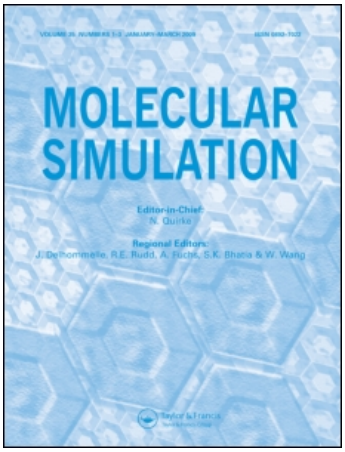

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Replica-exchange methods and predictions of helix configurations of membrane proteins

Hironori Kokubo[a]; Yuko Okamoto[b]
[a] Department of Chemistry, University of Houston, Houston, TX, USA [b] Department of Physics, Nagoya University, Nagoya, Aichi, Japan

To cite this Article Kokubo, Hironori and Okamoto, Yuko(2006) 'Replica-exchange methods and predictions of helix configurations of membrane proteins', Molecular Simulation, 32: 10, 791 — 801
To link to this Article: DOI: 10.1080/08927020601009591
URL: http://dx.doi.org/10.1080/08927020601009591

Molecular Simulation, Vol. 32, Nos. 10–11, 30 August–15 September 2006, 791–801

Taylor & Francis
Taylor & Francis Group

Replica-exchange methods and predictions of helix configurations of membrane proteins

HIRONORI KOKUBO†¶ and YUKO OKAMOTO‡*

†Department of Chemistry, University of Houston, Houston, TX 77204, USA
‡Department of Physics, Nagoya University, Nagoya, Aichi 464-8602, Japan

(Received June 2006; in final form September 2006)

A simulation in generalized ensemble is based on a non-Boltzmann weight factor and performs a random walk in potential energy space, which allows the simulation to avoid getting trapped in states of local-minimum energy states. In this article, we review uses of the generalized-ensemble algorithms. Three well-known methods, namely, multicanonical algorithm (MUCA), simulated tempering (ST) and replica-exchange method (REM), are described first. Both Monte Carlo (MC) and molecular dynamics (MD) versions of the algorithms are given. We then present the results of the application of replica-exchange MC method to the predictions of membrane protein structures.

Keywords: Replica-exchange method; Monte Carlo; Generalized-ensemble algorithm; Membrane proteins

1. Introduction

It is very difficult to obtain accurate canonical distributions at low temperatures for complex systems such as biomolecular systems by the conventional Monte Carlo (MC) and molecular dynamics (MD) simulations. This is because the simulations tend to get trapped in one of a huge number of local-minimum-energy states, which are separated by high-energy barriers. One way to overcome this multiple-minima problem is to perform a simulation in a *generalized ensemble* where each state is weighted by a non-Boltzmann probability weight factor so that a random walk in the potential energy space may be realized (for recent reviews, see, for instance, Refs. [1,2]). The random walk allows the simulation to escape from any energy barrier and to sample a much wider phase space than by conventional methods.

One of the most well-known generalized-ensemble algorithms is perhaps *multicanonical algorithm* (MUCA) [3,4]. Another well known such algorithm is *simulated tempering* (ST) [5,6]. A third popular generalized-ensemble algorithm is the *replica-exchange method* (REM) [7–9]. (REM is also referred to as *parallel tempering* [10].)

In this article, we give details of these three generalized-ensemble algorithms. As examples, we present the results of the applications of replica-exchange MC method to the predictions of membrane protein structures.

In Section 2, the details of the three generalized-ensemble algorithms are described. In Section 3, the results of the predictions of membrane protein structures are presented. Section 4 is devoted to conclusions.

2. Generalized-ensemble algorithms

2.1 Multicanonical algorithm and simulated tempering

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 $q \equiv \{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$ and $p \equiv \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, respectively. The Hamiltonian $H(q, p)$ of the system is the sum of the kinetic energy $K(p)$ and the

*Corresponding author. Email: okamoto@phys.nagoya-u.ac.jp
¶Email: kokubo@kitten.chem.uh.edu

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020601009591

potential energy $E(q)$:

$$H(q,p) = K(p) + E(q), \tag{1}$$

where

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

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

$$W_{\mathrm{B}}(x;T) = \mathrm{e}^{-\beta H(q,p)}, \tag{3}$$

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

$$\langle K(p)\rangle_T = \left\langle \sum_{k=1}^{N} \frac{\mathbf{p}_k^2}{2m_k} \right\rangle_T = \frac{3}{2} N k_{\mathrm{B}} T. \tag{4}$$

Because the coordinates q and momenta p are decoupled in equation (1), we can suppress the kinetic energy part and can write the Boltzmann factor as

$$W_{\mathrm{B}}(x;T) = W_{\mathrm{B}}(E;T) = \mathrm{e}^{-\beta E}. \tag{5}$$

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

$$P_{\mathrm{B}}(E;T) \propto n(E) W_{\mathrm{B}}(E;T). \tag{6}$$

In the MUCA [3,4], on the other hand, each state is weighted by a non-Boltzmann weight factor $W_{\mathrm{mu}}(E)$ (which we refer to as the *multicanonical weight factor*) so that a uniform energy distribution $P_{\mathrm{mu}}(E)$ is obtained:

$$P_{\mathrm{mu}}(E) \propto n(E) W_{\mathrm{mu}}(E) \equiv \text{constant}. \tag{7}$$

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.

From the definition in equation (7), the multicanonical weight factor is inversely proportional to the density of states and we can write it as follows:

$$W_{\mathrm{mu}}(E) \equiv \mathrm{e}^{-\beta_0 E_{\mathrm{mu}}(E;T_0)} = \frac{1}{n(E)}, \tag{8}$$

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

$$E_{\mathrm{mu}}(E;T_0) = k_{\mathrm{B}} T_0 \ln n(E) = T_0 S(E). \tag{9}$$

Here, $S(E)$ is the entropy in the microcanonical ensemble. Since 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 [3,4].

A multicanonical MC simulation is performed, for instance, with the usual Metropolis criterion [11]: The transition probability of state x with potential energy E to state x' with potential energy E' is given by

$$w(x \to x') = \begin{cases} 1, & \text{for } \Delta E_{\mathrm{mu}} \le 0, \\ \exp(-\beta_0 \Delta E_{\mathrm{mu}}), & \text{for } \Delta E_{\mathrm{mu}} > 0, \end{cases} \tag{10}$$

where

$$\Delta E_{\mathrm{mu}} \equiv E_{\mathrm{mu}}(E';T_0) - E_{\mathrm{mu}}(E;T_0). \tag{11}$$

The MD algorithm in multicanonical ensemble also naturally follows from equation (8), in which the regular constant temperature MD simulation (with $T = T_0$) is performed by solving the following modified Newton equation: [12,13]

$$\dot{\mathbf{p}}_k = -\frac{\partial E_{\mathrm{mu}}(E;T_0)}{\partial \mathbf{q}_k} = \frac{\partial E_{\mathrm{mu}}(E;T_0)}{\partial E} \mathbf{f}_k, \tag{12}$$

where $\mathbf{f}_k$ is the usual force acting on the k-th atom ($k = 1, \ldots, N$).

