Biophys Rev (2012) 4:27–44
DOI 10.1007/s12551-011-0063-6

REVIEW

Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes

Junichi Higo · Jinzen Ikebe · Narutoshi Kamiya · Haruki Nakamura

Received: 23 September 2011 / Accepted: 23 November 2011 / Published online: 11 January 2012

Abstract Protein folding and protein–ligand docking have long persisted as important subjects in biophysics. Using multicanonical molecular dynamics (McMD) simulations with realistic expressions, i.e., all-atom protein models and an explicit solvent, free-energy landscapes have been computed for several systems, such as the folding of peptides/proteins composed of a few amino acids up to nearly 60 amino-acid residues, protein–ligand interactions, and coupled folding and binding of intrinsically disordered proteins. Recent progress in conformational sampling and its applications to biophysical systems are reviewed in this report, including descriptions of several outstanding studies. In addition, an algorithm and detailed procedures used for multicanonical sampling are presented along with the methodology of adaptive umbrella sampling. Both methods control the simulation so that low-probability regions along a reaction coordinate are sampled frequently. The reaction coordinate is the potential energy for multicanonical sampling and is a structural identifier for adaptive umbrella sampling. One might imagine that this probability control invariably enhances conformational transitions among distinct stable states, but this study examines the enhanced conformational sampling of a simple system and shows that reasonably well-controlled sampling slows the transitions. This slowing is induced by a rapid change of entropy along the reaction coordinate. We then provide a recipe to speed up the sampling by loosening the rapid change of entropy. Finally, we report all-atom McMD simulation results of various biophysical systems in an explicit solvent.

J. Higo (✉) · J. Ikebe · N. Kamiya · H. Nakamura
Institute for Protein Research, Osaka University,
Suita, Osaka 565-0871, Japan
e-mail: higo@protein.osaka-u.ac.jp

Keywords Molecular dynamics · Enhanced sampling · Generalized ensemble · Multicanonical · Canonical ensemble · Free-energy landscape

Introduction

Large-scale intramolecular conformational motions are necessary for protein folding, with large intramolecular translational/rotational motions causing protein–ligand binding. With the rapidly increasing capabilities of computers, the study of these motions has come to be an important computational task. To trace large motions, fast computers specialized for molecular simulations, such as MDGRAPE-3 (Narumi et al. 2006) and ANTON (Shaw et al. 2007; Maragakis et al. 2008), might be useful. An alternative useful approach is the use of a source program that is especially coded for rapid processing, such as GROMACS (van der Spoel et al. 2005). A generalized ensemble method is also an alternative means to accelerate conformational sampling (Mitsutake et al. 2001). This algorithmic approach is useful whether or not fast computers or suitable programs are used.

Protein conformational sampling is equivalent to an exploration of a conformational space, which is an abstract space used to completely express the structural variety of a protein. When the protein consists of N_{aa} amino-acid residues, the number of degrees of freedom to specify any allowable protein structure is approximately proportional to N_{aa}. It is likely that a power law of N_{aa} approximates the volume V_{cs} of the conformational space for the protein as

$$V_{cs} \propto s_p^{Naa}, \tag{1}$$

where s_p is a constant that is specific to the protein system. Then, V_{cs} increases rapidly with increasing N_{aa}, and the conformational sampling is confronted with a difficulty.

Actually, rigorous all-atom computations of peptides in explicit solvent have demonstrated that three-residue prolongation of N_{aa} expands V_{cs} by tenfold (Ikebe et al. 2011a). Furthermore, numerous low-energy basins (energetically stable structures) or narrow energetic pinholes distributed widely in the conformational space trap the conformation during the simulation. A larger basin has a lower free energy than a smaller one. Consequently, it is desirable that the sampling method be able to both escape from basins and also be able to measure the basin size.

Protein simplification is a useful computational technique to study the overall features of protein folding (Go 1983; Dill 1985; Miyazawa and Jernigan 1985; Bryngelson et al. 1995). The Go-like model (Go 1983) modulates the potential energy in advance so that the native protein structure has the lowest energy. It can then predict the folding core regions (Koga and Takada 2001) and the folding kinetics (Munoz and Eaton 1999) for two-state proteins. The smoothing of the potential energy surface speeds up the protein conformational motions considerably. In an all-atom simulation, in contrast, the conformation is easily trapped in a local energy basin during a prolonged simulation time: once the conformation escapes from the basin, another basin traps it, and so on. This repetition of trapping-and-escape is likely to be the real picture of the protein folding occurring in a time scale that is too short to be identified experimentally. Consequently, all-atom simulation is an indispensable research step to ascertain the details of molecular events.

When a protein is bound to its partner molecule, the two molecules move in space to form a complex. In theory, numerous complex modes are possible. Additionally, once a complex is formed in the simulation, the thermodynamic stability of the complex should be examined. Consequently, the sampling is expected to be sufficiently powerful to produce various complexes and to estimate the binding free energy for each complex accurately. Intrinsically disordered proteins (IDPs), classified as a new protein group, are structurally disordered in the free state (unbound state) and adopt well-defined tertiary structures upon binding to their partner molecules (Wright and Dyson 1999; Sugase et al. 2007). In terms of IDP function, therefore, the binding is coupled indivisibly with the folding. As the time scale for this process is too short to be traced experimentally, a computer simulation is a key approach to study this process. However, one must solve the folding and binding in parallel, which necessitates higher sampling efficiency than solving either folding or binding alone.

As described in this paper, we review the multicanonical simulation, which is compatible with the all-atom treatment of proteins in explicit solvent. This method can assign free energies (i.e., statistical weights) to the energy basins. Therefore, a realistic free-energy landscape is obtained by mapping/projecting the sampled conformations in the conformational space. The multicanonical algorithm was originally introduced to study a physical system, namely, a two-dimensional Potts model (Berg and Neuhaus 1992), and was applied to polypeptide systems combined with a Monte Carlo simulation (Hansmann and Okamoto 1993; Kidera 1995). Subsequently, the algorithm was combined with a molecular dynamics (MD) simulation to study large fluctuations of a peptide (Hansmann et al. 1996; Nakajima et al. 1997b). Nakajima's MD version, denoted as McMD in this paper, solves the Newtonian equations in Cartesian coordinate space, while Hansmann's version integrates the equations in a dihedral-angular space. In the all-atom treatment, a protein consists of densely packed atoms, and solvent atoms tightly surround the protein. The Monte Carlo simulation is unsuitable for such a crowded-atom system because most trial conformations result in rejection by atomic bumps. Consequently, the adoption of MD is extremely important. The McMD simulation has been applied to various systems, from a two-residue peptide (Nakajima et al. 2000) to a 57-residue protein (Ikebe et al. 2011b), and applied to protein–ligand flexible docking (Nakajima et al. 1997a, b; Kamiya et al. 2008). A trajectory-parallelization method has been developed (Higo et al. 2009; Ikebe et al. 2011a) to increase the sampling efficiency still further, and this method has been applied to the coupled folding and binding of an IPD to generate the free-energy landscape (Higo et al. 2011).

Another useful simulation method used to generate the free-energy landscape is the replica-exchange method (REM). Herein we briefly mention this method, although we do not specifically examine this method in this review. REM was introduced to study an Ising spin glass system combined with the Monte Carlo simulation (Hukushima and Nemoto 1996) and applied to a biological system combined with canonical MD (Sugita and Okamoto 1999). A user executes multiple runs of the same system (replicas) at different temperatures in parallel and tries to exchange the temperatures among different replicas with reference to a physicochemical exchange probability between the replicas. When the exchange is accepted frequently, the replicas are relaxed thermally, and the sampled conformations are used to generate the free-energy landscape. To increase the exchange probability, the replica-exchange and multicanonical methods are combined (Sugita and Okamoto 2000) or a microcanonical MD version is used (Kar et al. 2009). An optimal choice of replica set has been discussed (Trebst et al. 2006). Furthermore, this method was generalized for exchanging parameters other than temperature (Sugita et al. 2000), and it was extended to a Hamiltonian-exchange form (Fukunishi et al. 2002). Another generalized-ensemble replica-exchange method has been proposed to focus on the first-order transition phenomena (Kim et al. 2010).

As explained later, the multicanonical simulation controls the sampling so that the energy distribution converges to a desired form. One can imagine a simulation in which a distribution function of another parameter than energy converges to a desired form, as introduced by Paine and Scheraga (1985) or Mezei (1987). This sampling method is now called 'the adaptive umbrella (AU) sampling'. The multicanonical and AU sampling methods therefore are similar in terms of their methodology. This review report describes the methodology of AU sampling as well as that of multicanonical sampling.

In this review, we begin with an introductory/preparative section ('Preparation') in which we provide a general explanation of conformational sampling; this is followed by two sections 'Adaptive umbrella sampling' and 'Multicanonical sampling', respectively, which describe in detail the theory of AU and multicanonical methods. We then solve a simple protein–ligand docking problem to show that the AU method does not always enhance sampling ('Traffic slowing in enhanced sampling') and provide a recipe to drastically increase the sampling efficiency ('Traffic enhancement'). This recipe provides an important supplement tor both the AU and multicanonical methods. Next, we explain actual procedures for multicanonical and AU methods (sections 'Actual procedure for multicanonical sampling' and 'Actual procedure for adaptive umbrella sampling') and provide some technical sections ('Methods to update the canonical distribution', 'TTP-multicanonical sampling', and 'Other computational techniques'). After the free-energy landscape ('Free-energy landscape') is explained we further describe the results of the McMD simulations of various biophysical systems expressed by the all-atom model in explicit solvent ('All-atom McMD simulations of various systems').

Preparation

In this section, we provide a general description of conformational sampling to produce a canonical ensemble, which is linked smoothly to the discussion in the next section on enhanced conformational sampling.

Consider that a system consists of biomolecular and solvent-molecular atoms. We express the position of atom i in the system by its Cartesian coordinates: x_i, y_i, and z_i. Then, a microscopic state of the system is expressed completely using a vector $\boldsymbol{r}$ as

$$r = [x_1, y_1, z_1, x_2, y_2, z_2, \ldots, z_N, y_N, z_N], \tag{2}$$

where N is the number of atoms in the system. Consequently, the microscopic state is assigned to a position of the N-dimensional conformational space. The conformational sampling is equivalent to move $\boldsymbol{r}$ in the N-dimensional space with a transition rule among the microscopic states. We schematically present the transition between two microscopic states $\mathrm{m_A}$ and $\mathrm{m_B}$ as

$$\mathrm{m_A} \underset{k_\mathrm{B}}{\overset{k_\mathrm{A}}{\rightleftarrows}} \mathrm{m_B}, \tag{3}$$

where k_A and k_B respectively represent the rate constants (kinetic constants) for the $\mathrm{m_A}$-to-$\mathrm{m_B}$ transition and its inverse process. We refer to the positions of $\mathrm{m_A}$ and $\mathrm{m_B}$ in the N-dimensional space as $\boldsymbol{r}_\mathrm{A}$ and $\boldsymbol{r}_\mathrm{B}$, respectively. Equation 3 is re-expressed using a couple of differential equations (reaction equations) as

$$\begin{cases} \frac{d\rho(r_\mathrm{A},t)}{dt} = -k_\mathrm{A}\rho(r_\mathrm{A},t) + k_\mathrm{B}\rho(r_\mathrm{B},t) \\ \frac{d\rho(r_\mathrm{B},t)}{dt} = k_\mathrm{A}\rho(r_\mathrm{A},t) - k_\mathrm{B}\rho(r_\mathrm{B},t) \end{cases}, \tag{4}$$

where $\rho(\boldsymbol{r},t)$ is a probability assigned to $\boldsymbol{r}$ at time t. We assume that the system reaches equilibrium for $t \to \infty$:

$$\lim_{t\to\infty} \rho(r,t) = \rho_\mathrm{c}(r). \tag{5}$$

Then, Eq. 4 is reduced to a single equation.

$$\frac{\rho_\mathrm{c}(r_\mathrm{A})}{\rho_\mathrm{c}(r_\mathrm{B})} = \frac{k_\mathrm{B}}{k_\mathrm{A}} \tag{6}$$

