Statistical temperature molecular dynamics: Application to coarse-grained β-barrel-forming protein models

Jaegil Kim,[a] John E. Straub, and Thomas Keyes
Department of Chemistry, Boston University, Boston, Massachusetts 02215

(Received 22 December 2006; accepted 1 February 2007; published online 2 April 2007)

Recently the authors proposed a novel sampling algorithm, "statistical temperature molecular dynamics" (STMD) [J. Kim *et al.*, Phys. Rev. Lett. **97**, 050601 (2006)], which combines ingredients of multicanonical molecular dynamics and Wang-Landau sampling. Exploiting the relation between the statistical temperature and the density of states, STMD generates a flat energy distribution and efficient sampling with a dynamic update of the statistical temperature, transforming an initial constant estimate to the true statistical temperature $T(U)$, with U being the potential energy. Here, the performance of STMD is examined in the Lennard-Jones fluid with diverse simulation conditions, and in the coarse-grained, off-lattice *BLN* 46-mer and 69-mer protein models, exhibiting rugged potential energy landscapes with a high degree of frustration. STMD simulations combined with inherent structure (IS) analysis allow an accurate determination of protein thermodynamics down to very low temperatures, overcoming quasiergodicity, and illuminate the transitions occurring in folding in terms of the energy landscape. It is found that a thermodynamic signature of folding is significantly suppressed by accurate sampling, due to an incoherent contribution from low-lying non-native IS in multifunneled landscapes. It is also shown that preferred accessibility to such IS during the collapse transition is intimately related to misfolding or poor foldability. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2711812]

I. INTRODUCTION

The potential energy landscape (PEL) of complex systems is characterized by a multitude of local minima separated by barriers.[1] Thus conventional canonical simulations using Monte Carlo (MC) or molecular dynamics (MD) algorithms can fail to sample the thermally significant phase space within a reasonable computational time, due to broken ergodicity.[2,3] To mitigate this problem, several advanced sampling methods have been developed such as multiple histogram reweighting,[4] multicanonical or entropic sampling,[5,6] replica exchange method or parallel tempering,[7–9] and the Wang-Landau (WL) random walk algorithm.[10] Recently, a feedback iteration algorithm has also been proposed to improve the efficiency of the broad histogram method[11] and parallel tempering.[12]

While the various algorithms differ substantially in detail, the common goal is to calculate the density of states, $\Omega(U)$, representing the degeneracy of available energy levels, where U is the potential energy;all thermodynamic quantities can then be calculated from the canonical partition function, $Z_{\text{cano}}(\beta)=\Sigma\Omega(U)e^{-\beta U}$, $\beta=1/k_BT$, k_B being the Boltzmann constant. Recently, Wang-Landau sampling[10,13] has attracted considerable interest, due to its conceptual framework and ability to facilitate fast equilibration by generating a flat distribution in potential energy. Several modified versions, combined with the configurational temperature[14] or transition-matrix algorithm,[15] have been applied to problems of phase transitions in lattice spins,[10] vapor-liquid equilibria of fluids,[14,15] biomolecules,[16–18] and Lennard-Jones[19] and spin glasses.[13]

The basic idea of WL sampling is most similar to that of multicanonical sampling[5] in that both methods achieve a flat energy distribution employing a weight, $w(U)=1/\Omega(U)$, which permits the system to visit otherwise rarely sampled regions via a random walk in energy. However, the density of states is not known *a priori*, so it is determined by an iterative procedure.[5,6] The distinguishing feature of WL sampling is its update scheme for the running estimate $\tilde{\Omega}(U)$. Every time that an energy U is visited, $\tilde{\Omega}(U)$ is multiplied by the modification factor $f(>1)$, $\tilde{\Omega}(U)\rightarrow f\tilde{\Omega}(U)$. This operation biases the acceptance probability $A(\mathbf{r}\rightarrow\mathbf{r}')=\min[1,\tilde{\Omega}(U)/\tilde{\Omega}(U')]$, where $U=U(\mathbf{r})$ and $U'=U(\mathbf{r}')$, and causes the system to move to less explored energy regions. In contrast to the recursive refinements for $\tilde{\Omega}(U)$ in multicanonical sampling, the dynamical update of $\tilde{\Omega}(U)$ enables a faster exploration of configuration space and a direct estimate of the true density of states in the asymptotic limit of $f\rightarrow 1$.

In spite of many successful applications of WL sampling, nontrivial modifications are required for continuum and large systems.[15,20] In particular, the discrete representation of $\Omega(U)$ on an energy grid can cause an accumulation problem[21,22] when the density of states shows a narrow entropic bottleneck. This tendency becomes more severe for large systems, where a huge number of energy bins is required to cover an increased range of $\Omega(U)$. To resolve this

[a]Electronic mail: jaegil@bu.edu

problem the original method has been recently modified to use a continuum density of states with a kernel function update.[23]

Another obstacle is the absence of a MD algorithm capable of handling the dynamic modification of the sampling weight. Prior implementations of WL sampling have been based upon MC. This reduces the applicability of the method to more complex systems where effective MC moves are not available. One attempt[17] has been made to use a short molecular dynamics simulation for the generation of trial moves using importance sampling with the density of states estimate $\tilde{\Omega}(U)$. However, this is still far from a genuine MD algorithm generating a deterministic trajectory.