After the optimal multicanonical weight factor is determined, one performs a long multicanonical simulation once. By monitoring the potential energy throughout the simulation, one can find the global-minimum-energy state. Moreover, by using the obtained histogram $N_{\mathrm{mu}}(E)$ of the potential energy distribution $P_{\mathrm{mu}}(E)$, the expectation value of a physical quantity A at any temperature $T = 1/k_{\mathrm{B}}\beta$ is calculated from

$$\langle A\rangle_T = \frac{\sum_E A(E) n(E)\, \mathrm{e}^{-\beta E}}{\sum_E n(E)\, \mathrm{e}^{-\beta E}}, \tag{13}$$

where the best estimate of the density of states is given by the single-histogram reweighting techniques (see equation (7)) [14]:

$$n(E) = \frac{N_{\mathrm{mu}}(E)}{W_{\mathrm{mu}}(E)}. \tag{14}$$

In the numerical work, we want to avoid round-off errors (and overflows and underflows) as much as possible. It is usually better to combine exponentials as follows (see equation (8)):

$$\langle A\rangle_T = \frac{\sum_E A(E) N_{\mathrm{mu}}(E)\, \mathrm{e}^{\beta_0 E_{\mathrm{mu}}(E;T_0) - \beta E}}{\sum_E N_{\mathrm{mu}}(E)\, \mathrm{e}^{\beta_0 E_{\mathrm{mu}}(E;T_0) - \beta E}}. \tag{15}$$

We now briefly review the original ST method [5,6]. In this method, the temperature itself becomes a dynamical variable and both the configuration and the temperature are updated during the simulation with a weight:

$$W_{\mathrm{ST}}(E;T) = \mathrm{e}^{-\beta E + a(T)}, \tag{16}$$

where the function $a(T)$ is chosen so that the probability distribution of temperature is flat:

$$P_{\mathrm{ST}}(T) = \int \mathrm{d}E\, n(E) W_{ST}(E;T) = \int \mathrm{d}E\, n(E)\, \mathrm{e}^{-\beta E + a(T)} = \text{constant}. \quad (17)$$

Hence, in ST the *temperature* is sampled uniformly. A free random walk in the temperature space is realized, which in turn induces a random walk in the potential energy space and allows the simulation to escape from states of energy local minima.

In the numerical work, we discretize the temperature in M different values, T_m $(m = 1, \ldots, M)$. Without loss of generality, we can order the temperature so that $T_1 < T_2 < \cdots < T_M$. The lowest temperature T_1 should be sufficiently low so that the simulation can explore the global-minimum-energy region and the highest temperature T_M should be sufficiently high so that no trapping in an energy-local-minimum state occurs. The probability weight factor in equation (16) is now written as

$$W_{\mathrm{ST}}(E;T_m) = \mathrm{e}^{-\beta_m E + a_m}, \quad (18)$$

where $a_m = a(T_m)$ $(m = 1, \ldots, M)$. The parameters a_m are not known *a priori* and have to be determined by iterations of short simulations. This process can be nontrivial and very difficult for complex systems. Note that from equations (17) and (18), we have

$$\mathrm{e}^{-a_m} \propto \int \mathrm{d}E\, n(E)\, \mathrm{e}^{-\beta_m E}. \quad (19)$$

The parameters a_m are therefore the "dimensionless" Helmholtz free energy at temperature T_m (i.e. the inverse temperature β_m multiplied by the Helmholtz free energy).

Once the parameters a_m are determined and the initial configuration and the initial temperature T_m are chosen, a ST simulation is then realized by alternately performing the following two steps [5,6]:

1. A canonical MC or MD simulation at the fixed temperature T_m is carried out for a certain MC or MD steps.
2. The temperature T_m is updated to the neighboring values $T_{m\pm1}$ with the configuration fixed. The transition probability of this temperature-updating process is given by the Metropolis criterion (equation (18)):

$$w(T_m \to T_{m\pm1}) = \begin{cases} 1, & \text{for } \Delta \le 0, \\ \exp(-\Delta), & \text{for } \Delta > 0, \end{cases} \quad (20)$$

where

$$\Delta = (\beta_{m\pm1} - \beta_m)E - (a_{m\pm1} - a_m). \quad (21)$$

Note that in Step 2, we exchange only pairs of neighboring temperatures in order to secure a sufficiently large acceptance ratio of temperature updates.

After the optimal ST weight factor is determined, one performs a long ST run once. From the results of this production run, one can obtain the canonical-ensemble average of a physical quantity A as a function of temperature from equation (13), where the density of states is given by the multiple-histogram reweighting techniques (also referred to as the weighted histogram analysis method) [15,16] as follows. Let $N_m(E)$ and n_m be, respectively the potential-energy histogram and the total number of samples obtained at temperature $T_m = 1/k_{\mathrm{B}}\beta_m$ $(m = 1, \ldots, M)$. The best estimate of the density of states is then given by [15,16]

$$n(E) = \frac{\sum_{m=1}^{M} g_m^{-1} N_m(E)}{\sum_{m=1}^{M} g_m^{-1} n_m\, \mathrm{e}^{f_m - \beta_m E}}, \quad (22)$$

where

$$\mathrm{e}^{-f_m} = \sum_E n(E)\, \mathrm{e}^{-\beta_m E}. \quad (23)$$

Here, $g_m = 1 + 2\tau_m$ and τ_m is the integrated autocorrelation time at temperature T_m. For many systems, the quantity g_m can safely be set to be a constant in the reweighting formulae [16] and we usually set $g_m = 1$. Note that equations (22) and (23) are solved self-consistently by iteration [15,16] to obtain the dimensionless Helmholtz free energy f_m (and the density of states $n(E)$). We remark that in the numerical work, it is often more stable to use the following equations instead of equations (22) and (23):

$$P_{\mathrm{B}}(E;T) = n(E)\, \mathrm{e}^{-\beta E} = \frac{\sum_{m=1}^{M} g_m^{-1} N_m(E)}{\sum_{m=1}^{M} g_m^{-1} n_m \mathrm{e}^{f_m - (\beta_m - \beta)E}}, \quad (24)$$

where

$$\mathrm{e}^{-f_m} = \sum_E P_{\mathrm{B}}(E;T_m). \quad (25)$$

The equations are solved iteratively as follows. We can set all the f_m $(m = 1, \ldots, M)$ to, e.g. zero initially. We then use equation (24) to obtain $P_{\mathrm{B}}(E;T_m)$ $(m = 1, \ldots, M)$, which are substituted into equation (25) to obtain next values of f_m and so on.

