Virtual states introduced for overcoming entropic barriers in conformational space

Junichi Higo and Haruki Nakamura

Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received May 6, 2012; accepted August 21, 2012

Free-energy landscape is an important quantity to study large-scale motions of a biomolecular system because it maps possible pathways for the motions. When the landscape consists of thermodynamically stable states (low-energy basins), which are connected by narrow conformational pathways (i.e., bottlenecks), the narrowness slows the inter-basin round trips in conformational sampling. This results in inaccuracy of free energies for the basins. This difficulty is not cleared out even when an enhanced conformational sampling is fairly performed along a reaction coordinate. In this study, to enhance the inter-basin round trips we introduced a virtual state that covers the narrow pathways. The probability distribution function for the virtual state was controlled based on detailed balance condition for the inter-state transitions (transitions between the real-state basins and the virtual state). To mimic the free-energy landscape of a real biological system, we introduced a simple model where a wall separates two basins and a narrow hole is pierced in the wall to connect the basins. The sampling was done based on Monte Carlo (MC). We examined several hole-sizes and inter-state transition probabilities. For a small hole-size, a small inter-state transition probability produced a sampling efficiency 100 times higher than a conventional MC does. This result goes against one’s intuition, because one considers generally that the sampling efficiency increases with increasing the transition probability. The present method is readily applicable to enhanced conformational sampling such as multicanonical or adaptive umbrella sampling, and extendable to molecular dynamics.

Key words: bottleneck, pathway, enhanced sampling, multicanonical, adaptive umbrella

Computer simulation is now widely used to explore the biomolecular conformational space. Free-energy (or energy) landscape is an important quantity obtainable from the simulation. The landscape provides distribution of thermodynamically stable states (low-energy basins) and pathways connecting the stable states. When the sampling is achieved in a wide conformational space, the landscape can be a road map for large-scale motions such as protein folding or protein-ligand binding. Figure 1 shows schematically the landscape, where the conformation passes through regions circled by broken lines, when an inter-basin transition occurs. To accurately estimate the free-energy difference between the low-energy basins, the conformational sampling should substantialize a number of transitions among the low-energy basins. However, when narrow pathways (bottlenecks) connect the low-energy basins, the frequency of transitions decreases, and the conformational sampling takes a long computing time to estimate accurately the probability (free energy) of each basin. Thus, generally, the computing time for the accurate free-energy estimation increases with narrowing the bottlenecks because the bottlenecks prevent inter-basin traveling. This difficulty lies in many sampling problems of biomolecular systems.

A generalized ensemble (GE) method, such as multicanonical sampling or adaptive umbrella sampling, generates an even (i.e. flat) probability distribution along a reaction coordinate. The flatness ensures that the conformational space is sampled widely along the reaction coordinate. In other words, the sampling is enhanced along the reaction coordinate. Thus, one may imagine that the GE method
can cause the passing through the bottlenecks effectively. However, we have shown that a fairly performed GE method provides even worse sampling efficiency than a conventional method when the conformational space on the reaction coordinate becomes narrow where the entropy suddenly decreases (i.e., the bottleneck appears on the pathway)\(^\text{13}\). Even the reaction coordinate is well designed so that the conformational changes along the reaction coordinate provide natural passage through the bottleneck, the conformation may be sluggish for long time in a basin before detecting the bottleneck.

In this study, we introduce a virtual state in the conformational space to ease passing the bottlenecks. Although this method is developed for sampling the biological systems, here we apply it to a simple system, which consists of two basins connected by a bottleneck. Because of the simplicity of the system, the free energies of the basins are computable analytically. We perform two Monte Carlo (MC) simulations called a “real-state MC” and a “virtual-state coupled MC”. The real-state MC is a conventional sampling method without the virtual state. In the virtual-state coupled MC, the virtual state covers the bottleneck. We show that the virtual state considerably enhances the inter-basin round trips.