Recently, we proposed the "statistical temperature molecular dynamics" (STMD) algorithm,[24] which integrates Wang-Landau sampling[10] and multicanonical MD (Ref. 25) via the statistical temperature. STMD relies on the well-known thermodynamic relationship between the density of states and the statistical temperature,[26]

$$T(U) = \left[\frac{\partial \ln \Omega(U)}{\partial U}\right]^{-1}. \qquad (1)$$

With use of Eq. (1), a flat energy distribution is obtained via the systematic refinement for the statistical temperature estimate, $\tilde{T}(U)$, rather than the density of states estimate, $\tilde{\Omega}(U)$. Applying the basic WL idea to the finite difference form of Eq. (1) yields a robust update scheme, transforming an initially constant $\tilde{T}(U)$ to the true statistical temperature $T(U)$. The updating of $\tilde{T}(U)$ is intrinsically nonlocal, refining $\tilde{\Omega}(U)$ concurrently at not only the visited state but also its neighborhood, and is easily implemented into molecular dynamics simulations through a force scaling combined with a Nose-Hoover thermostat.[27] STMD is applicable to complex systems with rough energy landscapes, overcoming the slow convergence and the unknown weight dependence of multicanonical MD.

In the present paper, we first examine the performance of STMD in the 110-particle Lennard-Jones fluid, with diverse simulation conditions. The accuracy of STMD simulations with a stepwise interpolation function for $\tilde{T}(U)$ is tested by comparing reweighted thermodynamic properties with the results of conventional canonical MD. We find that the rate of convergence of STMD can be considerably accelerated with the use of a large energy bin, with no deterioration of statistical accuracy.

Second, the applicability of STMD to biomolecular simulations is explored in coarse-grained *BLN* 46-mer[28] and 69-mer[29] protein models, in which a rugged potential energy landscape hampers the computation of thermodynamic behavior at low temperatures. It is shown that STMD yields accurate thermodynamic properties down to very low temperatures and that the thermodynamic signatures of folding are strongly influenced by the sampling efficiency in a multifunnel energy landscape.

The inherent structures[30] (IS) are the local minima of the PEL and serve as discrete states for continuous systems. An instantaneous configuration belongs to the IS to which it maps upon minimization. STMD sampling and the IS approach make a powerful combination for proteins. We refer to the lowest energy IS as the ground state and to the others as excited states; a subscript indexes the IS and their properties in order of increasing energy. Protein folding occurs when, by some criterion, the "native state" is reached. Native IS, including the ground state, share the structural motif of the native state and belong to the folding funnel.

We previously found[24,31] that the occupation probabilities of individual low-lying IS become finite below the collapse transition. Excited state occupation probabilities remain non-negligible below the previously estimated folding temperature, blurring the folding signature with an incoherent contribution of non-native IS to the thermodynamics. There exists a temperature interval in which the occupation of non-native IS exceeds the native occupation, and the extent of this interval strongly correlates with frustration on the PEL. The connection between the preferred occupation of non-native IS after collapse and poor foldability of β-barrel-forming proteins is further explored in the following.

The paper is organized as follows: In Sec. II, the basic formulation and the detailed simulation protocols of STMD are presented. The implicit connection of STMD to the original multicanonical MD and WL samplings is also verified through the one-to-one mapping between $\tilde{T}(U)$ and $\tilde{\Omega}(U)$ using stochastic sampling dynamics. In Sec. III, the basic advantages of STMD are examined for the 110-particle Lennard-Jones fluid by varying the energy bin size. In Sec. IV, STMD is applied to the *BLN* 46-mer and 69-mer model proteins. Folding is investigated in terms of equilibrium and inherent structure thermodynamics. The conclusion and a brief summary are presented in Sec. V.

II. THEORETICAL FORMULATION

A. Dynamical update method for the statistical temperature

Our study begins with the thermodynamic relationship between the microcanonical entropy, $S(U) = \ln \Omega(U)$, and the statistical inverse temperature, $\beta(U) = 1/T(U)$,

$$S(U) = \int^{U} \beta(U')dU'. \qquad (2)$$

Throughout this paper we set $k_B = 1$. Since $S(U)$ is uniquely determined, up to a constant, as a functional of $T(U)$, it is natural to seek a WL-type sampling driven by an iterative refinement of the statistical temperature, rather than the entropy or the density of states.