2.2 Replica-exchange method

The REM [7–9] was developed as an extension of ST. 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 is a one-to-one correspondence between replicas and temperatures; the label i $(i = 1, \ldots, M)$ for replicas is a permutation of the label m $(m = 1, \ldots, M)$ for temperatures and vice versa:

$$\begin{cases} i = i(m) \equiv f(m), \\ m = m(i) \equiv f^{-1}(i), \end{cases} \quad (26)$$

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. The state X is specified by the M sets of coordinates $q^{[i]}$ and momenta $p^{[i]}$ of N atoms in replica i at temperature T_m:

$$x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m. \tag{27}$$

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):

$$\begin{aligned} W_{\mathrm{REM}}(X) &= \exp\left\{-\sum_{i=1}^{M} \beta_{m(i)} H\left(q^{[i]}, p^{[i]}\right)\right\} \\ &= \exp\left\{-\sum_{m=1}^{M} \beta_m H\left(q^{[i(m)]}, p^{[i(m)]}\right)\right\}, \end{aligned} \tag{28}$$

where $i(m)$ and $m(i)$ are the permutation functions in equation (26).

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

$$\begin{aligned} X &= \left\{\ldots, x_m^{[i]}, \ldots, x_n^{[j]}, \ldots\right\} \to X' \\ &= \left\{\ldots, x_m^{[j]'}, \ldots, x_n^{[i]'}, \ldots\right\}. \end{aligned} \tag{29}$$

Here, i, j, m and n are related by the permutation functions in equation (26) and the exchange of replicas introduces a new permutation function f':

$$\begin{cases} i = f(m) \to j = f'(m), \\ j = f(n) \to i = f'(n). \end{cases} \tag{30}$$

The exchange of replicas can be written in more detail as

$$\begin{cases} x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m \to x_m^{[j]'} \equiv \left(q^{[j]}, p^{[j]'}\right)_m, \\ x_n^{[j]} \equiv \left(q^{[j]}, p^{[j]}\right)_n \to x_n^{[i]'} \equiv \left(q^{[i]}, p^{[i]'}\right)_n, \end{cases} \tag{31}$$

where the definitions for $p^{[i]'}$ and $p^{[j]'}$ will be given below. We remark that this process is equivalent to exchanging a pair of temperatures T_m and T_n for the corresponding replicas i and j as follows:

$$\begin{cases} x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m \to x_n^{[i]'} \equiv \left(q^{[i]}, p^{[i]'}\right)_n, \\ x_n^{[j]} \equiv \left(q^{[j]}, p^{[j]}\right)_n \to x_m^{[j]'} \equiv \left(q^{[j]}, p^{[j]'}\right)_m. \end{cases} \tag{32}$$

In the original implementation of the REM [7–9], MC algorithm was used and only the coordinates q (and the potential energy function $E(q)$) had to be taken into account. In the MD algorithm, on the other hand, we also have to deal with the momenta p. We proposed the following momentum assignment in equation (31) (and in equation (32)) [17]:

$$\begin{cases} p^{[i]'} \equiv \sqrt{\frac{T_n}{T_m}}\, p^{[i]}, \\ p^{[j]'} \equiv \sqrt{\frac{T_m}{T_n}}\, p^{[j]}, \end{cases} \tag{33}$$

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 equation (4) may be satisfied.

In order for this exchange process to converge towards an equilibrium distribution, it is sufficient to impose the detailed balance condition on the transition probability $w(X \to X')$:

$$W_{\mathrm{REM}}(X)\, w(X \to X') = W_{\mathrm{REM}}(X') w(X' \to X). \tag{34}$$

From equations (1), (2), (28), (33) and (34), we have

$$\begin{aligned} \frac{w(X \to X')}{w(X' \to X)} &= \exp\Big\{-\beta_m\Big[K\left(p^{[j]'}\right) + E\left(q^{[j]}\right)\Big] \\ &\quad -\beta_n\Big[K\left(p^{[i]'}\right) + E\left(q^{[i]}\right)\Big] \\ &\quad +\beta_m\big[K\left(p^{[i]}\right) + E\left(q^{[i]}\right)\big] \\ &\quad +\beta_n\big[K\left(p^{[j]}\right) + E\left(q^{[j]}\right)\big]\Big\}, \\ &= \exp\Big\{-\beta_m \frac{T_m}{T_n} K\left(p^{[j]}\right) - \beta_n \frac{T_n}{T_m} K\left(p^{[i]}\right) \\ &\quad +\beta_m K\left(p^{[i]}\right) + \beta_n K\left(p^{[j]}\right) \\ &\quad -\beta_m\big[E\left(q^{[j]}\right) - E\left(q^{[i]}\right)\big] \\ &\quad -\beta_n\big[E\left(q^{[i]}\right) - E\left(q^{[j]}\right)\big]\Big\}, = \exp(-\Delta), \end{aligned} \tag{35}$$

where

$$\Delta \equiv \left(\beta_n - \beta_m\right)\left(E\left(q^{[i]}\right) - E\left(q^{[j]}\right)\right), \tag{36}$$

and i, j, m and n are related by the permutation functions (in equation (26)) before the exchange:

$$\begin{cases} i = f(m), \\ j = f(n). \end{cases} \tag{37}$$

This can be satisfied, for instance, by the usual Metropolis criterion [11]:

$$\begin{aligned} w(X \to X') &\equiv w\left(x_m^{[i]} \middle| x_n^{[j]}\right) \\ &= \begin{cases} 1, & \text{for } \Delta \le 0, \\ \exp(-\Delta), & \text{for } \Delta > 0, \end{cases} \end{aligned} \tag{38}$$

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. Note that this is exactly the same criterion that was originally derived for MC algorithm [7–9].

Without loss of generality, we can again assume $T_1 < T_2 < \cdots < T_M$. A simulation of the REM [7–9] is then realized by alternately performing the following two steps:

1. Each replica in the 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 equation (38).

Note that in Step 2, we exchange only pairs of replicas corresponding to neighboring temperatures, because the acceptance ratio of the exchange process decreases exponentially with the difference of the two β's (equations (36) and (38)). Note also that whenever a replica exchange is accepted in Step 2, the permutation functions in equation (26) are updated.