**Methods**

Figure 1 is a scheme for the free-energy landscape of a biological system: See also Figure 1b of Ref. 3, which is the free-energy landscape of a β-hairpin peptide in explicit water. Thus, to increase the accuracy of the free energies for the low-energy basins, frequency of passing through the bottlenecks should be increased. To develop a useful method, we introduce a simple model explained below. A benefit of the simple model is that one can estimate analytically the free energies of the basins. We impose an important requirement on the model: a transition among the low-energy basins occurs through a narrow pathway. By varying the narrowness of the pathway, we can assess how our method is effective.

Figure 2 is the simple model, where the inner cavity of a rectangular cylinder is divided into two regions by a wall, and a narrow hole (circular cylinder-shaped hole) is pierced at the center of the wall (see Fig. 2A and B). The two wide regions, denoted as “state \(v_1\)” and “state \(v_2\)” in Figure 2C, resemble the low-energy basins, and the narrow hole does the bottleneck. A particle \(P_a\) is confined inside a rectangular cylinder and moves during a simulation to estimate free energies of the states \(v_1\) and \(v_2\). Thickness of the wall and radius of the hole are denoted as \(r_W\) and \(r_H\), respectively (see Fig. 2B). We define a cartesian coordinate system (see Fig. 2) so that the x-axis is the rectangular cylinder axis. The y-axis is parallel to one side of the basal planes of rectangular cylinder and the z-axis to the other side: The origin is set at the body center of the hole (i.e., the body center of the wall). The basal planes for states \(v_1\) and \(v_2\) are defined by \(x=x_1\) (\(x_1>0\)) and \(x=-x_2\) (\(x_2>0\)), respectively. The position of \(P_a\) is referred as to \(r=[x, y, z]\). For simplicity, we set the potential energy to be constant in the movable region of \(P_a\).
fore, the free-energy barrier caused by the hole is purely an entropic barrier.

First, we performed a conventional MC simulation and estimated numerically the volumes of the states \( v_1 \) and \( v_2 \), which are denoted \( V_{1\text{a}} \) and \( V_{2\text{a}} \) respectively. We call this sampling "real-state MC". Since the potential energy is constant, \( P_a \) can move unconditionally in the movable region. To count the round trips of \( P_a \) between \( v_1 \) and \( v_2 \), we introduced two regions, HR1 and HR2 (see neted areas in Fig. 1B), expressed by inequalities \( x \leq -x_0 \) and \( x \geq x_0 \) \((x_0 > 0)\), respectively. After \( P_a \) visits HR1 (or HR2), we wait till \( P_a \) returns to HR1, during which \( P_a \) visits HR2 at least once. We count this move as a round trip. When \( P_a \) has returned to HR1 without visiting HR2 (even with visiting state \( v_3 \)), this move is incomplete as a round trip. Then we wait further till \( P_a \) returns after visiting HR2. We denote the number of round trips from a long simulation as \( N_{RT} \).

Analytical values for \( V_{1\text{a}} \) and \( V_{2\text{a}} \) are given as:

\[
V_{1\text{a}}^\text{num} = S(x_r - h_\nu^2/2) + \pi r_\nu^2 r_\nu/2, \quad (i = 1, 2)
\]

(1)

where \( S \) is the area of the bases planes. The free-energy difference between \( v_1 \) and \( v_2 \) is:

\[
\Delta F = F_{v_2} - F_{v_1} = -\ln[V_{1\text{a}}^\text{num}/V_{2\text{a}}^\text{num}].
\]

We did not involve temperature in this expression by setting as \( k_B T = 1 \), because the potential energy is always constant in the movable region (i.e., \( V_{1\text{a}} \) and \( V_{2\text{a}} \) are independent of temperature). To check the convergence of sampling, we introduce a quantity, volume-ratio convergence, as:

\[
C(t) = \frac{V_{1\text{a}}^\text{num}/V_{2\text{a}}^\text{num}}{V_{1\text{a}}^\text{num}/V_{2\text{a}}^\text{num}},
\]

(2)

where \( V_{1\text{a}}^\text{num} \) is the numerically estimated volume for state \( v_i \) using a partial simulation trajectory from 0 to \( t \) steps. We denote a partial trajectory from \( t_i \) to \( t_j \) steps as \([t_i, t_j]\). Practically, \( V_{1\text{a}}^\text{num} \) is replaced by the number of snapshots where \( P_a \) exists in \( v_i \) in \([0, t]\).