If the canonical ensemble characterizes the equilibrium, then $\rho_\mathrm{c}(\boldsymbol{r})$ is given by the Boltzmann factor as

$$\rho_\mathrm{c}(r,T) = A_\mathrm{c}\, exp\left[-\frac{E(r)}{RT}\right], \tag{7}$$

where $E(\boldsymbol{r})$ denotes the potential energy at $\boldsymbol{r}$, T the temperature of the system, R represents the gas constant (energy is expressed in kcal/mol in this study), and A_c is a normalization constant (or an inverse of the partition function). The probabilities at $\boldsymbol{r}_\mathrm{A}$ and $\boldsymbol{r}_\mathrm{B}$ are then given formally and respectively as $\rho_\mathrm{c}(\boldsymbol{r}_\mathrm{A},T)=A_\mathrm{c}\ \exp[-\mathrm{E}(\boldsymbol{r}_\mathrm{A})\ /\ RT]$ and $\rho_\mathrm{c}(\boldsymbol{r}_\mathrm{B},T)=A_\mathrm{c}\ \exp[-\mathrm{E}(\boldsymbol{r}_\mathrm{B})\ /\ RT]$. We obtain the following relation for the rate constants:

$$\frac{k_\mathrm{B}}{k_\mathrm{A}} = \exp\left[-\frac{\Delta E}{RT}\right], \tag{8}$$

where $\Delta E = E(\boldsymbol{r}_\mathrm{A}) - E(\boldsymbol{r}_\mathrm{B})$. Equation 8 is usually called the *detailed valance* between microscopic states, and it does not determine k_A and k_B individually. Nonetheless, Eq. 8 guarantees that a sufficiently long simulation trajectory converges to the canonical ensemble (Eq. 7) independent of the initial simulation configuration. The ensemble from either set of $[k_\mathrm{A}, k_\mathrm{B}]$ and $[ck_\mathrm{A}, ck_\mathrm{B}]$ ($\mathrm{c}\neq 0$) converges to the same distribution sooner or later.

In a Monte Carlo (MC) simulation, the rate constants are usually set as shown below.

$$[k_\mathrm{A},\ k_\mathrm{B}] = \begin{cases} [e^{\Delta E/RT},\ 1] & (for\ E(r_\mathrm{A}) \le E(r_\mathrm{B})) \\ [1,\ e^{-\Delta E/RT}] & (for\ E(r_\mathrm{A}) > E(r_\mathrm{B})) \end{cases} \tag{9}$$

In a MD simulation, $\boldsymbol{r}$ moves according to the Newtonian equation, and Eq. 9 is not used. The force $\boldsymbol{f}_i$ acting on atom i is given by a gradient of the potential energy as

$$f_i = -grad_i E(r) = -e_x \frac{\partial E(r)}{\partial x_i} - e_y \frac{\partial E(r)}{\partial y_i} - e_z \frac{\partial E(r)}{\partial z_i}, \tag{10}$$

where $\boldsymbol{e}_x$, $\boldsymbol{e}_y$, and $\boldsymbol{e}_z$ respectively represent the unit vectors parallel to the x-, y- and z-coordinate axes. It has not been generally proved that an MD simulation trajectory always converges to the canonical distribution $\rho_c(\boldsymbol{r},T)$. However, many MD studies have assumed convergence because energy dissipation occurs extensively in the atom-crowded system (biological system) when the simulation temperature is controlled appropriately (Evans and Morriss 1983; Nose 1984; Hoover 1985).

The MD and MC methods described above are generally regarded as *canonical sampling* and the sampled conformations as a *canonical ensemble*. However, canonical sampling does not guarantee a quick convergence of the simulation trajectory to the canonical ensemble. In fact, very slow convergence is often experienced when a large and complicated system is simulated. To avoid this difficulty, enhanced conformational sampling has been proposed.

Adaptive umbrella sampling

The energy surface of a biological system is generally vast and bumpy. Therefore, acceleration for the sampling is crucial. Here we introduce a modified potential energy $h(\boldsymbol{r})$, which is an arbitrary single-valued function that is differentiable with respect to the atomic coordinates $\{x_1,\ldots,z_N\}$. Accordingly, the detailed balance between the microscopic states m_A and m_B is defined as

$$\frac{k_B}{k_A} = \exp\left[-\frac{\Delta h}{RT}\right], \tag{11}$$

where $\Delta h = h(\boldsymbol{r}_A) - h(\boldsymbol{r}_B)$. Then a long simulation trajectory converges to a non-Boltzmann distribution as

$$\rho_h(r,T) = A_h \exp\left[-\frac{h(r)}{RT}\right], \tag{12}$$

where A_h is a normalization constant. In performing an MD simulation, the force acting on atom i is given as $\boldsymbol{f}_i = -grad_i h(\boldsymbol{r})$. The canonical distribution $\rho_c(\boldsymbol{r},T)$ is computed readily from $\rho_h(\boldsymbol{r},T)$ as

$$\rho_c(r,T) = A_{ch} \exp\left[-\frac{E(r)-h(r)}{RT}\right] \rho_h(r,T), \tag{13}$$

where A_{ch} is a normalization constant.

The switching of the detailed balance from Eq. 8 to Eq. 11 varies the rate constants among the microscopic states. This variation might accelerate the sampling when the function form of $h(\boldsymbol{r})$ is set carefully. However, the adjustment of $h(\boldsymbol{r})$ for the acceleration is a difficult task because the detailed balance should be modulated consistently among a very large number of the microscopic states in the system. To control the sampling more practically, we contract $\rho_c(\boldsymbol{r},T)$ to a one-dimensional (1D) distribution for a structural parameter λ as

$$P_c(\lambda,T) = A_{\lambda c} \int D(a(r)-\lambda)\, \rho_c(r,T) dr, \tag{14}$$

where $A_{\lambda c}$ is a normalization constant, $a(\boldsymbol{r})$ is an arbitrary function of $\boldsymbol{r}$, and $D(a(\boldsymbol{r}) - \lambda)$ is defined as

$$D(a(r)-\lambda) = \begin{cases} 1/V_\lambda & (\text{for regions of } a(r)=\lambda) \\ 0 & (\text{elsewhere}) \end{cases}, \tag{15}$$

where V_λ is a volume of the regions of $a(\boldsymbol{r})=\lambda$ in the N-dimensional space expressed as

$$V_\lambda = \int_{a(r)=\lambda} dr. \tag{16}$$

Integration in this equation is taken over regions of $a(\boldsymbol{r})=\lambda$. When the equation $a(\boldsymbol{r})=\lambda$ represents an $(N-1)$-dimensional hypersurface in the N-dimensional space, $D(a(\boldsymbol{r}) - \lambda)$ is reduced to a delta function: $\delta(a(\boldsymbol{r}) - \lambda)$. Equation 14 shows that $P_c(\lambda,T)$ is an accumulation of the canonical probabilities $\rho_c(\boldsymbol{r},T)$ within the regions of $a(\boldsymbol{r})=\lambda$.

To control the 1D distribution, AU sampling (Paine and Scheraga 1985; Mezei 1987) was developed by introducing a potential function E_u as

$$E_u(r) = E(r) + RT \ln[P_c(\lambda,T)]. \tag{17}$$

Then, the equilibrated probability assigned to a microscopic state is given formally as

$$\rho_u(r,T) = A_u \exp\left[-\frac{E_u(r)}{RT}\right] = \frac{A_u}{P_c(\lambda,T)} \exp\left[-\frac{E(r)}{RT}\right] = A_u \frac{\rho_c(r,T)}{P_c(\lambda,T)}, \tag{18}$$

where A_u is a normalization constant. The 1D contraction of $\rho_u(\boldsymbol{r},T)$ on the parameter axis λ produces a uniform distribution as follows.

$$P_u(\lambda,T) = \int D(a(r)-\lambda)\rho_u(r,T)dr = \frac{A_u}{P_c(\lambda,T)} \int_{a(r)=\lambda} \rho_c(r,T)dr = const \tag{19}$$

Equation 19 shows that a sufficiently long simulation produces a flat distribution on the λ axis. This property of $E_u(\boldsymbol{r})$ may enhance the sampling in a following situation: presuming that the canonical distribution $P_c(\lambda,T)$ is a bimodal distribution function (broken line in Fig. 1a) where the conformation is stable at around λ_1 and λ_2 and unstable at around $\lambda=\lambda_{mid}$ (Fig. 1a), then the transitions between the stable states might be rare in the canonical sampling. In contrast, $P_u(\lambda,T)$ is flat (solid line in Fig. 1a). Therefore, we expect that the inter-state transitions using $E_u(\boldsymbol{r})$ are more frequent than those obtained by canonical sampling, as presented in Fig. 1b. Figure 1 was prepared so that λ, called the *reaction coordinate*, is a good parameter to discriminate the stable and unstable states.

The usual aim of AU sampling is to generate a flat 1D distribution on the λ axis at equilibrium (Eq. 19). However, one might want to generate a non-flat distribution instead of a flat one. We can see that the flat distribution is a particular case of the non-flat distribution. We therefore redefine the modified potential function $E_u(\boldsymbol{r})$ as

$$E_u(r) = E(r) + RT \ln\left[\frac{P_c(\lambda,T)}{g(\lambda)}\right], \tag{20}$$

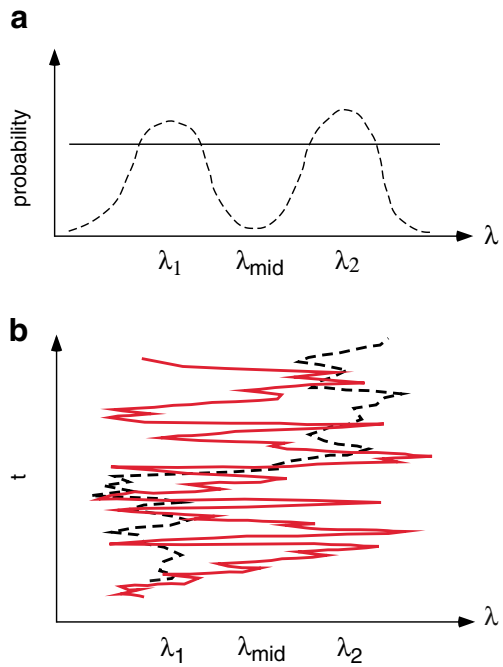

Fig. 1 **a** The one-dimensional (1D) probability distribution as a function of structural parameter λ. λ_1, λ_2, and λ_{mid} are explained in the main text. *Broken line* Canonical distribution P_c obtained from canonical sampling with the original potential energy $E(\boldsymbol{r})$, *solid line* flat distribution P_u from adaptive umbrella (AU) sampling with the modified potential energy $E_u(\boldsymbol{r})$. **b** Time (t) development of conformation on the λ axis. Broken and red solid lines represent results obtained using the canonical and AU sampling methods, respectively

where $g(\lambda)$ is an arbitrary single-valued function differentiable with respect to λ. The simulation generates the following distribution at equilibrium.

$$\begin{aligned} P_u(\lambda,T) &= A_u \int D(a(r)-\lambda)\exp\left[-\frac{E_u(r)}{RT}\right]dr \\ &= \frac{A_u g(\lambda)}{P_c(\lambda,T)}\int_{a(r)=\lambda}\exp\left[-\frac{E}{RT}\right]dr \\ &= A_u g(\lambda) \end{aligned} \tag{21}$$

The detailed valance for MC is

$$\frac{k_B}{k_A} = \exp\left[-\frac{\Delta E_u}{RT}\right], \tag{22}$$

where $\Delta E_u=E_u(\boldsymbol{r}_A)-E_u(\boldsymbol{r}_B)$. The force for MD is

$$f_i = -grad_i E_u(r) = -grad_i E(r) - RTgrad_i \ln\left[\frac{P_c(\lambda,T)}{g(\lambda)}\right]. \tag{23}$$

The term $-grad_i E(\boldsymbol{r})$ is the force derived from the original potential energy (Eq. 10). The other term can be arranged as

$$\begin{aligned} -RTgrad_i \ln\left[\frac{P_c(\lambda,T)}{g(\lambda)}\right] &= \frac{-RTg(\lambda)}{P_c(\lambda,T)} grad_i\left[\frac{P_c(\lambda,T)}{g(\lambda)}\right] \\ &= \frac{-RTg(\lambda)}{P_c(\lambda,T)}\frac{\partial}{\partial\lambda}\left[\frac{P_c(\lambda,T)}{g(\lambda)}\right] \\ &\quad \times grad_i a(r). \end{aligned} \tag{24}$$

We have not specified the function form of $a(\boldsymbol{r})$ because it should be set according to the problem to be solved. When the parameter λ is specific only to the protein conformation, $a(\boldsymbol{r})$ involves no solvent coordinates; then, the gradient with respect to the solvent coordinates is zero.