The REM simulation is particularly suitable for parallel computers. Because one can minimize the amount of information exchanged among nodes, it is best to assign each replica to each node (exchanging pairs of temperature values among nodes is much faster than exchanging coordinates and momenta). This means that we keep track of the permutation function $m(i;t) = f^{-1}(i;t)$ in equation (26) as a function of MC or MD step t during the simulation. After parallel canonical MC or MD simulations for a certain steps (Step 1), $M/2$ pairs of replicas corresponding to neighboring temperatures are simultaneously exchanged (Step 2) and the pairing is alternated between the two possible choices, i.e. (T_1, T_2), (T_3, T_4), ... and (T_2, T_3), (T_4, T_5), ...

The major advantage of REM over other generalized-ensemble methods such as MUCA [3,4] and ST [5,6] lies in the fact that the weight factor is *a priori* known (equation (28)), while in the latter algorithms the determination of the weight factors can be very tedious and time-consuming. A random walk in the "temperature space" is realized for each replica, which in turn induces a random walk in the potential energy space. This alleviates the problem of getting trapped in states of energy local minima. In REM, however, the number of required replicas increases as the system size N increases (according to $\sqrt{N}$) [7]. This demands a lot of computer power for complex systems.

3. Predictions of transmembrane helix configurations of membrane proteins

In this section, we present the results of the applications of REM MC simulations to the prediction of membrane protein structures [18–21].

It is estimated that 20–30% of all genes in most genomes encode membrane proteins [22]. However, only a small number of detailed structures have been obtained for membrane proteins because of technical difficulties in experiments such as high quality crystal growth. Therefore, it is desirable to develop a method for predicting membrane protein structures by computer simulations.

Our method consists of two parts. In the first part, the amino-acid sequences of the transmembrane helix regions of the target protein are identified. It is already established that the transmembrane helical segments can be predicted by analyzing mainly the hydrophobicity of amino-acid sequences, without having any information about the higher-order structures. There exist many WWW servers such as TMHMM [22], MEMSAT [23], SOSUI [24] and HMMTOP [25] in which given the amino-acid sequence of a protein they judge whether the protein is a membrane protein or not and (if yes) predict the regions in the amino-acid sequence that correspond to the transmembrane helices.

In the second part, we perform a REM simulation of these transmembrane helices that were identified in the first part. Given the amino-acid sequences of transmembrane helices, we first construct ideal canonical α-helices (3.6 residues per turn) of these sequences. For our simulations, we introduce the following rather drastic approximations: (1) we treat the backbone of the α-helices as rigid body and only side-chain structures are made flexible; (2) we neglect the rest of the amino acids of the membrane protein (such as loop regions); and (3) we neglect surrounding molecules such as lipids. In principle, we can also use the MD method, but we employ the MC algorithm here. We update configurations with rigid translations and rigid rotations of each α-helix and torsion rotations of side chains. We use a standard force field such as CHARMM [26,27] for the potential energy of the system. We also add the following simple harmonic constraints to the original force-field energy:

$$\begin{aligned} E_{\text{constr}} = & \sum_{i=1}^{N_{\text{H}}-1} k_1 \theta\big(r_{i,i+1} - d_{i,i+1}\big)\big[r_{i,i+1} - d_{i,i+1}\big]^2 \\ & + \sum_{i=1}^{N_{\text{H}}} \Big\{ k_2 \theta\big(\big|z_i^{\text{L}} - z_0^{\text{L}}\big| - d_i^{\text{L}}\big)\big[\big|z_i^{\text{L}} - z_0^{\text{L}}\big| - d_i^{\text{L}}\big]^2 \\ & + k_2 \theta\big(\big|z_i^{\text{U}} - z_0^{\text{U}}\big| - d_i^{\text{U}}\big)\big[\big|z_i^{\text{U}} - z_0^{\text{U}}\big| - d_i^{\text{U}}\big]^2 \Big\} \\ & + \sum_{\text{C}_\alpha} k_3 \theta\big(r_{\text{C}_\alpha} - d_{\text{C}_\alpha}\big)\big[r_{\text{C}_\alpha} - d_{\text{C}_\alpha}\big]^2, \end{aligned} \tag{39}$$

where N_{H} is the total number of transmembrane helices in the protein and $\theta(x)$ is the step function:

$$\theta(x) = \begin{cases} 1, & \text{for } x \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{40}$$

and k_1, k_2 and k_3 are the force constants of the harmonic constraints, $r_{i,i+1}$ is the distance between the C atom of the C-terminus of the i-th helix and the N atom of the N-terminus of the $(i+1)$-th helix, z_i^{L} and z_i^{U}

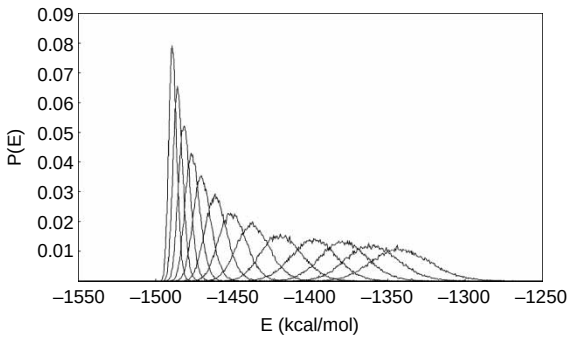

Figure 1. The canonical probability distributions of the total potential energy obtained from the replica-exchange MC simulation at 13 temperatures. The distributions correspond to the following temperatures (from left to right): 200, 239, 286, 342, 404, 489, 585, 700, 853, 1041, 1270, 1548, and 1888 K.

are the z-coordinate values of the C_α (or C) atom of the N-terminus (or C-terminus) of the i-th helix near the fixed lower boundary value z_0^{L} and the upper boundary value z_0^{U} of the membrane, respectively, r_{C_α} are the distance of C_α atoms from the origin and $d_{i,i+1}$, d_i^{L}, d_i^{U} and d_{C_α} are the corresponding central values of the harmonic constraints. The first term in equation (39) is the energy that constrains pairs of adjacent helices along the amino-acid chain not to be apart from each other too much (loop constraints). This term has a non-zero value only when the distance $r_{i,i+1}$ becomes longer than $d_{i,i+1}$.

The second term in equation (39) is the energy that constrains the helix N-teminus and C-terminus to be located near membrane boundary planes. This term has a non-zero value only when the C atom of each helix C-terminus and C_α atom of each helix N-terminus are apart more than d_i^{L} (or d_i^{U}). Base on the knowledge that most membrane proteins are placed in parallel, this constraint energy is included so that helices are not too apart from the perpendicular orientation with respect to the membrane boundary planes.

The third term in equation (39) is the energy that constrains all C_α atoms within the sphere (centered at the origin) of radius d_{C_α}. This term has a non-zero value only when C_α atoms go out of this sphere. The term is introduced so that the center of mass of the molecule stays near the origin. The radius of the sphere is set to a large value in order to guarantee that a wide configurational space is sampled.