Next, we introduce a spherical state \( v_3 \) centered at the coordinate origin with radius of \( r_\nu \) (Fig. 1C), where the left half \((x < 0)\) of \( v_3 \) overlaps with \( v_1 \), and the right half \((x \geq 0)\) with \( v_2 \). Therefore, \( v_3 \) is not a substantial state separated from \( v_1 \) and \( v_2 \), but a virtual state. Here, we introduce a virtual-state coupled MC simulation as follows: Suppose that \( P_a \) starts from a position in state \( v_3 \). During an interval \([0, \Delta t]\), we confine \( P_a \) to stay in \( v_3 \). When \( P_a \) is passing the \( v_1-v_3 \) boundary, this move is rejected, although \( P_a \) can pass the \( v_1-v_3 \) boundary freely. If \( P_a \) is inside of the \( v_1-v_3 \) boundary at the last step of \([0, \Delta t]\), \( P_a \) may transition to \( v_3 \) with a transition probability of \( p_{v_1-v_3} \) (the actual values for the transition probability is given later). Note that this inter-state transition alters the attribution of \( P_a \) (i.e. the state specifier) from \( v_1 \) to \( v_3 \) without changing the position of \( P_a \) in the rectangular cylinder. If \( P_a \) is outside the \( v_1-v_3 \) boundary at the last snapshot, no transition occurs. The particle \( P_a \) is confined again in \( v_1 \) during the next interval \([\Delta t, 2\Delta t]\), and the state transition is examined at the last step of \([\Delta t, 2\Delta t]\).

Once the inter-state transition \((v_1 \rightarrow v_3)\) has been accepted at the end of \([0, \Delta t]\), then \( P_a \) is confined in \( v_1 \) during the interval of \([\Delta t, 2\Delta t]\). Now, \( P_a \) can pass the \( v_1-v_3 \) boundary freely. However, moves toward outside of the \( v_1-v_3 \) and \( v_3-v_2 \) boundaries are rejected. At the last step of \([\Delta t, 2\Delta t]\), the state transition is examined as follows: \( P_a \) may return back to \( v_1 \) with a probability of \( p_{v_3-v_1} \) when \( P_a \) is in the region of \( x < 0 \). On the other hand, \( P_a \) may transition to \( v_3 \) with a probability of \( p_{v_3-v_2} \) when \( P_a \) is in the region of \( x > 0 \). Suppose that \( P_a \) has transitioned to \( v_3 \). Then \( P_a \) is confined in \( v_3 \) during the interval of \([2\Delta t, 3\Delta t]\), where \( P_a \) can pass the \( v_3-v_2 \) boundary freely, but moves passing the \( v_1-v_3 \) boundary are rejected. At the last step of \([2\Delta t, 3\Delta t]\), \( P_a \) may transition to \( v_3 \) with a transition probability of \( p_{v_3-v_2} \) when \( P_a \) is in the region of \( r \leq r_v \), where \( r = [x^2 + y^2 + z^2]^{1/2} \). Otherwise, \( P_a \) stays in \( v_3 \) for the next interval \([3\Delta t, 4\Delta t]\). By this way, the inter-state transition is examined at steps \( n\Delta t \) \((n = 1, 2, \ldots)\).

Since the motion of \( P_a \) from \( r \) to \( r' \) within \( v_i \) is unconditionally accepted, a long simulation yields an equation:

\[
P(r, v_i) = P(r', v_i) = \text{const},
\]

(3)

where \( P(r, v_i) \) is the probability distribution function of \( P_a \) at position \( r \) in \( v_i \). An inter-state transition from \( v_i \) to \( v_j \) at a position \( r \) is controlled by a transition probability \( p_{v_i-v_j} \). Then, the long simulation yields an equation:

\[
\frac{P(r, v_i)}{P(r, v_j)} = \frac{p_{v_i-v_j}}{p_{v_i-v_j}}.
\]

(4)

