W_{\text{MUCA}}^{\{m(i)\}}\left(E\left(x_{m(i)}^{[i]}\right)\right) = \prod_{m=1}^{\mathcal{M}} W_{\text{MUCA}}^{\{m\}}\left(E\left(x_{m}^{[i(m)]}\right)\right), \tag{1.49}$$

where we prepare the multicanonical weight factor (and the density of states) separately for $\mathcal{M}$ regions (see Eq. (1.17)):

$$W_{\text{MUCA}}^{\{m\}}\left(E\left(x_m^{[i]}\right)\right) = \exp\left[-\beta_m \mathcal{E}_{\text{MUCA}}^{\{m\}}\left(E\left(x_m^{[i]}\right)\right)\right] \equiv \frac{1}{n^{\{m\}}\left(E\left(x_m^{[i]}\right)\right)}. \tag{1.50}$$

Here, we have introduced $\mathcal{M}$ arbitrary reference temperatures T_m $(= 1/k_{\text{B}}\beta_m)$ $(m = 1, \ldots, \mathcal{M})$, but the final results will be independent of the values of T_m, as one can see from the second equality in Eq. (1.50) (these arbitrary temperatures are necessary only for MD simulations).

Each multicanonical weight factor $W_{MUCA}^{\{m\}}(E)$, or the density of states $n^{\{m\}}(E)$, is defined as follows. For each m $(m = 1, \ldots, \mathcal{M})$, we assign a pair of temperatures $(T_L^{\{m\}}, T_H^{\{m\}})$. Here, we assume that $T_L^{\{m\}} < T_H^{\{m\}}$ and arrange the temperatures so that the neighboring regions covered by the pairs have sufficient overlaps. Without loss of generality we can assume $T_{\text{L}}^{\{1\}} < \cdots < T_{\text{L}}^{\{\mathcal{M}\}}$ and $T_{\text{H}}^{\{1\}} < \cdots < T_{\text{H}}^{\{\mathcal{M}\}}$. We define the following quantities:

$$\begin{cases} E_{\text{L}}^{\{m\}} = <E>_{T_{\text{L}}^{\{m\}}}, \\ E_{\text{H}}^{\{m\}} = <E>_{T_{\text{H}}^{\{m\}}}, \ (m = 1, \ldots, \mathcal{M}). \end{cases} \tag{1.51}$$

Suppose that the multicanonical weight factor $W_{\mathrm{MUCA}}(E)$ (or equivalently, the multicanonical potential energy $E_{\mathrm{MUCA}}(E;T_0)$ in Eq. (1.18)) has been obtained as in REMUCA or by any other methods in the entire energy range of interest ($E_{\mathrm{L}}^{\{1\}} < E < E_{\mathrm{H}}^{\{\mathcal{M}\}}$). We then have for each m ($m = 1, \ldots, \mathcal{M}$) the following multicanonical potential energies (see Eq. (1.45)) [13]:

$$\mathcal{E}_{\mathrm{MUCA}}^{\{m\}}(E) \equiv \begin{cases} \left.\frac{\partial E_{\mathrm{MUCA}}(E;T_m)}{\partial E}\right|_{E=E_{\mathrm{L}}^{\{m\}}} \left(E - E_{\mathrm{L}}^{\{m\}}\right) + E_{\mathrm{MUCA}}\left(E_{\mathrm{L}}^{\{m\}}; T_m\right), & \text{for } E < E_{\mathrm{L}}^{\{m\}}, \\ E_{\mathrm{MUCA}}\left(E; T_m\right) & \text{for } E_{\mathrm{L}}^{\{m\}} \le E \le E_{\mathrm{H}}^{\{m\}}, \\ \left.\frac{\partial E_{\mathrm{MUCA}}(E;T_m)}{\partial E}\right|_{E=E_{\mathrm{H}}^{\{m\}}} \left(E - E_{\mathrm{H}}^{\{m\}}\right) + E_{\mathrm{MUCA}}\left(E_{\mathrm{H}}^{\{m\}}; T_m\right), & \text{for } E > E_{\mathrm{H}}^{\{m\}} \end{cases} \tag{1.52}$$

Finally, a MUCAREM simulation is realized by alternately performing the following two steps.

1. Each replica of the fixed multicanonical ensemble is simulated *simultaneously* and *independently* for a certain MC or MD steps.
2. A pair of replicas, say i and j, which are in neighboring multicanonical ensembles, say m-th and $(m+1)$-th, respectively, are exchanged:

$$X = \left\{\ldots, x_m^{[i]}, \ldots, x_{m+1}^{[j]}, \ldots\right\} \to X' = \left\{\ldots, x_m^{[j]}, \ldots, x_{m+1}^{[i]}, \ldots\right\}. \tag{1.53}$$

The transition probability of this replica exchange is given by the Metropolis criterion:

$$w\left(X \to X'\right) = \min\left(1, \exp\left(-\Delta\right)\right), \tag{1.54}$$

where we now have (see Eq. (1.37)) [13]

$$\begin{aligned} \Delta = \beta_m &\left\{\mathcal{E}_{\mathrm{MUCA}}^{\{m\}}\left(E\left(q^{[j]}\right)\right) - \mathcal{E}_{\mathrm{MUCA}}^{\{m\}}\left(E\left(q^{[i]}\right)\right)\right\} \\ &- \beta_{m+1}\left\{\mathcal{E}_{\mathrm{MUCA}}^{\{m+1\}}\left(E\left(q^{[j]}\right)\right) - \mathcal{E}_{\mathrm{MUCA}}^{\{m+1\}}\left(E\left(q^{[i]}\right)\right)\right\}. \end{aligned} \tag{1.55}$$

Here, $E(q^{[i]})$ and $E(q^{[j]})$ are the potential energy of the i-th replica and the j-th replica, respectively.