Multicanonical sampling

Because λ is an implicit function of $\boldsymbol{r}$, $E_u(\boldsymbol{r})$ controls the fluctuation of λ, but it cannot control the energy (E) fluctuations. To control these energy fluctuations, we introduce another modified potential energy E_{mc} as

$$E_{mc}(E) = E + RT\ln[P_c(E,T)], \tag{25}$$

where P_c is the canonical energy distribution at T (i.e., the contracted distribution on the energy axis).

$$P_c(E,T) = A_E \int n(E)\exp\left[-\frac{E(r)}{RT}\right]dr = A_E n(E)\exp\left[-\frac{E}{RT}\right] \tag{26}$$

In those equations, A_E is a normalization constant. The function $n(E)$ is the density of states: i.e., the number of

microscopic states in an iso-potential energy shell $[E, E+\Delta E]$ in the N-dimensional conformational space is given by $n(E)\mathrm{d}E$. The E_{mc} is rewritten using $n(E)$ as

$$E_{\mathrm{mc}}(E) = RT\ln[n(E)]. \tag{27}$$

A long simulation using E_{mc} gives the following energy distribution

$$P_{\mathrm{mc}}(E,T) = A_{\mathrm{mc}} n(E)\exp\left[-\frac{E_{\mathrm{mc}}}{RT}\right] = A_{\mathrm{mc}}\frac{n(E)}{n(E)} = const, \tag{28}$$

where A_{mc} is a normalization constant. This simulation is called multicanonical simulation or multicanonical sampling.

The aim of multicanonical sampling is to speed up energy relaxation. In this context, therefore, multicanonical sampling does not directly aim to speed up structural relaxation. However, energy barriers separate thermodynamically stable structures in the N-dimensional conformational space. Consequently, structural relaxation is related to energy relaxation. Figure 2a shows the conformational space characterized by E. In canonical sampling at a high temperature T_{high}, the conformation ascends into the high-energy regions without descending into energy barriers (red line in Fig. 2a), and the energy distribution $P_{\mathrm{c}}(E,T_{\mathrm{high}})$ is narrow (red line in Fig. 2b). Consequently, the room temperature (T_{room}) structures in the oblique-line region are seldom sampled. In contrast, at a low temperature T_{low}, the conformation is trapped in an energy basin (blue line in Fig. 2a), and the

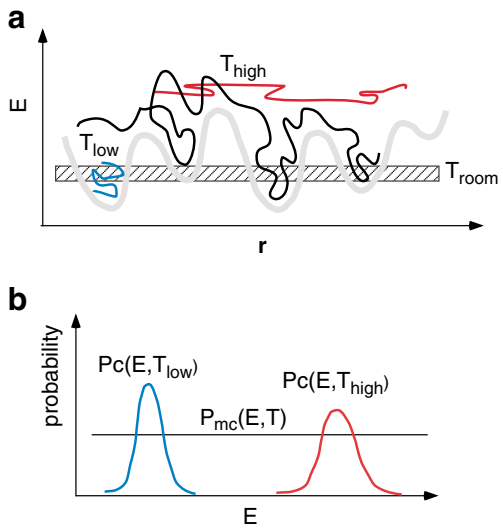

Fig. 2 **a** Energy (E) and structural (r) fluctuations from the high-temperature (T_{high}) canonical simulation (*red line*), low-temperature (T_{low}) canonical simulation (*blue*), and multicanonical simulation (*black*). *Gray line* Energy surface, *oblique-line region* room temperature (T_{room}) range. **b** The energy probability distribution $P_{\mathrm{c}}(E,T_{high})$ from the high-temperature canonical sampling (*red line*), the low-temperature sampling $P_{\mathrm{c}}(E,T_{low})$ (*blue*), and the distribution $P_{\mathrm{mc}}(E,T)$ from multicanonical sampling (*black*)

energy distribution $P_{\mathrm{c}}(E,T_{\mathrm{low}})$ is narrow (blue line in Fig. 2b). Therefore, the escape from the basin requires a considerably long simulation time. Although we might obtain some room temperature structures in this basin, we cannot judge whether those structures are biophysically more important than those in other basins because the trajectory visited only one basin. Multicanonical sampling explores both the high-energy regions and low-energy basins (black line in Fig. 2a), yielding a flat energy distribution P_{mc} (black line in Fig. 2b).

The modified potential E_{mc} involves P_{c} (see Eq. 25), but the function form of P_{c} is unknown when we start the simulation. Consequently, P_{c} is estimated self-consistently during the simulation, as explained later. At all events, once P_{c} is given accurately in a wide energy range, P_{mc} resultant from a long run is flat in this range. Although $P_{\mathrm{c}}(E,T)$ is the distribution specific to the simulation temperature T, we can convert it to $P_{\mathrm{c}}(E,T_{\mathrm{a}})$ at another temperature T_{a} as

$$\begin{aligned} P_{\mathrm{c}}(E,T_{\mathrm{a}}) &= A_{\mathrm{c}} n(E)\exp\left[-\frac{E}{RT_{\mathrm{a}}}\right] \\ &= A_{\mathrm{c}} P_{\mathrm{c}}(E,T)\exp\left[\frac{E}{RT}-\frac{E}{RT_{\mathrm{a}}}\right]. \end{aligned} \tag{29}$$

We used Eq. 26 to obtain this equation.

The final process in this section is to expand multicanonical sampling to yield a non-flat energy distribution $g(E)$, as was done for AU sampling (see Eq. 20). The modified potential energy is redefined as

$$E_{\mathrm{mc}}(E) = E + RT\ln\left[\frac{P_{\mathrm{c}}(E,T)}{g(E)}\right]. \tag{30}$$

The simulation trajectory with this potential energy converges to $g(E)$ as

$$\begin{aligned} P_{\mathrm{mc}}(E,T) &= A_{\mathrm{mc}}\int \delta(E'(r)-E)\exp\left[-\frac{E_{\mathrm{mc}}}{RT}\right]dr \\ &= A_{\mathrm{mc}}\frac{n(E)}{n(E)}g(E) = A_{\mathrm{mc}}g(E). \end{aligned} \tag{31}$$

For the McMD simulation at T, the atomic forces are defined as

$$\begin{aligned} f_i &= -grad_i E_{\mathrm{mc}}(r) = -grad_i E(r) - RTgrad_i\ln\left[\frac{P_{\mathrm{c}}(E,T)}{g(E)}\right] \\ &= -grad_i E(r) - RT\frac{g(E)}{P_{\mathrm{c}}(E,T)}\left[\frac{d}{dE}\frac{P_{\mathrm{c}}(E,T)}{g(E)}\right]grad_i E(r) \\ &= -grad_i E(r)\left[1 + RT\frac{g(E)}{P_{\mathrm{c}}(E,T)}\left\{\frac{d}{dE}\frac{P_{\mathrm{c}}(E,T)}{g(E)}\right\}\right]. \end{aligned} \tag{32}$$

The AU sampling procedure is effective when an essential reaction coordinate is known, along which biophysically

important structures are well discriminated. Multicanonical sampling is suitable to sample the entire conformational space and to generate the entire free-energy landscape. We can identify thermodynamically important energy basins and the free-energy barriers in the conformational ensemble at a desired temperature.

Traffic slowing in enhanced sampling

Enhanced conformational sampling controls the distribution as $P_u(\lambda,T)=g(\lambda)$ or $P_{mc}(E,T)=g(E)$. Therefore, the sampling indirectly controls the traffic of conformation along the λ or E axis as a by-product of the probability control (see Figs. 1b and 2a). Below we solve a simple protein–ligand docking problem by AU sampling and show that the probability-control does not always enhance the traffic, contrary to our expectations.

Here we consider a simple system mimicking protein–ligand binding. First, we prepare a large box designated as 'LC' in Fig. 3a. The x-, y-, and z-coordinate axes are defined so that they are parallel to the box sides, with the origin set on the body center of the LC box. Next, the LC box is divided into 3D cubic lattices with dimensions of 241^3, where 241 (= 2×120+1) lattice points line up along each of the coordinate axes. The ligand is represented as a particle (open circle in Fig. 3a) moving on the 3D lattice points. The ligand position (r_x, r_y, r_z) is then conditioned as $-120\leq r_x\leq 120$, $-120\leq r_y\leq 20$, and $-120\leq r_z\leq 20$. The smaller box, designated as 'PC' in Fig. 3a, is the protein of which the dimensions are 7^3: seven lattice points line up along each of the x-, y-, and z-axes. Figure 3b shows a cross-section (x–y plane with z=0) of the system. The body center of PC is set at the coordinate origin, at which the ligand-binding site (filled circle of Fig. 3b) is also set. A cuboid-shaped hole, mimicking the ligand binding cleft, is caved on the plane of x=3 of PC, for which the dimensions are 5×3×3 (Fig. 3b). The ligand then accesses the binding site through the hole. We also assume that the ligand can access the lattice points on the protein surface (open circles in Fig. 3b) but cannot enter into the protein interior (gray region in Fig. 3b). There are 89 sites in the inhibited region. The number of the accessible sites for the ligand is then 13,997,432 (= 241^3 – 89). We set the potential energy as zero (E=0) at any accessible site to assess an entropy effect in the sampling. Below we examine two sampling methods: non-enhanced sampling and AU sampling.

The non-enhanced sampling is a conventional Monte-Carlo sampling. The ligand was initially put randomly at a lattice point with excluding a case in which the ligand is buried in the protein interior and then moved randomly to the nearest neighbor lattice points. The moves were accepted unconditionally (remember that E is always zero), except

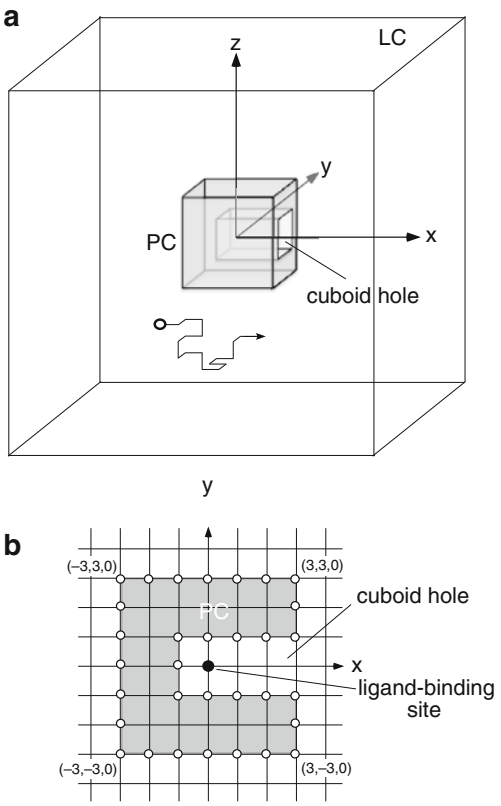

Fig. 3 **a** Overview of the system consisting of protein (*PC*) and ligand (*open circle*) confined in a large cubic box (*LC*). The origin of the x-, y-, and z-axes (*arrows x*, *y*, *z*, respectively) is set at the center of LC. The center of PC is also set on the origin. *Zigzag line* Ligand motions. The *cuboid hole* in PC mimics a cleft through which the ligand access to the ligand-binding site. **b** Cross section [x–y plane (*arrows x*, *y*, respectively, with z=0] of PC to show the cuboid hole and the ligand-binding site (*filled circle*). *Lattice points*, labeled such as (−3, 3, 0), are the edge positions of the PC, *open circles* PC-surface lattice points, at which the ligand can access. The ligand cannot access the interior of the PC (*gray region*)

for a motion to outside the LC box or the protein interior. We estimated the average interval for the reciprocation of the ligand between the binding site and a 'Far region' ($|r_x|\geq 100$, $|r_y|\geq 100$, or $|r_z|\geq 100$) presented in Fig. 4, as follows: once the ligand reached the binding site at a step number N_b, we memorized this number and waited until the ligand reached the Far region, for which the step number is denoted as N_f. Before reaching the Far region, the ligand might revisit the binding site. However, we did not reset N_b to the revisiting step. The trajectory interval for this motion was then defined as $\Delta N_{fb}=N_f-N_b$. We then waited until the ligand visited the binding site, for which the step number is denoted again as N_b. Before accessing the binding site, the ligand might revisit the Far region. However, we did not reset N_f to the revisiting step. We then calculated the interval for this motion as $\Delta N_{bf}=N_b-N_f$. The simulation was continued, the reciprocation was observed many times, and the average for the intervals was calculated as $\langle\Delta N\rangle=(\langle\Delta N_{fb}\rangle+\langle\Delta N_{bf}\rangle)/2$, where $\langle\Delta N_{fb}\rangle$ and $\langle\Delta N_{bf}\rangle$ represent the average of ΔN_{fb}