Combining Eqs. 3 and 4, the probability distribution function satisfies the following equation:

\[
\frac{P(r_1, v_1)}{P(r_2, v_2)} = \frac{P(r_1, v_3)P(r_2, v_3)P(r_3, v_3)}{P(r_1, v_3)P(r_2, v_3)P(r_3, v_3)}
\]

\[
= \frac{p_{v_3-v_1}p_{v_3-v_2}p_{v_3-v_2}}{p_{v_3-v_1}p_{v_3-v_2}p_{v_3-v_2}}
\]

(5)

where \( r_1 \) and \( r_2 \) are arbitrary positions in \( v_1 \) and \( v_2 \), respectively. Equation 5 indicates that the states \( v_1 \) and \( v_3 \) are indirectly linked via the virtual state \( v_3 \).

We note that Eq. 5 is not an equation to determine the absolute values for the inter-state rate constants. Equation 5 can be rewritten as:

\[
\frac{P(r_1, v_1)}{P(r_2, v_2)} = \frac{p_{v_3-v_1}p_{v_3-v_2}p_{v_3-v_2}}{p_{v_3-v_1}p_{v_3-v_2}p_{v_3-v_2}}
\]

(6)

where the inter-state transition probabilities are redefined as:

\[
p_{v_i-v_j} = kp_{v_i-v_j},
\]

(7)

where the coefficient \( k \) is any positive value. With increasing \( k \), the rate constant increases, and the simulation length to reach equilibrium becomes short. In MC scheme, then, the largest rate constant (the quickest convergence) may be result from \( p_{v_3-v_1} = 1 \). The parameter \( \Delta t \) also changes the rate constants: the larger the \( \Delta t \), the longer the simulation length...
to obtained an equilibrated probability distribution function.
Based on the above discussion, the parameter set of $[p_{t=v}, \Delta t]=[1, 1]$ may be the best to speed up the round trips. Note that this parameter set is similar with the simulation condition of the real-state MC (i.e., the conventional MC). In fact, we show below that the smaller the $\Delta t$, the quicker the convergence. However, against our better instincts, the larger the $p_{t=v}$, the worse the sampling efficiency for a narrow-hole system.

**Results and discussion**

We set the system parameters (non-dimensional quantities) as $S=2.0^2=4$, $r_w=0.2$, $x_1=1.0$, $x_2=2.0$ and $x_0=0.5$. Quantities $r_{RT}$, $\Delta t$, and $p_{t=v}$ are specific to the virtual state. We examined several values for $r_{RT}$ and $\Delta t$: $r_{RT}=[0.005, 0.01, \ldots, 0.16]=[d \times 2^0, d \times 2^1, \ldots, d \times 2^9]$ where $d=0.005$, and $\Delta t=[1, 10, \ldots, 10000]=[10^0, 10^1, \ldots, 10^5]$. For simplicity, the transition probabilities from the virtual state to the real states are set to a single value $p_v$, and those from the real states to the virtual state to 1.0: I.e., $p_{t=v}=p_{t=v}^r (0<p_v\leq 1)$ and $p_{t=v}=1.0$. We examined ten $p_v$ values as: $p_v=[2^{-9}, 2^{-8}, \ldots, 2^{-1}, 1]$. The size of the virtual state $v$ was set to 0.3. The total MC length (number of trials to move $P_j$) is $5 \times 10^{11}$ steps for all simulations.

In the real-state MC, $N_{RT}$ rapidly decreased with decreasing $r_{RT}$ (Fig. 3A). Figure 3B demonstrates the volume-ratio convergence $C(t)$ for $r_{RT}=0.16$ and $r_{RT}=0.005$. The convergence was quick for $r_{RT}=0.16$ and slow for $r_{RT}=0.005$.