Note that in Eq. (1.55) we need to newly evaluate the multicanonical potential energy, $\mathcal{E}_{\mathrm{MUCA}}^{\{m\}}\left(E\left(q^{[j]}\right)\right)$ and $\mathcal{E}_{\mathrm{MUCA}}^{\{m+1\}}\left(E\left(q^{[i]}\right)\right)$, because $\mathcal{E}_{\mathrm{MUCA}}^{\{m\}}(E)$ and $\mathcal{E}_{\mathrm{MUCA}}^{\{n\}}(E)$ are, in general, different functions for $m \neq n$.

In this algorithm, the m-th multicanonical ensemble actually results in a canonical simulation at $T = T_L^{\{m\}}$ for $E < E_L^{\{m\}}$, a multicanonical simulation for $E_L^{\{m\}} \le E \le E_H^{\{m\}}$, and a canonical simulation at $T = T_H^{\{m\}}$ for $E > E_H^{\{m\}}$, while the replica-exchange process samples states of the whole energy range ($E_{\mathrm{L}}^{\{1\}} \le E \le E_{\mathrm{H}}^{\{\mathcal{M}\}}$).

For obtaining the canonical distributions at any intermediate temperature T, the multiple-histogram reweighting techniques are again used. Let $N_m(E)$ and n_m be respectively the potential-energy histogram and the total number of samples obtained with the multicanonical weight factor $W_{MUCA}^{\{m\}}(E)$ $(m = 1, \ldots, \mathcal{M})$. The expectation value of a physical quantity A at any temperature T $(= 1/k_{\rm B}\beta)$ is then obtained from Eq. (1.25), where the best estimate of the density of states is obtained by solving the WHAM equations, which now read [13]

$$n(E) = \frac{\sum_{m=1}^{\mathcal{M}} N_m(E)}{\sum_{m=1}^{\mathcal{M}} n_m \exp\left(f_m\right) W_{\rm MUCA}^{\{m\}}(E)} = \frac{\sum_{m=1}^{\mathcal{M}} N_m(E)}{\sum_{m=1}^{\mathcal{M}} n_m \exp\left(f_m - \beta_m \mathcal{E}_{\rm MUCA}^{\{m\}}(E)\right)}, \tag{1.56}$$

where we have for each m $(= 1, \ldots, \mathcal{M})$

$$\exp\left(-f_m\right) = \sum_E n(E) W_{\rm MUCA}^{\{m\}}(E) = \sum_E n(E) \exp\left(-\beta_m \mathcal{E}_{\rm MUCA}^{\{m\}}(E)\right). \tag{1.57}$$

Note that $W_{MUCA}^{\{m\}}(E)$ is used instead of the Boltzmann factor $\exp(-\beta_m E)$ in Eqs. (1.40) and (1.41).

Moreover, ensemble averages of any physical quantity A (including those that cannot be expressed as functions of potential energy) at any temperature T $(= 1/k_{\rm B}\beta)$ can now be obtained from the "trajectory" of configurations of the production run. Namely, we first obtain f_m $(m = 1, \ldots, \mathcal{M})$ by solving Eqs. (1.56) and (1.57) self-consistently, and then we have [14]

$$<A>_T = \frac{\sum_{m=1}^{\mathcal{M}} \sum_{k=1}^{n_m} A(x_m(k)) \frac{1}{\sum_{l=1}^{\mathcal{M}} n_l \exp\left(f_l\right) W_{\rm MUCA}^{\{l\}}\left(E\left(x_m(k)\right)\right)} \exp[-\beta E(x_m(k))]}{\sum_{m=1}^{\mathcal{M}} \sum_{k=1}^{n_m} \frac{1}{\sum_{l=1}^{\mathcal{M}} n_l \exp\left(f_l\right) W_{\rm MUCA}^{\{l\}}\left(E\left(x_m(k)\right)\right)} \exp[-\beta E(x_m(k))]}, \tag{1.58}$$

where the trajectories $x_m(k)(k = 1, \ldots, n_m)$ are taken from each multicanonical simulation with the multicanonical weight factor $W_{MUCA}^{\{m\}}(E)$ $(m = 1, \ldots, \mathcal{M})$ separately.

As seen above, both REMUCA and MUCAREM can be used to obtain the multicanonical weight factor, or the density of states, for the entire potential energy range of interest. For complex systems, however, a single REMUCA or MUCAREM simulation is often insufficient. In such cases we can iterate MUCA (in REMUCA) and/or MUCAREM simulations in which the estimate of the multicanonical weight factor is updated by the single- and/or multiple-histogram reweighting techniques, respectively.

To be more specific, this iterative process can be summarized as follows. The REMUCA production run corresponds to a MUCA simulation with the weight factor $W_{\mathrm{MUCA}}(E)$. The new estimate of the density of states can be obtained by the single-histogram reweighting techniques of Eq. (1.27). On the other hand, from the MUCAREM production run, the improved density of states can be obtained by the multiple-histogram reweighting techniques of Eqs. (1.56) and (1.57).

The improved density of states thus obtained leads to a new multicanonical weight factor (see Eq. (1.17)). The next iteration can be either a MUCA production run (as in REMUCA) or MUCAREM production run. The results of this production run may yield an optimal multicanonical weight factor that yields a sufficiently flat energy distribution for the entire energy range of interest. If not, we can repeat the above process by obtaining the third estimate of the multicanonical weight factor either by a MUCA production run (as in REMUCA) or by a MUCAREM production run, and so on.

We remark that as the estimate of the multicanonical weight factor becomes more accurate, one is required to have a less number of replicas for a successful MUCAREM simulation, because each replica will have a flat energy distribution for a wider energy range. Hence, for a large, complex system, it is often more efficient to first try MUCAREM and iteratively reduce the number of replicas so that eventually one needs only one or a few replicas (instead of trying REMUCA directly from the beginning and iterating MUCA simulations).