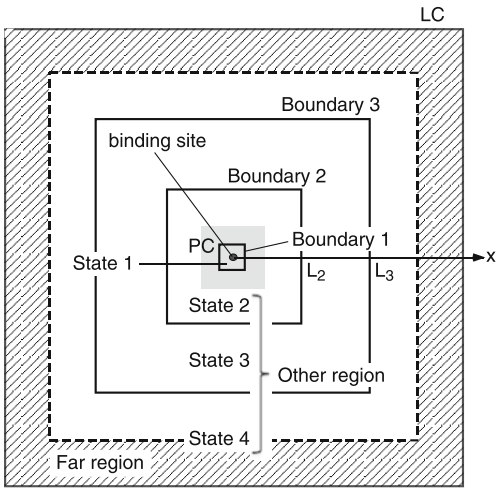

Fig. 4 Two-dimensional drawing to identify the space partitioning. Protein (*PC*) is located at the center of the LC box. *State 1* Ligand-binding cleft, with the ligand-binding site at the center of State 1, which overlaps the center of PC. In the two-state adaptive sampling, the region other than State 1 (i.e., State 2 + State 3 + State 4) is called the 'Other region'. Boundary 1 partitions States 1 and 2, Boundary 2 partitions States 2 and 3, and Boundary 3 partitions States 3 and 4. The 'Far region' is a part of State 4. Positions of Boundaries 2 and 3 are specified uniquely by L_2 and L_3, respectively, which are intercepts of the boundaries to the x-axis

and ΔN_{bf}, respectively. This simulation was performed four times, with each run executed for 5×10^{12} steps, discarding the initial 10^8 steps to compute$<\Delta N>$. We used the Mersenne twister MT19937 (Matsumoto and Nishimura 1998) to generate a random number series. The resultant value was $<\Delta N>= (2.53\pm0.02)\times 10^7$, where the ligand moved about 197,000 times between the binding site and the Far region.

The next step in the AU sampling is to enhance the probability of the ligand in the binding cleft. As such, we defined State 1 as having dimensions of $-1\le r_x\le 1$, $-1\le r_y\le 1$, and $-1\le r_z\le 1$, as portrayed in Fig. 4. The binding site is located at the center of State 1. We controlled the distribution as $P_{State1}=P_{Other}$, where P_{State1} and P_{Other} denote the probabilities of the ligand in State 1 and the 'Other region', respectively. States 2, 3, and 4 in Fig. 4 (States 2+3+4=Other region) are described in the next section. We designate this adaptive umbrella sampling as the 'two-state AU sampling' or simply 'two-state sampling'. Boundary 1 separates State 1 and the Other region (Fig. 4). The numbers (N_{State1} and N_{Other}) of ligand accessible sites are 27 and 13,997,405, respectively, for State 1 and the Other region. The transition probability of the ligand traversing Boundary 1 from State 1 to the Other region is N_{State1}/N_{Other}. This setting of the transition probability is explained later. The transition for the reversal process is accepted unconditionally. Other moves are always accepted. We repeated the simulation four times, with each run executed for 5×10^{12} steps and discarding the initial 10^8 steps. The resultant interval is$<\Delta N>= (2.81\pm0.01)\times 10^7$, where the ligand moved about 178,000 times between the binding site and the Far region. The probabilities were controlled well as $P_{State1}/P_{Other}=0.998$. This probability partitioning is in contrast to the result from the non-enhanced sampling: $P_{State1}/P_{Other}=0.193\times10^{-5}$. Consequently, the two-state AU sampling enhanced P_{State1}. In return, this sampling slowed traffic$<\Delta N>$, as shown above. For the ligand starting from the Far region, the probability of visiting State 1 is the same for each of the two simulations because all moves are unconditionally accepted in both. For the ligand starting from the binding site, all moves are also accepted unconditionally in the non-enhanced simulation. In contrast, in the two-state AU simulation, the ligand traverses Boundary 1 with the small transition probability N_{State1}/N_{Other}, which slows the traffic.

In the current protein–ligand docking model, the slow traffic does not cause a problem; i.e., the simulation is able to predict the correct complex structure once the ligand reaches the ligand-binding site. In an all-atom treatment, however, the ligand may enter the binding site with a different orientation from that in the correct complex structure or may weakly bind with non-binding sites on the protein surface. Those non-native complexes should be dissociated as quickly as possible during the enhanced sampling. Therefore, the slow traffic may cause a serious problem in conformational sampling. To increase the statistical significance, the traffic should be increased.

We survey the two-state AU sampling with the aim of more fully understanding a mechanism of the slowed-down traffic. The enhanced conformational sampling introduces the modified potential energy E_{mod} to control the probability distribution: $E_{mod}=E_u(\boldsymbol{r})$ and $E_{mc}(\boldsymbol{r})$ for the AU and multicanonical methods, respectively. The potential energy E is always zero in the present system. Therefore, the canonical ensemble assigns an equal probability to all accessible lattice points. The canonical distributions P_c(State1) and P_c(Other) are proportional, respectively, to N_{State1} and N_{Other}. The modified potential energy E_{mod} is then given as

$$\begin{cases} E_{mod}(\text{State1}) = \ln P_c(\text{State1}) = \ln N_{State1} \\ E_{mod}(\text{Other}) = \ln P_c(\text{Other}) = \ln N_{Other} \end{cases}, \tag{33}$$

where we set $RT=1$ because the temperature does not appear in this simulation. The rate constants then satisfy the following detailed balance as

$$\frac{k_{State1\to Other}}{k_{Other\to State1}} = \exp[-\Delta E_{mod}] = \frac{N_{State1}}{N_{Other}}, \tag{34}$$

where $\Delta E_{mod}=E_{mod}$(Other) – E_{mod}(State1) and the subscriptions for the rate constants represent the reaction processes. In the two-state AU sampling, $k_{State1\to Other}$ is considerably smaller than $k_{Other\to State1}$ because of $N_{State1}<< N_{Other}$. Finding the small region (State 1) for the ligand fluctuation in the Other region is arduous. To redress the balance between P_{State1}

and P_{Other}, the escape from State 1 should also be arduous, which makes $k_{\text{State1}\rightarrow\text{Other}}$ small; consequently the traffic slows.

Introducing entropy, ΔE_{mod} is rewritten as

$$\Delta E_{\text{mod}} = \ln\left[\frac{N_{\text{Other}}}{N_{\text{State1}}}\right] = S_{\text{Other}} - S_{\text{State1}}, \tag{35}$$

where S_{State1} and S_{Other} represent entropies for State 1 and the Other region, respectively: $S_{\text{State1}}=\ln N_{\text{State1}}$ and $S_{\text{Other}}=\ln N_{\text{Other}}$. Therefore, when traversing Boundary 1 from State 1 to the Other region, the ligand is expected to overcome a high-energy barrier ΔE_{mod} that originated from the entropy difference. As a general rule, in adequate enhanced sampling, the traffic slows down when the conformation traverses a boundary with a large change in entropy.

It is noteworthy that coercive traffic enhancement might result in a non-equilibrated ensemble, particularly in the all-atom treatment with explicit solvent where deep pinholes characterized by small entropies are distributed throughout the conformational space. The coercive enhancement pushes the conformation into a pinhole in the vicinity of the current conformation before the conformation takes a long trip to visit a wide energy basin.

Traffic enhancement

Is there any prescription for enhancing the traffic solely by controlling the distribution function? We now introduce States 2–3 partitioned by boundaries (Fig. 4). Boundary i is uniquely specified by six planes $x=\pm L_i$, $y=\pm L_i$, and $z=\pm L_i$, and eventually by a digit L_i. The probability of ligand in State i is denoted as $P_{\text{State}i}$ and the number of ligand accessible sites as $N_{\text{State}i}$. As we can calculate $N_{\text{State}i}$ exactly in advance, the detailed balance for transitions between States i and j for even probabilities ($P_{\text{State1}}=P_{\text{State2}}=P_{\text{State3}}=P_{\text{State4}}$) are set as

$$\frac{k_{\text{State}i\rightarrow\text{State}j}}{k_{\text{State}j\rightarrow\text{State}i}} = \frac{N_{\text{State}i}}{N_{\text{State}j}}. \tag{36}$$

We denote this AU sampling as 'four-state AU sampling' or simply 'four-state sampling'. The simulation was performed with various sets of Boundaries 2 and 3 (i.e., various values of L_2 and L_3) to investigate the dependency of the traffic on the boundary positions. We fixed State 1 (binding cleft) and the Far region as in the two-state sampling for all simulations. This simulation was performed four times at each set of boundary positions, and each run was executed for 1×10^{12} steps with the initial 10^8 steps being discarded to compute $<\Delta N>$.

Figure 5 shows the dependence of $<\Delta N>$ on the boundary set $[L_2,L_3]$. The probability was well controlled as $P_{\text{State}i}=25.00\pm0.01\%$. The traffic was enhanced considerably for all

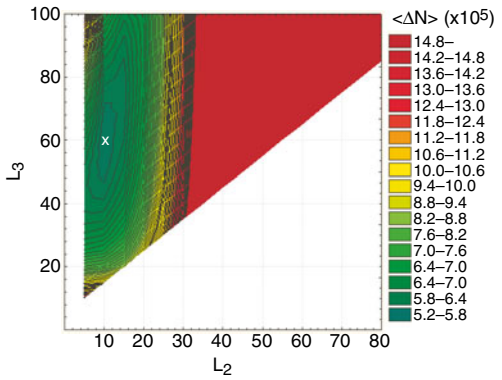

Fig. 5 Dependence of $<\Delta N>$ on the boundary set $[L_2,L_3]$, which is defined in Fig. 4. *x* Site at which the smallest $<\Delta N>$ was obtained

boundaries examined. The smallest $<\Delta N>$ (fastest traffic) was 5.21×10^5 steps at $[L_2,L_3]=[10,60]$, where the traffic was about 50-fold faster than that from two-state sampling because the introduced states (States 2–4) loosened the large entropy change: $|S_{\text{State}i} - S_{\text{State}i+1}| << |S_{\text{State1}} - S_{\text{Other}}|$. The modified potential energy of this system is funnel-like (red line in Fig. 6), contrasting to the golf hole-like potential of the two-state sampling (black broken line in Fig. 6). We also plot E_{mod} for two other sets $[L_2,L_3]=[5,15]$ (blue line) and $[40,80]$ (green line) in Fig. 6, for which $<\Delta N>$ was 9.45×10^5 and 24.1×10^5 steps, respectively. The former is more golf-hole-like than the red line, and the latter was more jar-like. Modulation of the boundaries to speed up the traffic is subtle—even for this simple sampling. We also examined simulations to produce uneven distributions ($g_1P_{\text{State1}}=g_2P_{\text{State2}}=g_3P_{\text{State3}}=g_4P_{\text{State4}}$) and found that $<\Delta N>$ depends on g_i. For instance, a set $[g_1, g_2, g_3, g_4]=[1,1,0.5,1]$ provided the fastest traffic ($<\Delta N>= 5.12\times10^5$ steps) for $[L_2,L_3]=[10,45]$.