In the first part of the present method, we obtain the amino-acid sequences of the transmembrane helix regions from existing WWW servers such as those in Refs. [22–25]. However, the precision of these programs in the WWW servers is about 85% and needs improvement. We thus focus our attention on the effectiveness of the second part of our method, leaving this improvement to the developers of the WWW servers. Namely, we use the experimentally known amino-acid sequence of helices (without relying on the WWW servers) and try to predict their conformations, following the prescription of the second part of our method. We selected the amino-acid sequence of the transmembrane dimer of glycophorin A (PDB code: 1AFO). The number of amino acids for each helix is 18 and the sequence is TLIIFGVMAGVIGTILLI.

At first, the ideal canonical α-helix (3.6 residues per turn) of this sequence was constructed. The N and C termini of this helix were blocked with the acetyl group and *N*-methyl group, respectively. The force field that we used is the CHARMM param19 parameter set (polar hydrogen model) [26,27]. No cutoff was introduced to the non-bonded energy terms and the dielectric constant

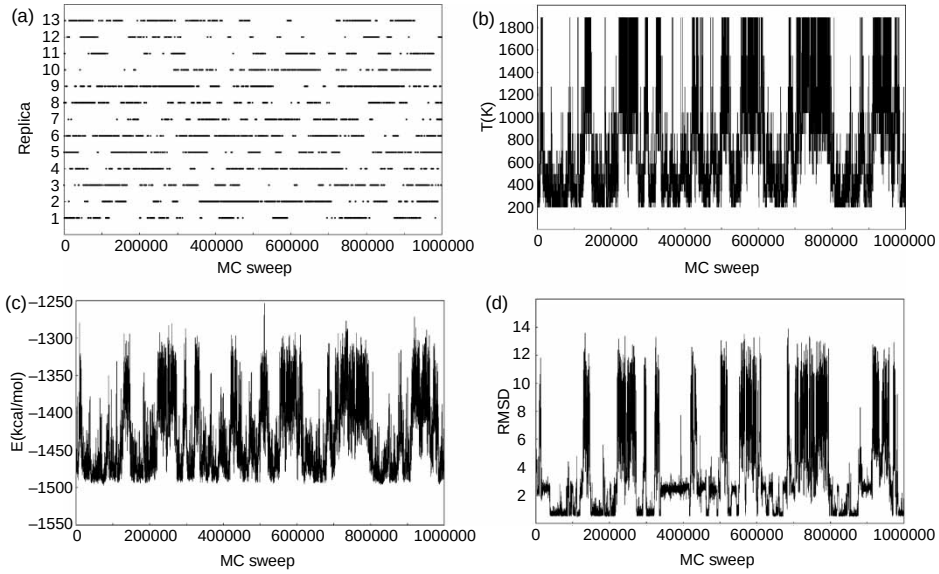

Figure 2. Time series of replica exchange at $T = 200$ K (a); temperature exchange for one of the replicas (Replica 8) (b); the total potential energy for Replica 8 (c); and the RMS deviation (in Å) of backbone atoms from the NMR structure for Replica 8 (d).

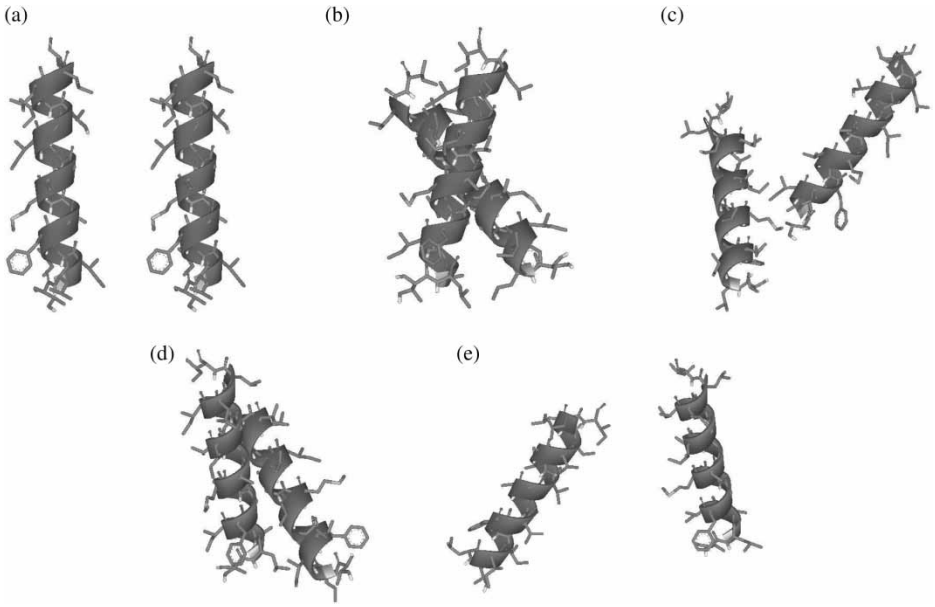

Figure 3. Typical snapshots from the REM simulation. (a) is the initial configuration.

ϵ was set equal to 1.0. The computer code based on the CHARMM macromolecular mechanics program [28] was used and the replica-exchange MC method was implemented in it.

The initial configuration for the REM simulation was that two α-helices of identical sequence and structure thus prepared were placed in parallel at a distance of 20 Å. These helices are quite apart from each other and the starting configuration is indeed very different from the native one. Note that the only information derived from the NMR experiments [29] is the amino-acid sequence of the individual helices.

The values of the constants for the constraints in equation (39) were set as follows: $N_{\mathrm{H}} = 2$, $k_1 = k_2 = 0.5\,\mathrm{kcal/(mol\,\mathring{A}^2)}$, $k_3 = 0.05\,\mathrm{kcal/(mol\,\mathring{A}^2)}$, $d_{i,i+1} = 20\,\mathring{\mathrm{A}}$, $z_0^{\mathrm{L}} = -13.35\,\mathring{\mathrm{A}}$, $z_0^{\mathrm{U}} = +13.35\,\mathring{\mathrm{A}}$, $d_i^{\mathrm{L}} = d_i^{\mathrm{U}} = 1.0\,\mathring{\mathrm{A}}$ and $d_{\mathrm{C}_\alpha} = 50\,\mathring{\mathrm{A}}$.