1.3 Simulation Results

We now present some examples of the simulation results by the algorithms described in the previous section. The computer code developed in Refs. [12, 13, 29, 30], which is based on the version 2 of PRESTO [31], was used after modifications that were necessary for each calculation.

The first example is the C-peptide of ribonuclease A in explicit water [32]. The N-terminus and the C-terminus of the C-peptide analogue were blocked with the acetyl group and the N-methyl group, respectively. The number of amino acids is 13 and the amino-acid sequence is: Ace-Ala-Glu$^-$-Thr-Ala-Ala-Ala-Lys$^+$-Phe-Leu-Arg$^+$-Ala-His$^+$-Ala-Nme [33, 34]. It is known by experiments that this peptide forms α-helix structures [33, 34]. The initial configuration of our simulation was first generated by a high temperature molecular dynamics simulation (at $T = 1{,}000$ K) in gas phase, starting from a fully extended conformation. We randomly selected one of the structures that do not have any secondary structures such as α-helix and β-sheet. The peptide was then solvated in a sphere of radius 22 Å, in which 1,387 water

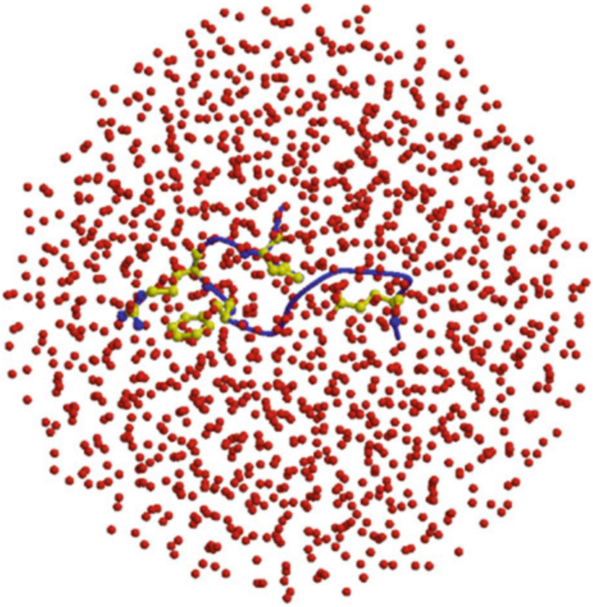

Fig. 1.1 The initial configuration of C-peptide in explicit water, which was used in all of the 32 replicas of the first REMD simulation (REMD1 in Table 1.1). The *red filled circles* stand for the oxygen atoms of water molecules. The number of water molecules is 1,387, and they are placed in a sphere of radius 22 Å. As for the peptide, besides the backbone structure (in *blue*), side chains of only Glu$^-$-2, Phe-8, Arg$^+$-10, and His$^+$-12 are shown (in *yellow*) (Reprinted from Ref. [32] with kind permission of Cell Press (2005))

Table 1.1 Summary of parameters in REMD, MUCAREM, and REMUCA simulations of C-peptide in explicit water[a]

Simulation	Number of replicas, M	Temperature, T_m (K) ($m = 1, \ldots, M$)	MD steps per replica
REMD1[b]	32	250, 258, 267, 276, 286, 295, 305, 315, 326, 337, 348, 360, 372, 385, 398, 411, 425, 440, 455, 470, 486, 502, 519, 537, 555, 574, 593, 613, 634, 655, 677, 700	2.0×10^5
MUCAREM1	4	360, 440, 555, 700	2.0×10^6
REMUCA1	1	700	3.0×10^7

[a] Reprinted from Ref. [32] with kind permission of Cell Press (2005)

[b] REMD1 stands for the replica-exchange molecular dynamics simulation, MUCAREM1 stands for the multicanonical replica-exchange molecular dynamics simulation, and REMUCA1 stands for the final multicanonical molecular dynamics simulation (the production run) of REMUCA. The results of REMD1 were used to determine the multicanonical weight factors for MUCAREM1, and those of MUCAREM1 were used to determine the multicanonical weight factor for REMUCA1

molecules were included (see Fig. 1.1). Harmonic restraint was applied to prevent the water molecules from going out of the sphere. The total number of atoms is 4,365. The dielectric constant was set equal to 1.0. The force-field parameters for protein were taken from the all-atom version of AMBER parm99 [35], which was found to be suitable for studying helical peptides [36, 37], and TIP3P model [38] was used for water molecules. The unit time step, Δt, was set to 0.5 fs. In Table 1.1 the parameter values in the simulations performed are summarized.

We first performed a REMD simulation with 32 replicas for 100 ps per replica (REMD1 in Table 1.1). During this REMD simulation, replica exchange was tried every 200 MD steps. Using the obtained potential-energy histogram of each replica as input data to the multiple-histogram analysis in Eqs. (1.40) and (1.41), we obtained the first estimate of the multicanonical weight factor, or the density of states. We divided this multicanonical weight factor into four multicanonical weight factors that cover different energy regions [13–15] and assigned these multicanonical weight factors into four replicas (the weight factors cover the potential energy ranges from −13791.5 to −11900.5 kcal/mol, from −12962.5 to −10796.5 kcal/mol, from −11900.5 to −9524.5 kcal/mol, and from −10796.5 to −8293.5 kcal/mol). We then carried out a MUCAREM simulation with four replicas for 1 ns per replica (MUCAREM1 in Table 1.1), in which replica exchange was tried every 1,000 MD steps. We again used the potential-energy histogram of each replica as the input data to the multiple-histogram analysis and finally obtained the multicanonical weight factor with high precision. As a production run, we carried out a 15-ns multicanonical MD simulation with one replica (REMUCA1 in Table 1.1) and the results of this production run were analyzed in detail.