The introduction of the intermediate states, States 2–4, corresponds to the adoption of a reaction coordinate suitable for weakening the entropy gaps. Because the present system

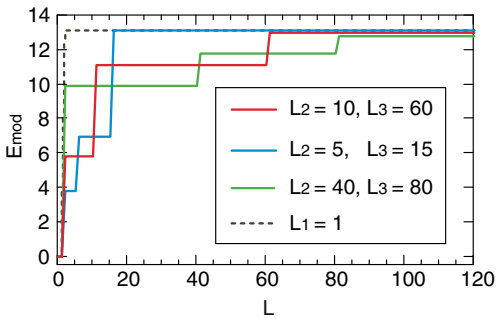

Fig. 6 Modified potential energy (E_{mod}) for four systems as a function of L, which is the distance from the ligand-binding site to a site on the x-axis (see caption of Fig. 4). *Colored lines* E_{mod} for the four-state AU sampling, for which L_2 and L_3 are shown in the *inset*, *broken line* E_{mod} for the two-state AU sampling

is extremely simple, its conformational space is also very simple. As a general rule, the choice of the reaction coordinate depends on the structure of the conformational space, and the structure remains unknown for most biological systems. It is also likely that the energy surface of a real biological system involves several low-energy basins, pinholes, and dead ends of conformational changes, which cause deep conformational trapping. Consequently, to define an effective reaction coordinate might be a difficult task for the realistic protein system. However, we generally note that the appropriate reaction coordinate drastically enhances the protein dynamics. For example, the tuning of $g(\lambda)$ (or $g(E)$ for multicanonical sampling) increases the sampling efficiency, as shown above. Then, starting with $g=1$, we can detect values of λ or E at which the traffic slows. We can then modify g at the values.

Actual procedure for multicanonical sampling

To perform enhanced conformational sampling, one should define the modified potential energy, $E_{\mathrm{mc}}(E)$ or $E_{\mathrm{u}}(\boldsymbol{r})$, which involves the canonical distribution $P_{\mathrm{c}}(E,T)$ (Eq. 30) or $P_{\mathrm{c}}(\lambda,T)$ (Eq. 20). This is self-contradictory because the canonical distribution is unknown in advance. To solve this problem, we iterate the simulation where the canonical distribution function gradually converges to the aimed function $g(E)$ or $g(\lambda)$. Below, we first explain practical procedures for multicanonical sampling and then explain those for the AU sampling.

First, we mention the energy range $[E_{\mathrm{low}}, E_{\mathrm{up}}]$ to be explored in the multicanonical simulation. From a general thermodynamic formula $1/RT=d \ln n(E)/dE$, we obtain the following relation.

$$\frac{1}{RT} = \frac{d}{dE}\left(\ln\left[\exp\left[\frac{E}{RT}\right] \times P_{\mathrm{c}}(E,T)\right]\right) = \frac{1}{RT} + \frac{\partial}{\partial E}\ln[P_{\mathrm{c}}(E,T)] \tag{37}$$

This equation yields the following.

$$\frac{\partial}{\partial E}\ln[P_{\mathrm{c}}(E,T)] = 0 \tag{38}$$

Consequently, solving Eq. 37 is equivalent to evaluating the energy value [denoted as $E_{P\mathrm{mx}}(T)$] at the maximum value of $P_{\mathrm{c}}(E,T)$. To ensure quick structural relaxation in the multicanonical simulation, the upper limit E_{up} should correspond to a high temperature T_{up}, at which the conformation overcomes high energy barriers. However, our biophysical interest is usually devoted to the conformations at room temperature (T_{room}). The lower energy limit E_{low} is therefore expected to correspond to a temperature T_{low} that is slightly lower than T_{room}. For that reason, the energy range $[E_{\mathrm{low}},E_{\mathrm{up}}]$ is determined as

$$[E_{\mathrm{low}}, E_{\mathrm{up}}] = [E_{P\mathrm{mx}}(T_{\mathrm{low}}), E_{P\mathrm{mx}}(T_{\mathrm{up}})]. \tag{39}$$

When we initiate multicanonical simulation, this energy range is unknown because $E_{P\mathrm{mx}}(T)$ is evaluated from $P_{\mathrm{c}}(E,T)$, which is unknown in advance. In the iterative procedure explained below, $P_{\mathrm{c}}(E,T)$ converges to an accurate function, following which the energy range is determined gradually.

The iterative procedure used to evaluate $P_{\mathrm{c}}(E,T)$ is as follows. A canonical simulation (denoted as 'pre-run') is first performed at T_{up} with setting $E_{\mathrm{mc}}=E$. This run produces a canonical energy distribution $P_{\mathrm{c}}^{\mathrm{pre}}(E,T_{\mathrm{up}})$, where the superscription 'pre' clarifies that the pre-run generated the distribution. Because the pre-run explores a narrow energy range around $E_{P\mathrm{mx}}(T_{\mathrm{up}})$, the distribution $P_{\mathrm{c}}^{\mathrm{pre}}(E,T_{\mathrm{up}})$ is accurate only in this range, denoted as $[E_{\mathrm{pre}}, E_{\mathrm{up}}]$ in Fig. 7a, which is narrower than the targeted range $[E_{\mathrm{low}},E_{\mathrm{up}}]$. E_{pre} can be determined quantitatively as $P_{\mathrm{c}}^{\mathrm{pre}}(E_{\mathrm{pre}},T_{\mathrm{up}}) / < P_{\mathrm{c}}^{\mathrm{pre}} \times (E,T_{\mathrm{up}}) >= d_{\mathrm{small}}$, where $< P_{\mathrm{c}}^{\mathrm{pre}}(E,T_{\mathrm{up}}) >$ is the average of $P_{\mathrm{c}}^{\mathrm{pre}}(E,T_{\mathrm{up}})$ over the range, and d_{small} is a value such as 0.1 or 0.05. Alternatively, E_{pre} may be intuitively set by viewing

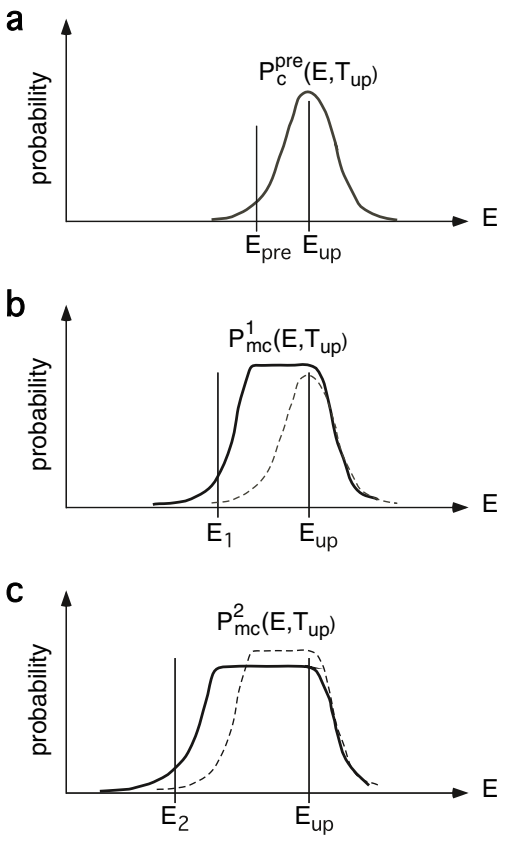

Fig. 7 The energy (E) probability distribution from the pre-run (**a**), the first multicanonical run (**b**), and the second multicanonical run (**c**). *Broken lines* in **b** and **c** correspond to the *solid lines* in **a** and **b**, respectively. ValuesE_{up}, E_1, and E_2 are described in the main text

the shape of $P_c^{pre}(E, T_{up})$. To increase the operability of $P_c^{pre} \times (E, T_{up})$, one might approximate $\ln[P_c^{pew}(E, T_{up})]$ or $P_c^{pre} \times (E, T_{up})$ by a polynomial of E or other differentiable functions. We do not reset the upper limit because E_{up} is always the upper limit for all iterative runs. The function $P_c^{pre}(E, T_{up})$ outside $[E_{pre}, E_{up}]$ is a linear function of E for which the gradient is determined by the following condition.

$$\begin{cases} \frac{dP_c^{pre}(E,T_{up})}{dE} = \frac{dP_c^{pre}(E_{pre},T_{up})}{dE_{pre}} & (\text{for } E < E_{pre}) \\ \frac{dP_c^{pre}(E,T_{up})}{dE} = \frac{dP_c^{pre}(E_{up},T_{up})}{dE_{up}} & (\text{for } E > E_{up}) \end{cases} \quad (40)$$

This equation sets walls in the energy axis so that the conformation only slightly extends outside $[E_{pre}, E_{up}]$. Below E_{pre}, the sampling is equivalent to a canonical simulation at temperature T satisfying $E_{Pmx}(T)=E_{pre}$; above E_{up}, the sampling is that at T satisfying $E_{Pmx}(T)=E_{up}$.

To expand the energy range in which the canonical distribution is determined accurately, we perform the first multicanonical run at T_{up} using the following modified potential:

$$E_{mc}^{pre}(E) = E + RT_{up} \ln\left[\frac{P_c^{pre}(E, T_{up})}{g(E)}\right]. \quad (41)$$

The initial conformation for this run is the final conformation of the pre-run. This choice of the initial conformation is important for quick relaxation of the system. This run produces an energy distribution $P_{mc}^1(E, T_{up})$ that is related formally to $P_c^{pre}(E, T_{up})$ as

$$P_{mc}^1(E, T_{up}) = n(E) \exp\left[-\frac{E_{mc}^{pre}}{RT_{up}}\right] = \frac{g(E)}{P_c^{pre}(E, T_{up})} \times n(E) \exp\left[-\frac{E}{RT_{up}}\right]. \quad (42)$$

If $P_c^{pre}(E, T_{up})$ had been determined sufficiently accurately in $[E_{pre}, E_{up}]$ and if the first multicanonical run is sufficiently long, then $P_{mc}^1(E, T_{up})$ converges to $g(E)$ in this energy range. Figure 7b portrays a flat distribution for $P_{mc}^1 \times (E, T_{up})$ assuming that $g(E)=1$. In practice, however, $P_c^{pre} \times (E, T_{up})$ might not be sufficiently accurate, where $P_c^{pre} \times (E, T_{up})$ deviates appreciably from $g(E)$. For the second multicanonical run, we define the canonical energy distribution $P_c^1(E, T_{up})$ using Eq. 42 as

$$P_c^1(E, T_{up}) = n(E) \exp\left[-\frac{E}{RT_{up}}\right] = \frac{P_{mc}^1(E, T_{up}) P_c^{pre}(E, T_{up})}{g(E)}. \quad (43)$$

Regarding that equation, $P_c^1(E, T_{up})$ is uniquely determined because $P_c^{pre}(E, T_{up})$ and $P_{mc}^1(E, T_{up})$ are computed numerically from the pre-run and the first multicanonical run, respectively, and $g(E)$ is given definitely by users. The distribution $P_{mc}^1(E, T_{up})$ decreases outside the range $[E_{pre}, E_{up}]$ (Fig. 7b) because of the energy walls (Eq. 40). The sampling range can then be expanded to $[E_1, E_{up}]$, where E_1 might be set as $P_{mc}^1(E_1, T_{up}) / < P_{mc}^1(E_{pre}, T_{up}) >= d_{small}$, where $< P_{mc}^1(E_{pre}, T_{up}) >$ is the average of $P_{mc}^1(E, T_{up})$ over the range $[E_{pre}, E_{up}]$. Equation 43 defines $P_c^1(E, T_{up})$ only in this energy range, and its outside range is determined as shown below.

$$\begin{cases} \frac{dP_c^1(E,T_{up})}{dE} = \frac{dP_c^1(E_1,T_{up})}{dE_1} & (for\ E < E_1) \\ \frac{dP_c^1(E,T_{up})}{dE} = \frac{dP_c^1(E_{up},T_{up})}{dE_{up}} & (for\ E > E_{up}) \end{cases} \quad (44)$$

Next, we define the modified potential energy as

$$E_{mc}^1(E) = E + RT_{up} \ln\left[\frac{P_c^1(E, T_{up})}{g(E)}\right]. \quad (45)$$

The second multicanonical run using E_{mc}^1 produces numerically the distribution function $P_{mc}^2(E, T_{up})$ (Fig. 7c). This procedure is repeated until the energy range reaches $[E_{low}, E_{up}]$, at which point the energy distribution converges to $g(E)$.