Current study presents a recipe to enhance the sampling by introducing the virtual state for a narrow-hole system. The $p_{t=v}$-$N_{RT}$ relation (Fig. 4A) for the narrowest-hole system ($r_{RT}=0.005$) at various $\Delta t$ manifests that the virtual state enhances the sampling because $N_{RT}$ is larger than that from the real-state MC (broken line). The only exception was found at $[p_v, \Delta t]=[1, 10^0]$. For $\Delta t\leq 10^4$, $N_{RT}$ increased monotonically with decreasing $p_v$. The highest efficiency was found at $p_v=2^{-9}$, where $N_{RT}$ was about 100 times larger than that from the real-state MC. Figure 4B plots $C(t)$ from $[p_v, \Delta t]=[2^{-9}, 1]$ and $[1, 1]$. The convergence was quick for $p_v=2^{-9}$ and slow for $p_v=1$. The mechanism for the enhancement is simple: With decreasing $p_v$, the probability $P(r_{t=v})$ of $r_{t=v}$ in the virtual state increases, and accordingly the hole-passing chance increases.

The black solid line ($\Delta t=10^0$) of Figure 4A had different behavior from the other lines ($\Delta t\leq 10^4$): $N_{RT}$ had a peak at $p_v=2^{-7}$. Let us consider a situation that $p_v$ falls to zero ($p_v\to 0$). In this extremity, $N_{RT}$ should decay to zero ($N_{RT}\to 0$) because $P_r$ cannot escape from $v_1$ once $P_r$ is trapped in $v_1$. Then, we get two inequalities: $dN_{RT}/dp_v|_{p_v>0}>0$ and $dN_{RT}/dp_v|_{p_v=1}<0$. These inequalities result in that $N_{RT}$ has a peak. The $N_{RT}$ showed no peak for $\Delta t\leq 10^4$ because the peak position is below $p_v=2^{-7}$. Here, one may raise a question: Why does the peak position for $\Delta t=10^0$ was larger than those for the other $\Delta t$? This is because increment of $\Delta t$ decreases the inter-state transitions (the rate constants). In this regime, a large $p_v$ (transition probability from $v_1$ to $v_1$ or $v_v$) plays a role of enhancer for inter-state transitions. Accordingly, the peak position shifts positively.

The positive shift of the peak position is clearly shown for a middle hole $r_{RT}=0.02$ (Fig. 5A): All curves had a peak, and the peak position shifted positively with increasing $\Delta t$. Figure 5B demonstrates the $p_{t=v}$-$N_{RT}$ relation for the largest hole $r_{RT}=0.16$, where the peak position was merge into the edge of $p_v$ ($p_v=1$) for $\Delta t\geq 10^4$.

The virtual-state coupling is readily applicable to GE methods, such as multicanonical or adaptive umbrella sampling. The GE enhances the sampling along a reaction coordinate $x$ with introducing an effective potential energy $E_{eff}=E+k_BT\ln[P_e(x,T)]$, where $E$ is the original potential energy, $P_e(x,T)$ a canonical distribution of $x$ at temperature $T$, and $k_B$ the Boltzmann constant. The reaction coordinate is a function of
position (or conformation) $r: x=x(r)$. The positional transition probability from $r_1$ to $r_2$ is simply given by $\exp[-\Delta E_{\text{eff}}/k_BT]$, where $\Delta E_{\text{eff}} = E_{\text{eff}}(x(r_2)) - E_{\text{eff}}(x(r_1))$. The phase point fluctuating in the entire conformational space may spend a long time before running into the narrow bottleneck. One can set the virtual state covering the bottleneck, where the virtual volume is larger than the bottleneck size. Then, the chance that the phase point finds the bottleneck increases. In the virtual state, the phase point can find bottleneck readily because the phase point is confined in the virtual state for a while.

We note that the current method is expandable readily to molecular dynamics (MD) by computing an effective force $f_{i}^{\text{eff}}$ acting on atom $i$ as: $f_i = -\text{grad}[E_{\text{eff}}(x, T)]$. With introducing the virtual state, a canonical MD simulation at temperature $T$ with $f_{i}^{\text{eff}}$ becomes virtual-state coupled multicanonical or adaptive umbrella sampling. Conformational changes in each time interval $\Delta t$ are done by MD scheme, and the inter-state transitions examined at the end of time intervals ($t = n\Delta t$) are achieved by MC scheme.