In Fig. 1.2 we show the probability distributions of potential energy that were obtained from the above three generalized-ensemble simulations, namely, REMD1, MUCAREM1, and REMUCA1. We see in Fig. 1.2a that there are enough overlaps between all pairs of neighboring canonical distributions, suggesting that there were sufficient numbers of replica exchange in REMD1. We see in Fig. 1.2b that there are good overlaps between all pairs of neighboring multicanonical distributions, implying that MUCAREM1 also performed properly. Finally, the multicanonical distribution in Fig. 1.2c is completely flat between around −13,000 kcal/mol and around −8,000 kcal/mol. The results suggest that a free random walk was realized in this energy range.

In Fig. 1.3a we show the time series of potential energy from REMUCA1. We indeed observe a random walk covering as much as 5,000 kcal/mol of energy range. We show in Fig. 1.3b the average potential energy as a function of temperature, which was obtained from the trajectory of REMUCA1 by the reweighting techniques. The average potential energy monotonically increases as the temperature increases.

The accuracy of average quantities calculated depend on the "quality" of the random walk in the potential energy space, and the measure for this quality can be given by the number of tunneling events [7, 15]. One tunneling event is defined by a trajectory that goes from E_{H} to E_{L} and back, where E_{H} and E_{L} are the values near the highest energy and the lowest energy, respectively, which the random walk can reach. If E_{H} is sufficiently high, the trajectory gets completely uncorrelated when it reaches E_{H}. On the other hand, when the trajectory reaches near E_{L}, it tends to get trapped in local-minimum states. We thus consider that the more tunneling events we observe during a fixed number of MC/MD steps, the more efficient the method is as a generalized-ensemble algorithm (or, the average quantities obtained by the reweighting techniques are more reliable). Here, we took $E_{\mathrm{H}} = -8{,}250$ kcal/mol and $E_{\mathrm{L}} = -12{,}850$ kcal/mol for the measurement of the tunneling events. The

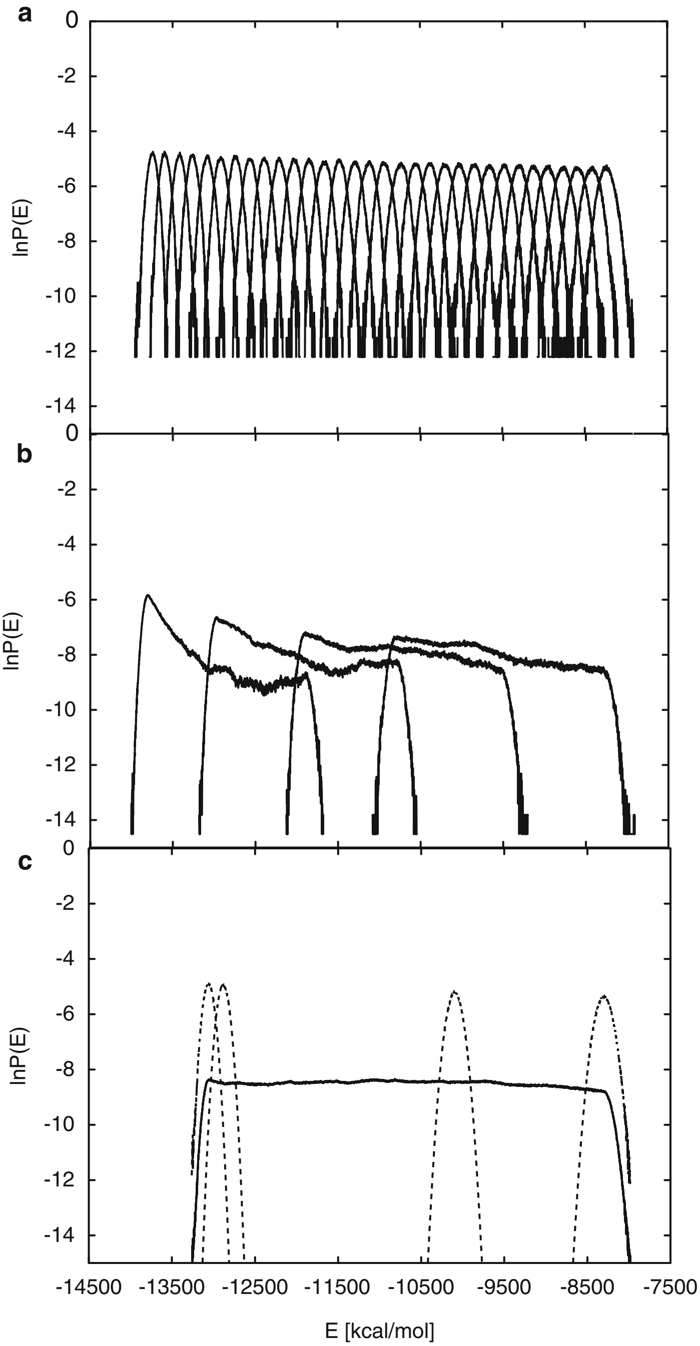

Fig. 1.2 Probability distributions of potential energy of the C-peptide system obtained from (**a**) REMD1, (**b**) MUCAREM1, and (**c**) REMUCA1. See Table 1.1 for the parameters of the simulations. *Dashed curves* in (**c**) are the reweighted canonical distributions at 290, 300, 500, and 700 K (from *left* to *right*) (Reprinted from Ref. [32] with kind permission of Cell Press (2005))