Generally the i-th multicanonical run produces $P_{mc}^i \times (E, T_{up})$ numerically, and the canonical distribution $P_c^i \times (E, T_{up})$ is computed as

$$P_c^i(E, T_{up}) = \frac{P_{mc}^i(E, T_{up}) P_c^{i-1}(E, T_{up})}{g(E)}. \quad (46)$$

Then, the modified potential energy for the $(i+1)$-th multicanonical run is defined as

$$E_{mc}^i(E) = E + RT_{up} \ln\left[\frac{P_c^i(E, T_{up})}{g(E)}\right]. \quad (47)$$

In the McMD simulation, the derivatives of $\ln[P_c^i(E, T_{up})]$ or $P_c^i(E, T_{up})$ should be computed (Eq. 32). Similar to the process used for $P_c^{pre}(E, T_{up})$, one might approximate $\ln[P_c^i \times (E, T_{up})]$ or $P_c^i(E, T_{up})$ by a polynomial of E or other differentiable functions. The derivatives are then computed analytically.

Actual procedure for AU sampling

In multicanonical sampling, all microscopic states of the same energy E contribute evenly to $P_c(E,T)$. For this reason, the density of states $n(E)$ appears in the formulae (Eqs. 26 and 27). Although AU sampling has some similarity to multicanonical sampling, $n(E)$ does not appear in the former formulation because microscopic states of the same λ, which contribute to $P_c(\lambda,T)$, have various energies. Therefore, the procedures for the AU sampling are somewhat different from those for multicanonical sampling.

We denote the sampled range for λ as $[\lambda_{low}, \lambda_{up}]$, where the canonical distribution $P_c(\lambda,T)$ should be determined

accurately. If the structures at λ_{low} and λ_{up} belong to different stable states (different energy basins) at T, then the sampling might provide possible pathways for the conformational changes between the states. To start with, a canonical run (again demoted as 'pre-run') is performed using the original potential energy at T, which is usually an interesting temperature such as room temperature. In this run, we restrict the sampling in a range $[\lambda_{low}^{pre}, \lambda_{up}^{pre}]$, which is usually narrower than $[\lambda_{low}, \lambda_{up}]$, by setting artificial walls outside the range, and the initial simulation conformation is better in this range. The pre-run produces a canonical distribution $P_c^{pre}(\lambda, T)$, which is accurate only in $[\lambda_{low}^{pre}, \lambda_{up}^{pre}]$. Then we extrapolate $P_c^{pre}(\lambda, T)$ to a wider range $[\lambda_{low}^{1}, \lambda_{up}^{1}]$, where $\lambda_{low}^{1} \le \lambda_{low}^{pre}$ and $\lambda_{up}^{1} \ge \lambda_{up}^{pre}$. Subsequently, we set the modified potential energy as

$$E_u^{pre}(r) = E(r) + RT \ln\left[\frac{P_c^{pre}(\lambda, T)}{g(\lambda)}\right], \tag{48}$$

and perform the first AU run at T. The numerically obtained distribution function $P_u^1(\lambda, T)$ is related to $P_c^{pre}(\lambda, T)$ as

$$\begin{aligned} P_u^1(\lambda, T) &= \int D(f(r) - \lambda) \exp\left[-\frac{E_u^{pre}(r)}{RT}\right] dr \\ &= \frac{g(\lambda)}{P_c^{pre}(\lambda, T)} \int_{a(r)=\lambda} \exp\left[-\frac{E}{RT}\right] dr \\ &= \frac{g(\lambda) P_c^1(\lambda, T)}{P_c^{pre}(\lambda, T)}, \end{aligned} \tag{49}$$

where the normalization constant (A_u in Eq. 21) is omitted. The 1D canonical distribution $P_c^1(\lambda, T)$ is then determined as

$$P_c^1(\lambda, T) = \frac{P_u^1(\lambda, T) P_c^{pre}(\lambda, T)}{g(\lambda)}. \tag{50}$$

We now expand again the range for $P_c^1(\lambda, T)$ to $[\lambda_{low}^2, \lambda_{up}^2]$, reset the walls, and define the modified potential energy as

$$E_u^1(r) = E(r) + RT \ln\left[\frac{P_c^1(\lambda, T)}{g(\lambda)}\right]. \tag{51}$$

The second AU run is then performed at T. The procedure is repeated until the sampling covers the intended range $[\lambda_{low}, \lambda_{up}]$. The expansion of the sampling range and the simulation length should be determined carefully with progression of the iteration.

Methods to update the canonical distribution

In the methods described above, the modified potential energy is invariant during an iterative run, and the canonical distribution function $P_c(q,T)$, in which $q=E$ or λ, is updated after completion of the iterative run. The i-th run should be performed sufficiently long to generate $P_c^i(q, T)$ as accurately as possible in a given range $[q_{low}^i, q_{up}^i]$. Consequently, the simulation is categorized in equilibrium sampling when the initial relevant simulation conformation is prepared. We have designated this updating method as 'the every-run update method'.

An alternative means is to update P_c slightly at every step of the simulation. When the conformation is detected in a bin $[q, q+\Delta q]$, P_c is modified by a small increment Δq as

$$P_c(q, T) \rightarrow P_c(q, T) + \Delta P_c(q). \tag{52}$$

This method is called the Wang–Landau sampling for $q=E$ (Wang and Landau 2001) and the metadynamics (Laio and Parrinello 2002) or the filling potential sampling method (Fukunishi et al. 2003) for $q=\lambda$. The increment ΔP_c is usually positive and restricted in the detected bin or bins in the vicinity of the detected bin. When the sampling is based on MD, ΔP_c should be differentiable with respect to q. The modified potential energy is modified at every simulation step. Therefore, this simulation is categorized in non-equilibrium sampling independently of the initial simulation conformation. During the simulation, the conformation feels a repulsive force to escape from bins that have been visited. With progress of the simulation, increment ΔP_c decreases gradually, ultimately vanishing: $\Delta P_c \rightarrow 0$. One expects convergence of P_c to the accurate distribution function at this final stage. We designate this updating method as 'the every-step update method'.

The benefit of the every-step update is its ease of automation: Once the protocol for setting ΔP_c is determined, one can perform the simulation without manual operations until ΔP_c vanishes. However, the conformational space of a large biological system is vast, within which numerous energy basins, pinholes, and energy barriers can be distributed. In this case, the emerging repulsive force might push the conformations within a local region of the conformational space before the conformation fluctuates toward vast regions that have not yet been visited. In other words, the conformation wanders among a small number of basins/pinholes without overcoming energy barriers to visit the new regions.

To avoid this delicate problem, a force-biased multicanonical sampling (Kim et al. 2004) has been proposed. In this method, the modified potential energy is maintained for a long interval of the simulation, during which ΔP_c is summed up ($\Delta P_{sum} = \sum_i \Delta P_c^i$, where i specifies the simulation step) but not added to P_c at every time step. Then, after executing the interval, ΔP_{sum} is added to P_c. This method is categorized in the equilibrated sampling in each interval. We designate this updating method as 'the every-interval update method'. If the interval length is sufficiently

long, then the method is substantially equivalent to the every-run update. However, if the interval is short, this method reaches the every-step update. The benefit of the every-interval update is its ease of automation, where the setting of the interval length controls the entire sampling process.

All of these methods target the accurate estimation of the canonical distribution $P_c(q,T)$. Consequently, a long final run (production or sampling run) is required while using the converged canonical distribution without another update. This additional procedure is important for checking whether the converged distribution can produce the aimed distribution $g(q)$ [usually $g(q)=1$)] Coincidentally, the sampled conformations from the production run are used for analyses.

A conventional MD (canonical MD) at a temperature provides a canonical energy distribution, $P_c(E,T)$, which is accurate only in a narrow energy range, from which a partially accurate density of states $n(E)$ is obtained. Terada et al. (2003) performed several canonical MD runs at different temperatures, obtained the fractions of $n(E)$, and constructed the entire density of states by integrating the fractions. When the computed system is small, the constructed $n(E)$ is useful for the production run of multicanonical simulation without the iterative procedure. With increasing system size, however, the accuracy of the constructed $n(E)$ decreases because a canonical run at a temperature sample only involves a restricted region of the conformational space. However, this method provides the first approximation of $n(E)$, which can be refined via iterative multicanonical runs.

TTP-multicanonical sampling

The methods described above guarantee that the accuracy of $P_c(q,T)$ increases concomitantly with increased simulation length. However, the volume of the conformational space increases rapidly with increased system size (Eq. 1), while the moving speed of the conformation in the conformational space remains almost unchanged despite the system size. Consequently, equilibration becomes unachievable in an actual computational time with increased system size.

Trajectory-parallelization methods have recently been developed for use in the multicanonical simulation of a large system in which many runs are performed, starting from various initial conformations (Higo et al. 2009; Sugihara et al. 2009). In the trivial trajectory-parallelization multicanonical molecular dynamics (TTP-McMD), the multiple trajectories are simply connected, where each trajectory might be short. Importantly, the integrated long trajectory can be regarded theoretically as a single simulation trajectory because the detailed balance is satisfied at the inter-trajectory connection points (Ikebe et al. 2011a). Because the initial conformations spread in the conformational space in advance, the sampled space is wider than that by the single multicanonical simulation, even though the length of the integrated trajectory is equal to or shorter than the single simulation trajectory (Fig. 8). To substantialize the wide sampling, trajectory parallelization is done from the pre-run stage, where the conformations are randomized in the high-energy region. The next multiple runs (first multicanonical runs) are then initiated from the last snapshots of the pre-runs, and so on. This method has been used for the coupled folding and binding of an intrinsically disordered protein (Higo et al. 2011).

Parallel computing to speed up a single run by a number of computing nodes is effective when the time development of the system is of interest. In multicanonical sampling (and in any of the generalized ensemble methods) the simulation trajectory does not provide realistic time development of the system. The purpose of multicanonical sampling is to obtain the conformational ensemble. To increase the statistics of the ensemble, N runs should be executed when there are N computing nodes. In fact, the computing nodes do not communicate during the simulation. In other words, the parallelization efficiency is always 100% in the TTP method.

Other computational techniques

The enhanced sampling methods explained in this review are those that control the sampling by a 1D distribution $P_c(q,T)$. This can be extended naturally to a multi-dimensional version where $P_c(q_1, q_2, \ldots; T)$ controls the sampling. Some 2D versions have already been proposed (Higo et al. 1997; Iba et al. 1998; Nakajima 1998; Okumura and Okamoto 2004),

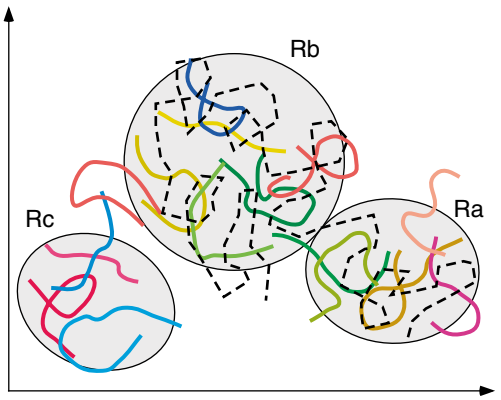

Fig. 8 Scheme for trivial trajectory-parallelization (TTP) multicanonical sampling. *Differently colored lines* Different trajectories. Conformational space is represented two-dimensionally. The multiple trajectories are distributed in three *gray regions* (R_a, R_b, R_c). *Broken line* Long single-simulation trajectory that does not visit R_c

such as multi-dimensional AU sampling (Bartels and Karplus 1997), multi-dimensional replica exchange (Sugita et al. 2000), and multi-dimensional AU/multicanonical sampling (Zheng et al. 2008). If the sampling is performed for a sufficiently long period to determine $P_c(q_1, q_2, T)$ accurately, then the generated conformational ensemble provides more information than the 1D distribution.