We performed a REM MC simulation of 1,000,000 MC sweeps, starting from this parallel configuration. We used the following 13 temperatures: 200, 239, 286, 342, 404, 489, 585, 700, 853, 1041, 1270, 1548, and 1888 K, which are distributed almost exponentially. The highest temperature was chosen sufficiently high so that no trapping in local-minimum-energy states occurs. This temperature

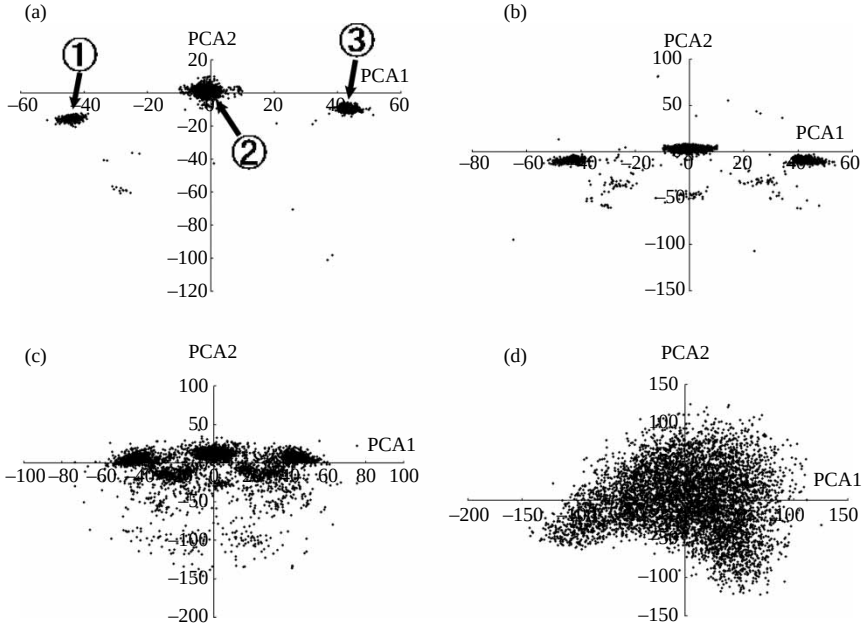

Figure 4. The projection of the sampled structures from the replica-exchange simulation onto the 1st and 2nd principal axes at the following temperatures: 200 K (a); 342 K (b); 585 K (c); and 1888 K (d). The circled numbers "1", "2", and "3" in (a) stand for Clusters 1, 2, and 3, respectively.

distribution was chosen so that all the acceptance ratios are almost uniform and sufficiently large (>10%) for computational efficiency. The backbone structures were fixed during simulations and the MC move types were taken to be rigid translation of each helix, rigid rotation of each helix and torsion-angle rotations of side chains.

In figure 1, the canonical probability distributions of the total potential energy obtained at the chosen 13 temperatures from the REM simulation are shown. We see that there are enough overlaps between all neighboring pairs of distributions, indicating that there will be sufficient numbers of replica exchange between pairs of replicas.

In figure 2(a), we show the "time series" of replica exchange at the lowest temperature ($T = 200$ K). We see that every replica takes the lowest temperature many times and we indeed observe a random walk in the replica space. The complementary picture to this is the temperature exchange for each replica. The results for one of the replicas (Replica 8) are shown in figure 2(b). We again observe a random walk in the temperature space between the lowest and highest temperatures. Other replicas perform random walks in the same way. In figure 2(c), the corresponding time series of the total potential energy is shown. We see that a random walk in the potential energy space between low and high energies is also realized. Note that there is a strong correlation between the behaviors in figure 1(b) and (c) as there should. All these results confirm that the present REM simulation has been properly performed.

We now study how widely the configurational space is sampled during the present simulation. For this purpose, we plot the time series of the root-mean-square (RMS) deviation of the backbone atoms from the NMR structure [29] in figure 2(d). When the temperature becomes high, the RMS deviation takes a large value (the largest value in figure 2(d) is 13.9 Å and the maximum value among all the replicas is 15.7 Å) and when the temperature becomes low, the RMS deviation takes a small value (the smallest value in figure 2(d) is 0.48 Å and the minimum value among all the replicas is 0.47 Å). By comparing figure 2(c) and (d), we see that there is a strong correlation between the total potential energy and the RMS deviation values. In particular, it is remarkable that when the energy is the lowest (around 1490 kcal/mol), most of the RMS values are as small as about 0.5 Å. This implies that the global-minimum-energy state is indeed very close to the native structure. Note also that the RMS values around 2.5 Å are also sampled at low temperatures. This suggests that there exists a local-minimum free energy state around 2.5 Å.

In figure 3, typical snapshots from the present REM simulation are shown. Figure 3(a) is the initial configuration of this simulation, in which the two helices are placed in parallel. We see that this simulation sampled many non-native configurations such as those in figure 3(c) and (e). At low temperatures low-energy configurations such as that in figure 3(b) and (d) with the side chains packed are sampled. We see that the REM simulation performs a random walk not only in energy space but also in conformational space and that it does not get trapped in one of a huge number of local-minimum-energy states.