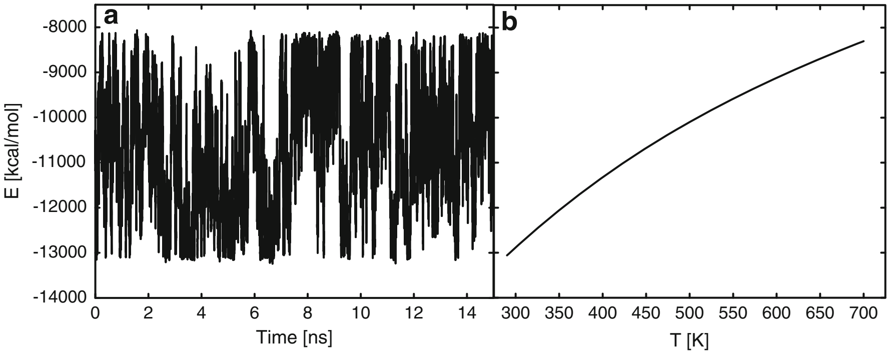

Fig. 1.3 Time series of potential energy of the C-peptide system from the REMUCA production run (REMUCA1 in Table 1.1) (**a**) and the average potential energy as a function of temperature (**b**). The latter was obtained from the trajectory of REMUCA1 by the single-histogram reweighting techniques (Reprinted from Ref. [32] with kind permission of Cell Press (2005))

random walk in REMUCA1 yielded as many as 55 tunneling events in 15 ns. The corresponding numbers of tunneling events for REMD1 and for MUCAREM1 were 0 in 3.2 ns and 5 in 4 ns, respectively. Hence, REMUCA is the most efficient and reliable among the three generalized-ensemble algorithms.

In Fig. 1.4 the potential of mean force along the first two principal component axes at 300 K is shown. There exist three distinct minima in the free-energy landscape, which correspond to three local-minimum-energy states. We show representative conformations at these minima in Fig. 1.5. The structure of the global-minimum free-energy state (GM) has a partially distorted α-helix with the salt bridge between Glu^{-}-2 and Arg^{+}-10. The structure is in good agreement with the experimental structure obtained by both NMR and X-ray experiments. In this structure there also exists a contact between Phe-8 and His^{+}-12. This contact is again observed in the corresponding residues of the X-ray structure. At LM1 the structure has a contact between Phe-8 and His^{+}-12, but the salt bridge between Glu^{-}-2 and Arg^{+}-10 is not formed. On the other hand, the structure at LM2 has this salt bridge, but it does not have a contact between Phe-8 and His^{+}-12. Thus, only the structures at GM satisfy all of the interactions that have been observed by the X-ray and other experimental studies.

The next example is the C-terminal β-hairpin of streptococcal protein G B1 domain [39]. This peptide is sometimes referred to as G-peptide [40] and is known by experiments to form β-hairpin structures in aqueous solution [41, 42]. The number of amino acids is 16 and the amino-acid sequence is: Gly-Glu^{-}-Trp-Thr-Tyr-Asp^{-}-Asp^{-}-Ala-Thr-Lys^{+}-Thr-Phe-Thr-Val-Thr-Glu^{-}. The N-terminus and C-terminus were set to be in the zwitter ionic form ($\text{NH}_3{}^{+}$ and COO^{-}), following the conditions in the experiments. GROMOS96 (43a1) force field [43] was used for the solute molecule. SPC model [44] was employed for solvent water molecules according to the GROMOS prescription. We first performed a REMD simulation of

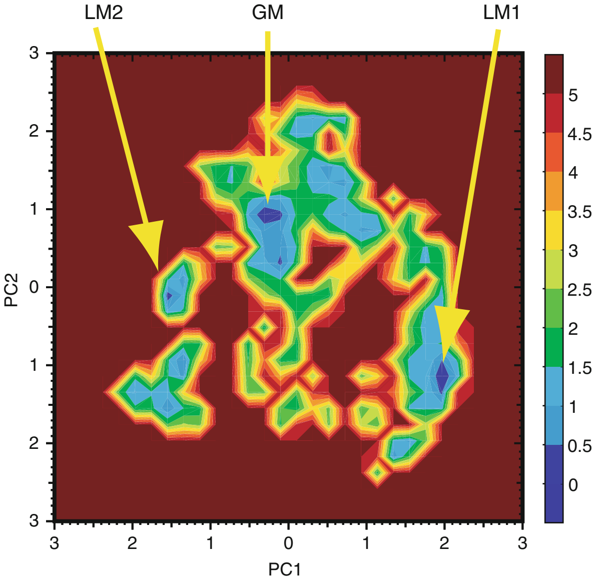

Fig. 1.4 Potential of mean force (kcal/mol) of the C-peptide system along the first two principal components at 300 K. The free energy was calculated from the results of REMUCA production run (REMUCA1 in Table 1.1) by the single-histogram reweighting techniques and normalized so that the global-minimum state (GM) has the value zero. GM, LM1, and LM2 represent three distinct minimum free-energy states (Reprinted from Ref. [32] with kind permission of Cell Press (2005))

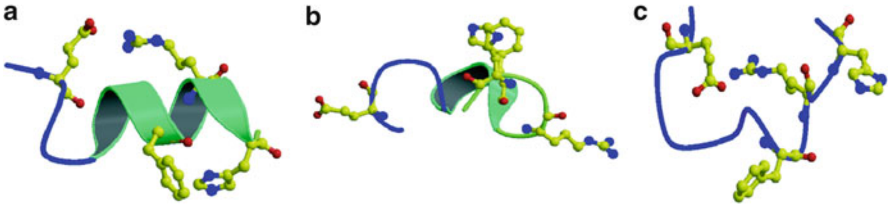

Fig. 1.5 The representative structures at the global-minimum free-energy state ((**a**) GM) and the two local-minimum states ((**b**) LM1 and (**c**) LM2). As for the peptide structures, besides the backbone structure, side chains of only Glu^{-}-2, Phe-8, Arg^{+}-10, and His^{+}-12 are shown in ball-and-stick model (Reprinted from Ref. [32] with kind permission of Cell Press (2005))