We now introduce two computational techniques: the mass-scaling and puddle-skimming methods. Although these methods are not categorized in the generalized ensemble method, they can be combined to the AU or multicanonical sampling. In the mass-scaling method, atomic masses are varied to speed up the sampling. Feenstra et al. (1999) scaled up the mass of hydrogen atoms in a system and increased the time step Δt to integrate the Newtonian equations because fast motions related to the hydrogen atoms are slowed by mass scaling. In contrast, Gee and van Gunsteren (2006) scaled down the masses of the solvent atoms, with the result that the viscosity decreased and the peptide moved quickly. One might point out that the system kinetics changes through mass scaling, suggesting that this method is less useful for tracing the time series of the system motions. However, the equilibrated distribution converges to the canonical ensemble (i.e., after a long simulation) irrespective of the unrealistic kinetics. Mass-scaling can help the generalized ensemble method to speed up the sampling.

A protein is a long polypeptide chain in which the atoms are connected by covalent bonds. Therefore, once the chain has misfolded during a simulation, the structure should unfold to restart the folding. In the puddle-skimming method, energy that is higher than a given value E_b is reset to E_b (Steiner et al. 1998; Rahman and Tully 2002a, b). When E_b is set to a high value, conformations with energies larger than E_b do not influence the equilibrated ensemble at room temperature. This method might allow self-overlapping of the polypeptide chain, i.e., misfolded structure refolds without unfolding. A simplified protein model has shown that the self-overlapping considerably enhances the structure relaxation when the overlap is controlled well (Iba et al. 1998).

Free-energy landscape

The enhanced conformational sampling is used for constructing the free-energy landscape. The free-energy landscape visualizes conformational clusters (low free-energy basins) and inter-cluster pathways. The free energy assigned to cluster i is defined as $F_i = -RT_a \ln[N_i]$, where N_i is the number of conformations involved in the cluster and T_a is the temperature at which the conformational ensemble is obtained (detail is described later). Therefore, the largest cluster (i.e., the cluster involving the most conformations) has the lowest free energy, and the free-energy difference between clusters i and j is calculated as $\Delta F = F_j - F_i = -RT_a \ln[N_j / N_i]$. When a conformational distribution $P(q_1,q_2,\ldots)$ is computed from the ensemble, where the set of parameters $[q_1,q_2,\ldots]$ specifies a position in the conformational space, the free-energy map is defined as $F(q_1,q_2,\ldots) = -RT_a \ln[P(q_1,q_2,\ldots)]$. In the map, a cluster corresponds to a low free-energy region, and free-energy barriers are identified among the low free-energy regions.

In multicanonical sampling, the entire conformational ensemble, denoted as Q_{all}, is characterized by a wide energy distribution. A canonical conformational ensemble $Q_c(T_a)$ at temperature T_a is generated as follows: first, we pick a conformation from Q_{all}, for which energy is denoted as E_{pic}, and assign a probability $P_c(E_{pic},T_a)$ to the selected conformation as

$$p_c(E_{pic}, T_a) = P_c(E_{pic}, T_a)/P_c^{max}(T_a), \quad (53)$$

where $P_c^{max}(T_a)$ is the maximum value of $P_c(E,T_a)$. If $p_c(E_{pic}, T_a)$ is larger than a random number distributed uniformly in [0,1], then the chosen conformation is registered in $Q_c(T_a)$. Repeating this procedure for all conformations in Q_{all}, the ensemble $Q_c(T_a)$ is generated. The most biophysically interesting ensemble is usually that at room temperature: $Q_c(T_{room})$. We can generate a visible free-energy landscape by projecting the structures in $Q_c(T_{room})$ onto a low-dimensional conformational space. The low-dimensional space might be constructed by overall structural identifiers, such as the radius of gyration, solvent accessible surface area, or root mean square deviation measured from a given structure, or by abstract coordinate axes derived from principal component analysis (PCA). Ono et al. (1999) constructed a fine free-energy landscape for the *cis*/*trans*-imide isomerization of a peptide dimer, –Ala–Pro–.

The TTP-McMD produces short trajectories, and a long trajectory is generated connecting the short trajectories. Since the long trajectory can be regarded as a single multicanonical trajectory, the snapshots recorded in the long trajectory construct the entire conformational ensemble Q_{all}. The distribution $P_c(E,T_a)$ is also computed from the long trajectory and then the ensemble $Q_c(T_a)$ is computed with the method explained above.

We note a disadvantage of the overall structural identifiers to generate the free-energy landscape: widely different protein tertiary structures can have the same value as the structural identifier. This structural ambiguity leads to a misleading interpretation of a free-energy barrier. We experienced that free-energy barriers identified in the PCA space completely vanish in the space constructed by the overall structural identifiers (Kamiya et al. 2002; Higo et al. 2011).

All-atom McMD simulations of various systems

Lastly in this paper, we describe our all-atom McMD studies of various biophysical systems. In these studies, we gradually increased the system size to be sampled and determined that at the present time the McMD method is applicable to the 57-residue system. We first applied McMD to a two-residue peptide and produced a free-energy landscape in which possible conformations were identified as clusters (Nakajima et al. 2000). The clusters were separated by free-energy barriers and might be bridged by free-energy pathways. This work revealed that McMD is useful to study biological systems.A seven-residue peptide (DNA-binding segment of a DNA binding protein) was subsequently solved (Higo et al. 2001b). Although this segment adopts a helix in the protein framework, it is disordered in the isolated state. We have shown that the free-energy landscape consists of various secondary structures, such as helices, hairpins, and other disordered conformations. It is particularly interesting that a cluster was found whose structure is the same as that in the protein framework. A similar result was obtained in McMD simulations of a nine-residue segment taken from a distal β-hairpin of a SH3 domain (Ikeda et al. 2003). These results suggest that the segment structure in the protein framework is metastable, even in the disordered state. The McMD simulations of a seven-residue β segment (Higo et al. 2001a; Kamiya et al. 2002) revealed that three β-hairpin clusters exist in the free-energy landscape and that each cluster is characterized by a different number of inter-strand hydrogen bonds. Therefore, hydrogen bond formation accompanies a jump in a free-energy barrier. A similar result was obtained in the work described above (Higo et al. 2001b). The McMD simulation of a 16-residue chameleon sequence (a part of DNA binding protein) showed that this sequence has a strong propensity to fold into α-helix or β-hairpin, each of which correlated well with the experimentally determined polytypic structures (Ikeda and Higo 2003). The free-energy landscape visualized probable pathways for conformational changes between the α and β structures, suggesting that the actually selected structure (α or β) is determined by an interaction between the DNA-binding protein and DNA.

We then proceeded to longer peptides, which might exist as a single chain state without a protein framework. The McMD simulation of a 25-residue segment from the Alzheimer's β amyloid peptide (Aβ) in a TFE/water co-solvent showed that this peptide folds into the experimentally determined helical structure (Kamiya et al. 2007), although it is disordered in water (Ikebe et al. 2007a). The free-energy landscape was funnel-like above 325 K, where the funnel bottom corresponded to the experimental structure, and the landscape transitioned abruptly to a rugged one below 325 K. This work might have captured a general property of the temperature-induced structural transition exhibited by many peptides/proteins. The effect of solvent on the polypeptide conformation is an interesting issue in biophysics. A 24-residue peptide, humanin, is disordered in water and adopts a helical structure in the TFE/water co-solvent. We performed McMD simulations of this peptide in both solvents (Yagisawa et al. 2008). The results obtained showed good agreement with the experiment in which we discussed details of the interactions among the peptide, TFE, and water. McMD simulation of a 40-residue protein, the C-terminal domain of H-NS, in explicit water has also ben performed (Ikebe et al. 2007b). This small protein consists of α and β secondary-structure elements in the native structure. The obtained conformational ensemble involved a small cluster, which corresponded to the native structure, and a large cluster, where half of the protein (helical region) folded well to the native structure but the other half (β region) adopted a distorted β-hairpin. Analyses showed that the two regions were incorrectly packed together. The analyses also suggested that the force field might not be sufficiently accurate. Nevertheless, the existence of the small native-like cluster was encouraging because the McMD approach proved to be powerful for the protein. It is likely that the small cluster grows as the largest cluster if an accurate force field is used.

Based on those studies, we proceeded to a 57-residue protein, the first repeat of human glutamyl-prolyl-tRNA synthetase (EPRS-R1), surrounded by an explicit solvent (Ikebe et al. 2011b). This protein comprises two long helices adopting a helix–hairpin fold in its native NMR structure (Jeong et al. 2000). The force field was set carefully so that it prefers either a α or β secondary structure depending on the sequence (Kamiya et al. 2005), although EPRS-R1 is the helical protein. Starting from a fully extended conformation, the McMD simulation produced conformational ensembles at several temperatures. The protein was disordered at a high temperature (600 K for instance). In contrast, the ensemble at 300 K was characterized by two helical regions, which corresponded to those observed in the NMR structure. This room temperature ensemble was subjected to a structure clustering, resulting in 20 clusters. Importantly, the largest cluster (lowest free-energy cluster) showed the most native-like structure of all clusters. Subsequent analyses revealed that the hydrophobic core formation between the two helical regions drives the conformation toward the native fold with exclusion of water molecules from the protein interior.

The McMD simulation was used to study protein–ligand flexible docking. The first application was done on the binding of a short proline-rich peptide to a Src homology 3 (SH3) domain (Nakajima et al. 1997a). Although the protein and ligand were put into a vacuum, a conformational cluster corresponded to the native complex. In the flexible docking of lysozyme and its inhibitor in explicit solvent, a

number of different clusters were obtained (Kamiya et al. 2008). Importantly, the largest cluster (lowest free-energy cluster) corresponded to the native complex form, and the native cluster was discriminated from the other minor clusters by a free-energy barrier.

The McMD simulation was applied on an IDP system consisting of a 15-residue IDP segment (NRSF/REST) and its receptor protein mSin3 (Higo et al. 2011). Its native complex structure was resolved through an nuclear magnetic resonance experiment in which NRSF/REST adopts a helix when it binds to the deep binding cleft of mSin3 (Nomura et al. 2005). Starting from a conformation where NRSF/REST was disordered and apart from the receptor in explicit solvent, an ensemble at 300 K was obtained. The free-energy landscape revealed that NRSF/REST can bind to mSin3, adopting various conformations, with cluster analysis showing that the largest is the native-complex cluster. The other minor clusters are non-native ones. In the non-native clusters, NRSF/REST adopts bent or extended structures in the binding cleft of mSin3, with some of these adopting the opposite orientation against NRSF/REST in the native complex. The free-energy landscape exhibited two free-energy barriers. Analyses have shown that NRSF/REST changes the chain orientation or the end-to-end distance to overcome free-energy barriers. Additional McMD simulations of single-chain NRSF/REST have revealed that NRSF/REST is disordered in solution and that the various conformations in the complex state also appear in this free state. Therefore, NRSF/REST is highly flexible in both the complex and free states. We have proposed a mechanism for this system in which the coupled folding and binding is achieved through coupling of the population shift (Bosshard 2001; Yamane et al. 2010) and induced folding (Monod et al. 1965; Spolar and Record 1994).

Conclusion

With the rapidly increasing capabilities of computers, the study of the conformational sampling of large biological systems is becoming important. In this context, the enhancement of sampling is of crucial importance in the exploration of the energy surface with statistical significance,. In this article, we have explained the methodology of multicanonical and AU sampling methods, which are categorized in generalized ensemble methods. These methods directly control the probability distribution and indirectly control the transition probability (rate constant) among different states. Studies of various biophysical systems, expressed as all-atom models, were reported here. The results show that enhanced sampling might slow the large motions of the system, even when the enhancement is performed fairly, because the entropy largely varies at a position of the reaction coordinate. We have demonstrated that the loosening of the large entropy change drastically enhances the sampling.