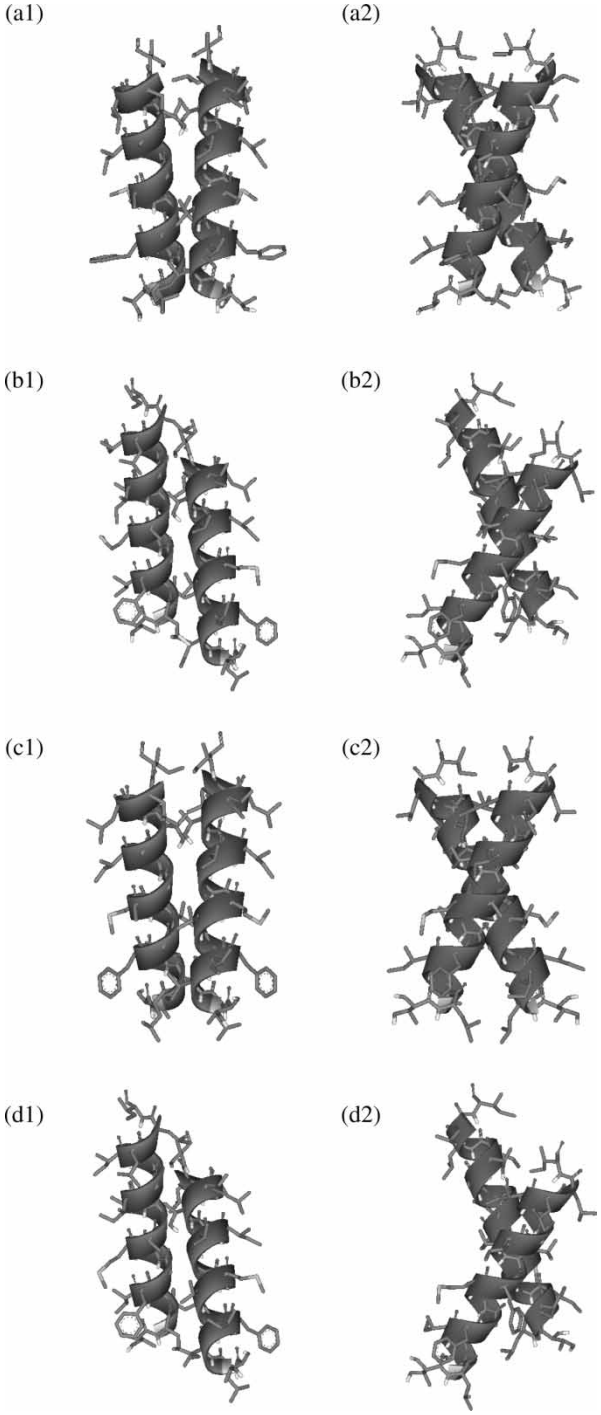

Figure 5. The NMR structure (Model 16 of the PDB code 1AFO) and typical cluster structures of the principal component analysis at the lowest temperature (200 K) from the REM simulation. (a1) and (a2) are the same structure viewed from different angles. Similarly, (b1) and (b2), (c1) and (c2) and (d1) and (d2) are the same structures viewed from different angles, respectively. (a) is the NMR structure. (b), (c) and (d) are the typical configurations of Cluster 1, Cluster 2, and Cluster 3, respectively. The figures were created with Viewer Lite.

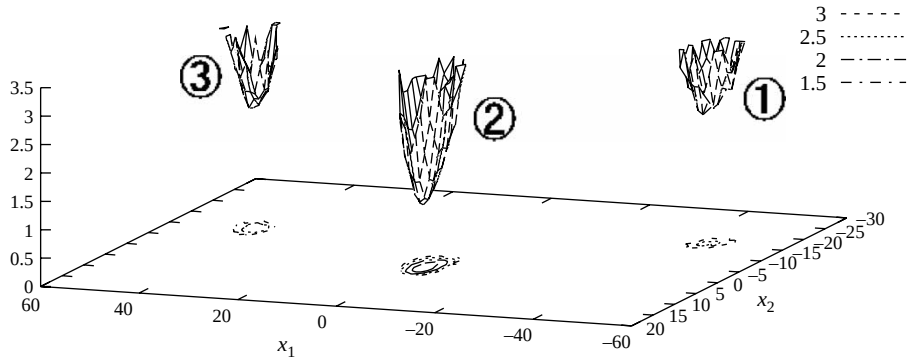

Figure 6. The free energy landscape $P(x_1, x_2)$ at 200 K calculated from the replica-exchange MC simulation. x_1 and x_2 correspond to the first principal component and the second principal component, respectively. The circled numbers "1", "2", and "3" stand for Clusters 1, 2, and 3, respectively.

In order to classify the low-energy conformations, we applied the principal component analysis (PCA) [30–34]. In figure 4, the structures obtained from the replica-exchange simulation are projected on the first and second principal component axes at chosen four temperatures. There are three distinct clusters at the lowest temperature in figure 4(a). As the temperature becomes higher, these clusters become less distinct and the first and second principal components become larger (note that the scales of both axes are expanded). This implies that as the temperature becomes higher, a wider conformational space is sampled without getting trapped in local-minimum free energy states. If we perform constant temperature simulations at the lowest temperature, the simulations will get trapped in one of the clusters in figure 4(a), depending on the initial configurations of the simulations. However, each replica of the replica-exchange simulations will not get trapped in the

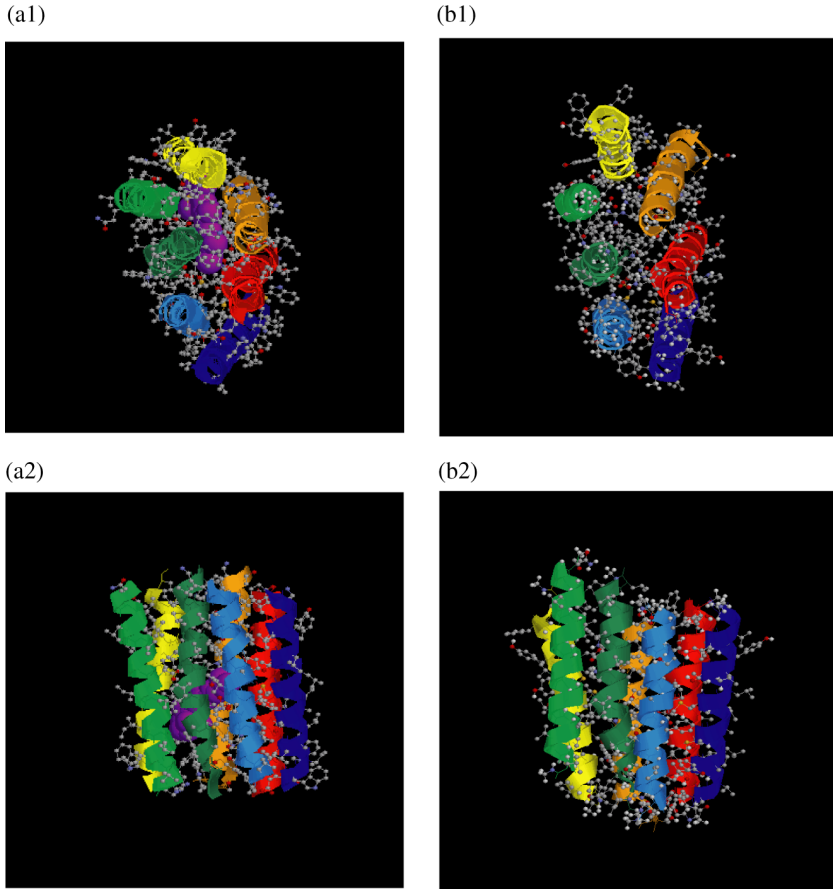

Figure 7. (a) The PDB structure of bacteriorhodopsin (PDB code: 1C3W) with retinal. (b) The smallest RMSD configuration that was obtained by the REM simulation. (a1), (a2) and (b1), (b2) are the same structures viewed from different angles (from top and from side), respectively. Purple-color atoms in (a) represent the retinal. (a) was drawn by eliminating the loop regions and lipids from the PDB file in online version. The RMSD of the structure in (b) from the native structure of (a) is 4.42 Å with respect to all C_α atoms. The figures were created with RasMol [35].

local-minimum free energy states, because the temperature of each replica goes up and down by temperature exchange. The three clusters in figure 4(a) lie in the ranges $(-60 \sim -30, -25 \sim -10)$, $(-20 \sim 20, -10 \sim 10)$ and $(30 \sim 60, -15 \sim 0)$, which we refer to as Cluster 1, Cluster 2, and Cluster 3, respectively.