G-peptide without explicit solvents from a fully extended polypeptide conformation. In the simulation, we used the distance-dependent dielectric constant. We then selected the final conformation in the replica that was simulated at the highest temperature at the end of the simulation. This conformation was soaked in a water cap whose radius was 26 Å. Before starting the MUCAREM simulation, we performed a 100-ps REMD simulation with 64 replicas twice. (One of them was done for optimization of temperature table for the second REMD.) Using the results of the second REMD, we determined the initial multicanonical weight

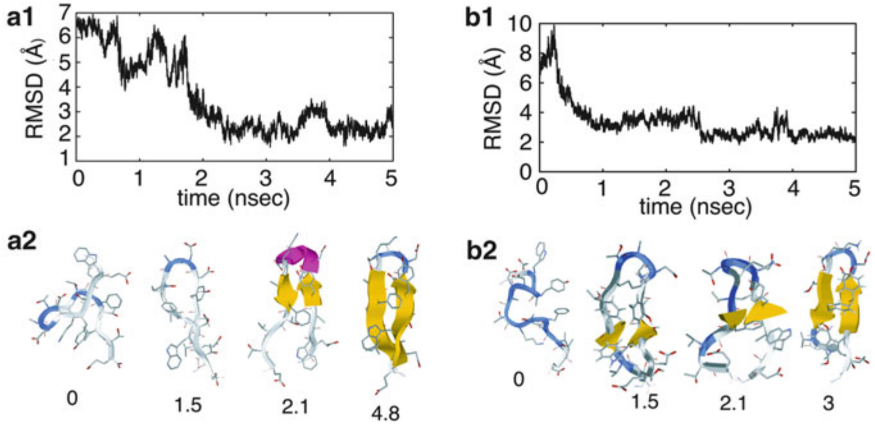

Fig. 1.6 Time series of structural properties of two folding events of G-peptide, resulting in the native-like β-hairpin structures. Those during 5-ns time windows are shown. (**a1**) and (**b1**) are the time series of heavy-atom RMSD values for Replica 4 and Replica 8, respectively. Likewise, (**a2**) and (**b2**) are representative snapshot structures observed in these 5-ns time windows. The numbers written under the snapshot structures represent the time when it was observed (Reprinted from Ref. [39] with kind permission of Wiley (2007))

factor. By iterating cycles of a short MUCAREM with 8 replicas and an update to a new weight factor [15], we refined the multicanonical weight factor. After that we performed a MUCAREM MD with 8 replicas for 34.75 ns (per replica) as a production run. Thus, the total production MD length was 278 ns. In total, three independent folding events were observed in three different replicas. Thus, the average simulation length per one observed folding event was 92.7 ns. This suggests that MUCAREM can accelerate G-peptide folding more than 60 times than the conventional MD simulations, because the experimental folding time of G-peptide is 6 μs [45].

Figure 1.6 shows the time series of the heavy-atom Root Mean Square Deviation (RMSD) from the native configuration (coordinates in the PDB entry 2GB1) and representative snapshot structures observed in the folding events are shown for two replicas. They indeed folded into native-like conformations.

We also evaluated the canonical expectation values of secondary-structure contents (β-bridge contents) of each residue at 320 K using the multiple-histogram reweighting techniques in Eqs. (1.56), (1.57), and (1.58). The results are shown in Fig. 1.7. These results are qualitatively similar to the previous ones that were derived from shorter MUCAREM simulations [36, 37]. They clearly imply that the β-hairpin structures are formed at this temperature.

The third example is the chicken villin headpiece subdomain in explicit water [46]. The number of amino acids is 36. The force field CHARMM22 [47] with CMAP [48, 49] and TIP3P water model [38, 47] were used. The number of water molecules was 3,513. The MD time step was 1.0 fs. We made two production runs of about 1 μs, each of which was a MUCAREM simulation with eight replicas.

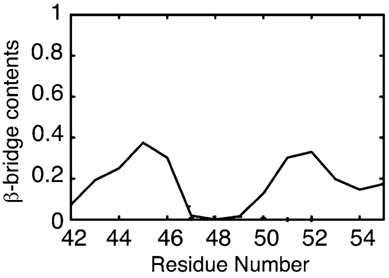

Fig. 1.7 Canonical expectation values of the β-bridge contents of G-peptide at 320 K as a function of the residue number. Values are evaluated by the multiple-histogram reweighting techniques (Reprinted from Ref. [39] with kind permission of Wiley (2007))

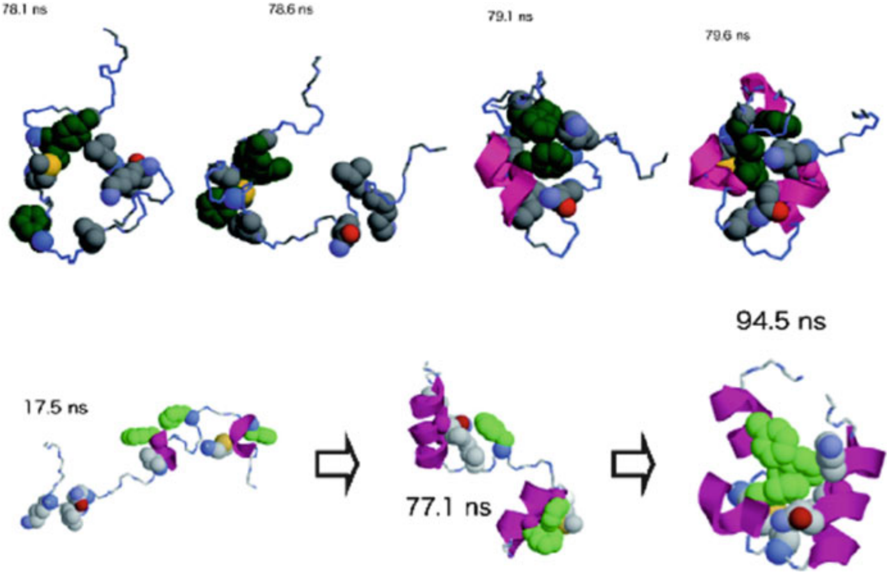

Fig. 1.8 Snapshots of villin headpiece during the MUCAREM production runs that folded into native-like conformations: MUCAREM1 (*above*) and MUCAREM2 (*below*)