Acknowledgments H.N. was supported by Grants-in-Aid for Scientific Research (B) (20370061) from the Japan Society for the Promotion of Science, and for Scientific Research on Priority Areas 'Structures of Biological Macromolecular Assemblies' (513–20051013) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan. J.H. was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (21113006) from MEXT. N.K. was supported by a Grant-in-Aid for Scientific Research (22570160) from MEXT. J.H., N.K., and H.N. were supported by grants from the New Energy and Industrial Technology Development Organization (NEDO) Japan.

Conflict of interest None

References

Bartels C, Karplus M (1997) Multidimensional adaptive umbrella sampling: applications to main chain and side chain peptide conformations. J Comput Chem 18:1450–1462

Berg BA, Neuhaus T (1992) Multicanonical ensemble: a new approach to simulate first-order phase transitions. Phys Rev Lett 68:9–12

Bosshard HR (2001) Molecular recognition by induced fit: how fit is the concept? News Physiol Sci 16:171–173

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 21:167–195

Dill K (1985) Theory for the folding and stability of globular proteins. Biochemistry 24:1501–1509

Evans DJ, Morriss GP (1983) The isothermal/isobaric molecular dynamics ensemble. Phys Lett A 98:433–436

Feenstra KA, Hess B, Berendsen HJC (1999) Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. J Comput Chem 20:786–798

Fukunishi H, 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

Fukunishi Y, Mikami Y, Nakamura H (2003) The filling potential method: a method for estimating the free energy surface for protein–ligand docking. J Phys Chem B 107:13201–13210

Gee PJ, van Gunsteren WF (2006) Numerical simulation of the effect of solvent viscosity on the motions of a β-peptide heptamer. Chem-A Eur J 12:72–75

Go N (1983) Theoretical studies of protein folding. Annu Rev Biophys Bioeng 12:183–210

Hansmann UHE, Okamoto Y (1993) Prediction of peptide conformation by multicanonical algorithm: new approach to the multiple-minima problem. J Comp Chem 14:1333–1338

Hansmann UHE, Okamoto Y, Eisenmenger F (1996) Molecular dynamics, Langevin, and hybrid Monte Carlo simulations in multicanonical ensemble. Chem Phys Lett 259:321–330

Higo J, Nakajima N, Shirai H, Kiedra A, Nakamura H (1997) Two-component multicanonical Monte Carlo method for effective conformational sampling. J Comp Chem 18:2086–2092

Higo J, Galzitskaya OV, Ono S, Nakamura H (2001a) Energy landscape of a beta-hairpin peptide in explicit water studied by multicanonical molecular dynamics. Chem Phys Lett 377:169–175

Higo J, Ito N, Kuroda M, Ono S, Nakajima N, Nakamura H (2001b) Energy landscape of a peptide consisting of alpha-helix, 3(10)-helix, beta-turn, beta-hairpin, and other disordered conformations. Protein Sci 10:1160–1171

Higo J, Kamiya N, Sugihara T, Yonezawa Y, Nakamura H (2009) Verifying trivial parallelization of multicanonical molecular dynamics for conformational sampling of a polypeptide in explicit water. Chem Phys Lett 473:326–329

Higo J, Nishimura Y, Nakamura H (2011) A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. J Am Chem Soc 133:10448–10458

Hoover WG (1985) Canonical dynamics: equilibrium phase-space distributions. Phys Rev A 31:1695–1697

Hukushima K, Nemoto K (1996) Exchange Monte Carlo method and application to spin glass simulations. J Phys Soc Jpn 65:1604–1608

Iba Y, Chikenji G, Kikuchi M (1998) Simulation of lattice polymers with multi-self-overlap ensemble. J Phys Soc Jpn 67:3327–3330

Ikebe J, Kamiya N, Ito J, Shindo S, Higo J (2007a) Simulation study on the disordered state of an Alzheimer's β amyloid peptide Aβ(12–36) in water consisting of random-structural, β-structural, and helical clusters. Protein Sci 16:1596–1608

Ikebe J, Kamiya N, Shindo H, Nakamura H, Higo J (2007b) Conformational sapling of a 40-residue protein consisting of α and β secondary-structure elements in explicit solvent. Chem Phys Lett 443:364–368

Ikebe J, Umezawa K, Kamiya N, Sugihara T, Yonezawa T, Takano Y, Nakamura H, Higo J (2011a) Theory for trivial trajectory parallelization of multicanonical molecular dynamics and application to a polypeptide in water. J Comput Chem 32:1286–1297

Ikebe J, Standley DM, Nakamura H, Higo J (2011b) Ab initio simulation of a 57-residue protein in explicit solvent reproduces the native conformation in the lowest free-energy cluster. Protein Sci 20:187–196

Ikeda K, Higo J (2003) Free-energy landscape of a chameleon sequence in explicit water and its inherent α/β bifacial property. Protein Sci 12:2542–2548

Ikeda K, Galzitskaya OV, Nakamura H, Higo J (2003) Beta-hairpins, alpha-helices, and the intermediates among the secondary structures in the energy landscape of a peptide from a distal beta-hairpin of SH3 domain. J Comput Chem 24:310–318

Jeong EJ, Hwang GS, Kim KH, Kim MJ, Kim S, Kim KS (2000) Structural analysis of multifunctional peptide motifs in human bifunctional tRNA synthetase: identification of RNA-binding residues and functional implications for tandem repeats. Biochemistry 39:15775–15782

Kamiya N, Higo J, Nakamura H (2002) Conformational transition states of beta-hairpin peptide between the ordered and disordered conformations in explicit water. Protein Sci 11:2297–2307

Kamiya N, Watanabe YS, Ono S, Higo J (2005) AMBER-based hybrid force field for conformational sampling of polypeptides. Chem Phys Lett 401:312–317

Kamiya N, Mitomo D, Shea J-E, Higo J (2007) Folding of the 25 residue Aβ(12–36) peptide in TFE/water: temperature dependent transition from a funneled free-energy landscape to a rugged one. J Phys Chem B 111:5351–5356

Kamiya N, Yonezawa Y, Nakamura H, Higo J (2008) Protein-inhibitor flexible docking by a multicanonical sampling: native complex structure with the lowest free energy and a free-energy barrier distinguishing the native complex from the others. Proteins 70:41–53

Kar P, Nadler W, Hansmann UHE (2009) Microcanonical replica exchange molecular dynamics simulation of proteins. Phys Rev E 80:056703

Kidera A (1995) Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide. Proc Natl Acad Sci USA 92:9886–9889

Kim JG, Fukunishi Y, Kidera A, Nakamura H (2004) Multicanonical molecular dynamics algorithm employing an adaptive force-biased iteration scheme. Phys Rev E 70:057103

Kim J, Keyes T, Straub JE (2010) Generalized replica exchange method. J Chem Phys 132:224107

Koga N, Takada S (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. J Mol Biol 313:171–180

Laio A, Parrinello M (2002) Escaping free-energy minima. Proc Natl Acad Sci USA 99:12562–12566

Maragakis P, Lindorff-Larsen K, Eastwood MP, Dror RO, Klepeis JL, Arkin IT, Jensen MØ, Xu H, Trbovic N, Friesner RA, Iii AG, Shaw DE (2008) Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. J Phys Chem B 112:6155–6158

Matsumoto M, Nishimura T (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Trans Model Computer Simulations 8:3–30

Mezei M (1987) Adaptive umbrella sampling: self-consistent determination of the non-Boltzmann bias. J Comput Phys 68:237–248

Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers Peptide Sci 60:96–123

Miyazawa S, Jernigan RL (1985) Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552

Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: a plausible model. J Mol Biol 12:88–118

Munoz V, Eaton WA (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 96:11311–11316

Nakajima N (1998) A selectively enhanced multicanonical molecular dynamics method for conformational sampling of peptides in realistic water molecules. Chem Phys Lett 288:319–326

Nakajima N, Higo J, Kiedra A, Nakamura H (1997a) Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation. Chem Phys Lett 278:297–301

Nakajima N, Nakamura H, Kidera A (1997b) Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. J Phys Chem 101:817–824

Nakajima N, Higo J, Kiedra A, Nakamura H (2000) Free energy landscape of peptides by enhanced conformational sampling. J Mol Biol 296:197–216

Narumi T, Ohno Y, Okimoto N, Koishi T, Suenaga A, Futatsugi N, Yanai R, Himeno R, Fujikawa S, Ikei M, Taiji M (2006) A 55 TFLOPS simulation of amyloid-forming peptides from yeast prion Sup35 with the special-purpose computer system MDGRAPE-3. In: Proc Supercomputing 2006 (SC06). Tampa, Florida

Nomura M, Uda-Tochio H, Murai K, Mori N, Nishimura Y (2005) The neural repressor NRSF/REST binds the PAH1 domain of the Sin3 corepressor by using its distinct short hydrophobic helix. J Mol Biol 354:903–915

Nose S (1984) A unified formulation of the constant temperature molecular-dynamics methods. J Chem Phys 81:511–519

Okumura H, Okamoto Y (2004) Molecular dynamics simulations in the multibaric–multithermal ensemble. Chem Phys Lett 391:248–253

Ono S, Nakajima N, Higo J, Nakamura H (1999) The multicanonical weighted histogram analysis method for the free-energy landscape along structural transition paths. Chem Phys Lett 312:247–254

Paine GH, Scheraga HA (1985) Prediction of the native conformation of a polypeptide by a statistical–mechanical procedure. I. Backbone structure of enkephalin. Biopolymers 24:1391–1436

Rahman JA, Tully JC (2002a) Puddle-skimming: an efficient sampling of multidimensional configuration space. J Chem Phys 116:8750–8760

Rahman JA, Tully JC (2002b) Puddle-jumping: a flexible sampling algorithm for rare event systems. Chem Phys 285:277–287

Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, Eastwood MP, Gagliardo J, Grossman JP, Ho CR, Ierardi DJ, Kolossváry I, Klepeis JL, Layman T, McLeavey C, Moraes MA, Mueller R, Priest EC, Shan Y, Spengler J, Theobald M, Towles B, Wang SC (2007) Anton: a special-purpose machine for molecular dynamics simulation. In: Proc. 34th Int Symp Computer Architecture (ISCA'07). San Diego, CA, pp 1–12

Spolar RS, Record MT Jr (1994) Coupling of local folding to site-specific binding of proteins to DNA. Science 263:777–784

Steiner MM, Genilloud PA, Wilkins JW (1998) Simple bias potential for boosting molecular dynamics with the hyperdynamics scheme. Phys Rev B 57:10236–10239

Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447:1021–1025

Sugihara T, Higo J, Nakamura H (2009) Parallelization of Markov chain generation and its application to the multicanonical method. J Phys Soc Jpn 78:074003

Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151

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

Sugita Y, Kitao A, Okamoto Y (2000) Multidimensional replica-exchange method for free-energy calculations. J Chem Phys 113:6042–6051

Terada T, Matsuo Y, Kidera A (2003) A method for evaluating multicanonical potential function without iterative refinement: application to conformational sampling of a globular protein in water. J Chem Phys 118:4306–4311

Trebst S, Troyer M, Hansmann UHE (2006) Optimized parallel tempering simulations of proteins. J Chem Phys 124:174903

Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718

Wang F, Landau DP (2001) Determining the density of states for classical statistical models: a random walk algorithm to produce a flat histogram. Phys Rev E 64:056101

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 293:321–331

Yagisawa R, Kamiya N, Ikebe J, Umezawa K, Higo J (2008) Structure dependency of a 24-residue peptide humanin on solvent and preferential solvation by trifluoroethanol studied by multicanonical sampling. Chem Phys Lett 455:293–296

Yamane T, Okamura H, Nishimura Y, Kidera A, Ikeguchi M (2010) Side-chain conformational changes of transcription factor PhoB upon DNA binding: a population-shift mechanism. J Am Chem Soc 132:12653–12659

Zheng L, Chen M, Yang W (2008) Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. Proc Natl Acad Sci USA 105:20227–20232