Because this membrane protein is a dimer and the two helices consist of the same amino-acid sequences, one configuration can have two different numbering of atoms. Therefore, the PCA treats the same configurations as different clusters if it does not have the C_2-symmetry in structure. Cluster 2, which is close to the native structure, is C_2-symmetric and therefore becomes only one cluster, while Clusters 1 and 3 have no such symmetry and actually have the same structure.

In figure 5, the typical structure of each cluster and the solution NMR structure (PDB code: 1AFO [29]) are shown. We confirm that the structures of Clusters 1 and 3 are almost the same and the structures of Cluster 2 are indeed very close to the experimental one. The structures of Clusters 1 and 3 are a little off from the membrane boundary. In figure 6, the free energy surface with respect to the first and second principal component axes at 200 K is shown. We use the following equation to calculate the free energy as a function of the first and second principal components x_1 and x_2:

$$F(x_1, x_2) = -k_B T \ln P(x_1, x_2), \tag{41}$$

where k_B is the Boltzmann constant, T is absolute temperature and $P(x_1, x_2)$ is the probability to find the structure with the first and second principal component values x_1 and x_2. We see that Cluster 2 is the lowest free energy state. This figure shows that the cluster which is very close to the native structure has indeed the global-minimum free energy structure.

Finally, we present the results of a more complicated system, namely, bacteriorhodopsin [21]. We performed a REM MC simulation of 168,000,000 MC steps. We used the following 32 temperatures: 200, 218, 238, 260, 284, 310, 338, 369, 410, 455, 505, 561, 623, 691, 768, 853, 947, 1052, 1125, 1202, 1285, 1374, 1469, 1642, 1835, 2051, 2293, 2679, 3132, 3660, 4278, and 5000 K. This temperature distribution was chosen so that all the acceptance ratios are almost uniform and sufficiently large ($>10\%$) for computational efficiency. The highest temperature was chosen sufficiently high so that no trapping in local-minimum-energy states occurs. Replica exchange was attempted once at every 50 MC steps. Here, we just compare one of the local-minimum energy structure with the native structure in figure 7. They indeed are quite similar to each other.

4. Conclusions

In this article, we have reviewed the uses of the generalized-ensemble algorithm, which is a generic term for a powerful simulation algorithm that overcomes the multiple-minima problem based on non-Boltzmann weight factors. Detailed formulations of the three well-known generalized-ensemble algorithms, namely, MUCA, ST and REM, were given.

As examples, we presented the results of the application of the replica-exchange MC method to the predictions of membrane protein structures. We have shown that this method is indeed effective for molecular simulations of biomolecular systems.

Acknowledgements

The computations were performed on the computers at the Research Center for Computational Science, Institute for Molecular Science, Japan Atomic Energy Research Institute. This work was supported, in part, by the Grants-in-Aid for the Next Generation Super Computing Project, Nanoscience Program and for Scientific Research in Priority Areas, "Water and Biomolecules", from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] A. Mitsutake, Y. Sugita, Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolym. (Pept. Sci.)*, **60**, 96 (2001).
[2] Y. Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.*, **22**, 425 (2004).
[3] B.A. Berg, T. Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett.*, **B267**, 249 (1991).
[4] B.A. Berg, T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.*, **68**, 9 (1992).
[5] A.P. Lyubartsev, A.A. Martinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, **96**, 1776 (1992).
[6] E. Marinari, G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, **19**, 451 (1992).
[7] K. Hukushima, K. Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.*, **65**, 1604 (1996).
[8] K. Hukushima, H. Takayama, K. Nemoto. Application of an extended ensemble method to spin glasses. *Int. J. Mod. Phys. C*, **7**, 337 (1996).
[9] C.J. Geyer. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface*, E.M. Keramidas (Ed.), pp. 156–163, Interface Foundation, Fairfax Station (1991).
[10] E. Marinari, G. Parisi, J.J. Ruiz-Lorenzo. In *Spin Glasses and Random Fields*, A.P. Young (Ed.), pp. 59–98, World Scientific, Singapore (1998).
[11] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087 (1953).
[12] U.H.E. Hansmann, Y. Okamoto, F. Eisenmenger. Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble. *Chem. Phys. Lett.*, **259**, 321 (1996).
[13] N. Nakajima, H. Nakamura, A. Kidera. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B*, **101**, 817 (1997).

[14] A.M. Ferrenberg, R.H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.*, **61**, 2365 (1988); *ibid.* **63**, 1658 (1989).
[15] A.M. Ferrenberg, R.H. Swendsen. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, **63**, 1195 (1989).
[16] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman, J.M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules I. The method. *J. Comput. Chem.*, **13**, 1011 (1992).
[17] Y. Sugita, Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314**, 141 (1999).
[18] H. Kokubo, Y. Okamoto. Prediction of transmembrane helix configurations by replica-exchange simulations. *Chem. Phys. Lett.*, **383**, 397 (2004).
[19] H. Kokubo, Y. Okamoto. Prediction of membrane protein structures by replica-exchange Monte Carlo simulations: Case of two helices. *J. Chem. Phys.*, **120**, 10837 (2004).
[20] H. Kokubo, Y. Okamoto. Classification and prediction of low-energy membrane protein helix configurations by replica-exchange Monte Carlo method. *J. Phys. Soc. Jpn.*, **73**, 2571 (2004).
[21] H. Kokubo, Y. Okamoto. Self-assembly of transmembrane helices of bacteriorhodopsin by a replica-exchange Monte Carlo simulation. *Chem. Phys. Lett.*, **392**, 168 (2004).
[22] A. Krogh, B. Larsson, G. v. Heijne, E.L.L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567 (2001).
[23] D.T. Jones, W.R. Taylor, J.M. Thornton. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038 (1994).
[24] T. Hirokawa, S. Boon-Chieng, S. Mitaku. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378 (1998).
[25] G.E. Tusnady, I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489 (1998).
[26] W.E. Reiher III. Theoretical studies of hydrogen bonding. Ph.D. Thesis, Department of Chemistry, Harvard University, Cambridge, MA, USA (1985).
[27] E. Neria, S. Fischer, M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, **105**, 1902 (1996).
[28] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187 (1983).
[29] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman. A transmembrane helix dimer: structure and implications. *Science*, **276**, 131 (1997).
[30] M.M. Teeter, D.A. Case. Harmonic and quasiharmonic descriptions of crambin. *J. Phys. Chem.*, **94**, 8091 (1990).
[31] A. Kitao, F. Hirata, N. Go. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.*, **158**, 447 (1991).
[32] A.E. Garcia. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, **68**, 2696 (1992).
[33] R. Abagyan, P. Argos. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.*, **225**, 519 (1992).
[34] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen. Essential dynamics of proteins. *Proteins*, **17**, 412 (1993).
[35] R.A. Sayle, E.J. Milner-White. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374 (1995).