They are referred to as MUCAREM1 and MUCAREM2. The former consisted of 1.127 μs covering the temperature range between 269 and 699 K, and the latter 1.157 μs covering the temperature range between 289 and 699 K.

We consider that the backbone folded into the native structure from unfolded ones if the mainchain RMSD becomes less than or equal to 3.0 Å. The folding event is counted separately if it goes through an unfolded structure (with the backbone RMSD greater than or equal to 6.5 Å). With this criterion, we observed 11 folding events in seven different replicas (namely, Replicas 5, 7, and 8 in MUCAREM1 and Replicas 1, 2, 4, and 5 in MUCAREM2). In Fig. 1.8 we show the snapshots of the

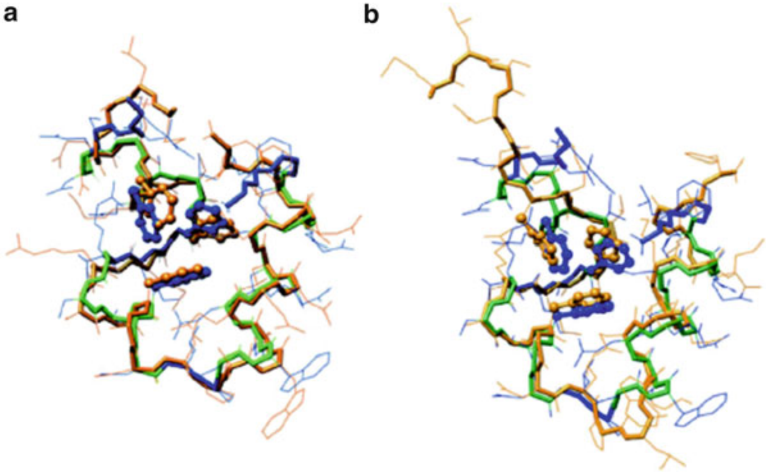

Fig. 1.9 Low-RMSD conformations of villin headpiece subdomain HP36 obtained in MUCAREM1 and MUCAREM2 (colored in *orange*). The X-ray structure (PDB ID: 1YRF) is also superimposed (colored in *blue* and *green*). Here, the α-helices in the X-ray structure are colored in *green* and the rest in *blue*. Three phenylalanine side chains (Phe7, Phe11, and Phe18), which form a hydrophobic core, are shown in *ball*-and-*stick* representation. (**a**) The lowest-bakcbone-RMSD conformation observed in the two MUCAREM production runs (Replica 5 of MUCAREM2). The backbone RMSD value is 1.1 Å (for non-terminal 34 residues). (**b**) A low-RMSD conformation observed in MUCAREM1 (Replica 8). The RMSD value is 1.0 Å for residues 9–32 and 3.3 Å for non-terminal 34 residues (Reprinted from Ref. [46] with kind permission of Cell Press (2010))

replicas folding into native-like conformations for the two MUCAREM production runs. In Fig. 1.9 we compare the obtained low-RMSD conformations and the native structure. They are indeed very close to the native structure.

1.4 Conclusions

In this article we introduced four powerful generalized-ensemble algorithms, namely, multicanonical algorithm (MUCA), replica-exchange method (REM), replica-exchange multicanonical algorithm (REMUCA), and multicanonical replica-exchange method (MUCAREM), which can greatly enhance conformational sampling of biomolecular systems. The results of protein folding simulations by these methods were presented. Because it is very difficult to determine the multicanonical weight factors for very large systems, MUCAREM is the most promising method among the four methods for large biomolecular systems.

Acknowledgments This work was supported, in part, by Grants-in-Aid for Scientific Research on Innovative Areas, "Transient Macromolecular Complex" (Y.S.) and "Fluctuations and Biological Functions" (Y.O.), for Computational Materials Science Initiative (Y.O.), and for High Performance Computing Infrastructure (HPCI) (Y.S. and Y.O.) from MEXT, Japan