In the present study, the bottleneck position in the conformational space is known in advance. However, the bottleneck position is generally unknown a priori. Then, before introducing the virtual state, pre-sampling is required, which may be a "real-state" multicanonical or adaptive umbrella sampling. In general, shape and volume of the virtual state are arbitrary depending on the system.

We note that the current method is expandable readily to molecular dynamics (MD) by computing an effective force $f_{i}^{\text{eff}}$ acting on atom $i$ as: $f_i = -\text{grad}[E_{\text{eff}}(x, T)]$. With introducing the virtual state, a canonical MD simulation at temperature $T$ with $f_{i}^{\text{eff}}$ becomes virtual-state coupled multicanonical or adaptive umbrella sampling. Conformational changes in each time interval $\Delta t$ are done by MD scheme, and the inter-state transitions examined at the end of time intervals ($t = n\Delta t$) are achieved by MC scheme.

In the present study, the bottleneck position in the conformational space is known in advance. However, the bottleneck position is generally unknown a priori. Then, before introducing the virtual state, pre-sampling is required, which may be a "real-state" multicanonical or adaptive umbrella sampling. In general, shape and volume of the virtual state are arbitrary depending on the system.

The current model (Fig. 2) was defined in the three-dimensional space. Then, one may doubt if the current method is useful for biological systems, because the biological systems are defined in a high-dimensional space.
However, as shown in all-atom model of polypeptides\textsuperscript{1–3}, the bottlenecks are well identified by projecting the high-dimensional distribution in a low-dimensional (2D or 3D) conformational space.

Last, we note that the virtual state acts as a lens to view the probability distribution function at the bottleneck. To increase the events passing through the bottleneck, we set $p_t$ to be small, which makes the probability distribution for the virtual state large. Thus, the virtual-state coupled sampling can be used to estimate both free energies of major basins and the bottlenecks (i.e., free-energy barriers).

Acknowledgement

H. N. was supported by Grant-in-Aid for Scientific Research (B) (23370071) from the Japan Society for the Promotion of Science. J. H. was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (21113006) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan. J. H. and H. N. were supported by grants from the New Energy and Industrial Technology Development Organization (NEDO) Japan.

References

1. Ono, S., Nakajima, N., Higo, J. & Nakamura, H. The multicanonical weighted histogram analysis method for free energy landscape along structural transition paths. *Chem. Phys. Lett.* **312**, 247–254 (1999).
2. Higo, J., Galzitskaya, O. V., Ono, S. & Nakamura, H. Energy landscape of a β-hairpin peptide in explicit water studied by multicanonical molecular dynamics. *Chem. Phys. Lett.* **337**, 169–175 (2001).
3. Kamiya, N., Higo, J. & Nakamura, H. Conformational transition states of a β-hairpin peptide between the ordered and disordered conformations in explicit water. *Protein Sci.* **11**, 2297–2307 (2002).
4. Kamiya, N., Mitomo, D., Shea, J.-E. & Higo, J. Folding of the 25 residue AB(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 (2007).
5. Higo, J., Nishimura, Y. & Nakamura, H. 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 (2011).
6. Berg, B. A. & Neuhaus, T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.* **68**, 9–12 (1992).
7. Hansmann, U. H. E. & Okamoto, Y. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem. *J. Comp. Chem.* **14**, 1333–1338 (1993).
8. Kidera, A. Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide. *Proc. Natl. Acad. Sci. USA* **92**, 9886–9889 (1995).
9. Nakajima, N., Nakamura, H. & Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B* **101**, 817–824 (1997).
10. Iba, Y., Chikenji, G. & Kikuchi, M. Simulation of Lattice Polymers with multi-self-overlap ensemble. *J. Phys. Soc. Jpn.* **67**, 3327–3330 (1998).
11. Paine, G. H. & Scheraga, H. A. Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. I. Backbone structure of enkephalin. *Biopolymers* **24**, 1391–1436 (1985).
12. Mezei, M. Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *J. Comp. Phys.* **68**, 237–248 (1987).
13. Higo, J., Ikebe, J., Kamiya, N. & Nakamura, H. Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. *Biophys. Rev.* **4**, 27–44 (2012).