References

1. Hansmann UHE, Okamoto Y (1999) New Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 9:177–183
2. Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 60:96–123
3. Sugita Y, Okamoto Y (2002) Free-energy calculations in protein folding by generalized-ensemble algorithms. In: Schlick T, Gan HH (eds) Lecture notes in computational science and engineering. Springer, Berlin, pp 304–332; e-print: cond-mat/0102296
4. Okumura H, Itoh SG, Okamoto Y (2012) Generalized-ensemble algorithms for simulations of complex molecular systems. In: Leszczynski J, Shukla MK (eds) Practical aspects of computational chemistry II: an overview of the last two decades and current trends. Springer, Dordrecht, pp 69–101
5. Sugita Y, Miyashita N, Li P-C, Yoda T, Okamoto Y (2012) Recent applications of replica-exchange molecular dynamics simulations of biomolecules. Curr Phys Chem 2:401–412
6. Berg BA, Neuhaus T (1991) Multicanonical algorithms for 1st order phase transitions. Phys Lett B267:249–253
7. Berg BA, Neuhaus T (1992) Multicanonical ensemble: a new approach to simulate first-order phase transitions. Phys Rev Lett 68:9–12
8. Hansmann UHE, Okamoto Y (1993) Prediction of peptide conformation by multicanonical algorithm – new approach to the multiple-minima problem. J Comput Chem 14:1333–1338
9. Hukushima K, Nemoto K (1996) Exchange Monte Carlo method and application to spin glass simulations. J Phys Soc Jpn 65:1604–1608
10. Marinari E, Parisi G, Ruiz-Lorenzo JJ (1997) Numerical simulations of spin glass systems. In: Young AP (ed) Spin glasses and random fields. World Scientific, Singapore, pp 59–98
11. Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281:140–150
12. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
13. Sugita Y, Okamoto Y (2000) Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. Chem Phys Lett 329:261–270
14. Mitsutake A, Sugita Y, Okamoto Y (2003) Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test. J Chem Phys 118:6664–6675
15. Mitsutake A, Sugita Y, Okamoto Y (2003) Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. II. Application to a more complex system. J Chem Phys 118:6676–6688
16. Sugita Y, Kitao A, Okamoto Y (2000) Multidimensional replica-exchange method for free-energy calculations. J Chem Phys 113:6042–6051
17. Fukunishi F, Watanabe O, Takada S (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. J Chem Phys 116:9058–9067
18. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
19. Nosé S (1984) A molecular dynamics method for simulations in the canonical ensemble. Mol Phys 52:255–268
20. Nosé S (1984) A unified formulation of the constant temperature molecular dynamics methods. J Chem Phys 81:511–519
21. Hansmann UHE, Okamoto Y, Eisenmenger F (1996) Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble. Chem Phys Lett 259:321–330
22. Nakajima N, Nakamura H, Kidera A (1997) Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. J Phys Chem B 101:817–824

23. Ferrenberg AM, Swendsen RH (1988) New Monte Carlo technique for studying phase transitions. Phys Rev Lett 61:2635–2638; (1989). *ibid., 63*, 1658
24. Berg BA (2004) Markov chain Monte Carlo simulations and their statistical analysis. World Scientific, Singapore, p 253
25. Berg BA (2003) Multicanonical simulations step by step. Comput Phys Commun 153:397–406
26. Mori Y, Okamoto Y (2010) Replica-exchange molecular dynamics simulations for various constant temperature algorithms. J Phys Soc Jpn 79:074001
27. Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. Phys Rev Lett 63:1195–1198
28. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. J Comput Chem 13:1011–1021
29. Sugita Y, Kitao A (1998) Improved protein free energy calculation by more accurate treatment of nonbonded energy: application to chymotrypsin inhibitor 2, V57A. Proteins 30:388–400
30. Kitao A, Hayward S, Go N (1998) Energy landscape of a native protein: jumping-among-minima model. Proteins 33:496–517
31. Morikami K, Nakai T, Kidera A, Saito M, Nakamura H (1992) PRESTO (protein engineering simulator): a vectorized molecular dynamics program for biopolymers. Comput Chem 16:243–248
32. Sugita Y, Okamoto Y (2005) Molecular mechanism for stabilizing a short helical peptide studied by generalized-ensemble simulations with explicit solvent. Biophys J 88:3180–3190
33. Shoemaker KR, Kim PS, York EJ, Stewart JM, Baldwin RL (1987) Tests of the helix dipole model for stabilization of alpha-helices. Nature 326:563–567
34. Shoemaker KR, Faiman R, Schultz DA, Robertson AD, York EJ, Stewart JM, Baldwin RL (1990) Side-chain interactions in the C-peptide helix: Phe 8 ... His 12^+. Biopolymers 29:1–11
35. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. J Comput Chem 21:1049–1074
36. Yoda T, Sugita Y, Okamoto Y (2004) Comparisons of force fields for proteins by generalized-ensemble simulations. Chem Phys Lett 386:460–467
37. Yoda T, Sugita Y, Okamoto Y (2004) Secondary-structure preferences of force fields for proteins evaluated by generalized-ensemble simulations. Chem Phys 307:269–283
38. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
39. Yoda T, Sugita Y, Okamoto Y (2007) Cooperative folding mechanism of a β-hairpin peptide studied by a multicanonical replica-exchange molecular dynamics simulation. Proteins 66:846–859
40. Honda S, Kobayashi N, Munekata E (2000) Thermodynamics of a β-hairpin structure: evidence for cooperative formation of folding nucleus. J Mol Biol 295:846–859
41. Blanco FJ, Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable β-hairpin in aqueous solution. Nat Struct Biol 1:584–589
42. Kobayashi N, Honda S, Yoshii H, Uedaira H, Munekata E (1995) Complement assembly of two fragments of the streptococcal protein G B1 domain in aqueous solution. FEBS Lett 366:99–103
43. van Gunsteren WF, Billeter SR, Eising AA, Hunenberger PH, Kruger P, Mark AE, Scott WRP, Tironi IG (1996) Biomolecular simulation: the GROMOS96 manual and user guide. Vdf Hochschulverlag AG an der ETH, Zurich
44. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Reidel, Dordrecht, pp 331–342
45. Munoz V, Thompson PA, Hofrichter J, Eaton WA (1997) Folding dynamics and mechanism of β-hairpin formation. Nature 390:196–199

46. Yoda T, Sugita Y, Okamoto Y (2010) Hydrophobic core formation and dehydration in protein folding studied by generalized-ensemble simulations. Biophys J 99:1637–1644
47. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WEIII, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
48. MacKerell AD Jr, Feig M, Brooks CL III (2004) Improved treatment of the protein backbone in empirical force fields. J Am Chem Soc 126:698–699
49. Mackerell AD Jr, Feig M, Brooks CL III (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 25:1400–1415