Thus, we introduce the running estimate for the statistical temperature,

$$\tilde{\beta}(U) = 1/\tilde{T}(U) = [\partial \tilde{S}/\partial U], \qquad (3)$$

where $\tilde{S}(U) = \ln \tilde{\Omega}(U)$ is the entropy estimate. On an equally spaced energy grid $U_j = G(U/\Delta)\Delta$, with bin size Δ and $G(x)$ returning the nearest integer to x, the multiplicative WL operation of $\tilde{\Omega}_j \to f\tilde{\Omega}_j$ reduces to $\tilde{S}_j \to \tilde{S}_j + \ln f$ for a visit to "state" j with energy U_j. Combining this algebraic operation with the central finite difference approximation for Eq. (3),

$$\partial_U \tilde{S}|_{U_j} = \tilde{\beta}_j = 1/\tilde{T}_j \simeq (\tilde{S}_{j+1} - \tilde{S}_{j-1})/2\Delta, \tag{4}$$

we obtain the dynamic update scheme for the inverse temperature,

$$\tilde{\beta}'_{j\pm1} = \tilde{\beta}_{j\pm1} \mp \delta f, \tag{5}$$

with $\delta f = \ln f/2\Delta$, and the prime denotes the updated values.

Except for phase transition regions in finite size systems,[33] the true inverse temperature $\beta(U)$ monotonically decreases to zero as the energy increases. This means that the estimate could go negative, with consecutive subtractions by δf, when the system visits the same energy state repeatedly. To avoid this problem, Eq. (5) is further transformed by taking the inverse and rewriting it in terms of $\tilde{T}(U)$,

$$\tilde{T}'_{j\pm1} = \alpha_{j\pm1}\tilde{T}_{j\pm1}, \tag{6}$$

where $\alpha_{j\pm1} = 1/(1 \mp \delta f \tilde{T}_{j\pm1})$.

Equation (6) is very suggestive in that (i) the scaling operations of decreasing $\tilde{T}_{j-1}$ and increasing $\tilde{T}_{j+1}$ transform $\tilde{T}(U)$ so that it converges to the monotonically increasing $T(U)$, (ii) the scaling factor $\alpha_{j\pm1}$ approaches unity at low temperature, allowing a fine tuning of $\tilde{T}(U)$ even with repeated visits to the same energy state due to a localized energy distribution and trapping in local minima, and (iii) the "edge" effect[34] can be avoided by restricting updates for $T_l < \tilde{T}_j < T_h$ and maintaining $\tilde{T}_j = T_l$ and T_h beyond lower and upper temperature bounds T_l and T_h, respectively. The constant temperature estimates at both ends of the energy region under study cause the system to sample canonical ensembles with temperatures T_l and T_h, respectively.

B. Implementation into molecular dynamics simulation

Since STMD updates the intensive variable $\tilde{T}(U)$, rather than $\tilde{\Omega}(U) \sim O(e^N)$, it can be naturally combined with a molecular dynamics algorithm using the generalized ensemble simulation technique,[25] in which a non-Boltzmann sampling is attained by scaling the potential and maintaining the kinetic energy at a reference inverse temperature $\beta_0 = 1/T_0$. Considering the generalized ensemble associated with the weight,

$$w(U) = e^{-\int^U (1/\tilde{T}(U'))dU'} = e^{-\beta_0 v(U)}, \tag{7}$$

as a canonical ensemble with the effective potential $v(U) = T_0\tilde{S}(U)$, MD with the effective potential plus the dynamic modification for the temperature estimate yields STMD. The forces are constantly adjusted by an energy dependent scaling factor, $\gamma(U) = T_0/\tilde{T}(U)$,

$$\tilde{\mathbf{f}}_i = \gamma(U)\mathbf{f}_i, \tag{8}$$

where $\mathbf{f}_i$ is the force on particle i without the scaling. The velocity distribution is maintained at the reference temperature T_0 using a Nose-Hoover thermostat.[27] The result is a trajectory sampling the weight $w(U)$ in configurational space,[35] with a probability density function (PDF) obeying

$$P(U) \sim e^{S(U)-\tilde{S}(U)} = e^{\int^U \delta\beta(U')dU'}, \tag{9}$$

where $\delta\beta(U) = \beta(U) - \tilde{\beta}(U) = \delta T(U)/(T(U)\tilde{T}(U))$, $\delta T(U) = \tilde{T}(U) - T(U)$.

Initially the force scaling factor, $\gamma(U)$, is changing and the trajectory is not in equilibrium.[10] Every time a flat energy distribution is obtained with a given f, the simulation is repeated with a reduced modification factor, $\sqrt{f}$, starting from the previous temperature estimate. In the asymptotic limit of $f \to 1$ (or $\delta f \to 0$), detailed balance is recovered and the running estimate $\tilde{T}(U)$ converges to the true $T(U)$, producing an equilibrium trajectory subject to the weight in Eq. (7).

The stationary sampling dynamics leading to Eq. (9) can[36,37] be described as diffusion in energy, modeled by a Langevin equation,

$$\partial_t U = \delta\beta(U) + \eta(t), \tag{10}$$

where $\eta(t)$ is the random force. The coincidence of $\tilde{T}(U)$ with $T(U)$ realizes a random walk by canceling the deterministic force, $\delta\beta(U)$. For a weakly nonstationary case, where the modification of the temperature estimate per time step is small due to a small δf, i.e., $\delta f \tilde{T}^2(U) \ll 1$, one may see how the systematic bias in $\delta\beta(U)$ allows escape from a trap. For example, if the system gets confined in some energy region near U_j, the accumulated operations of Eq. (6) on the several corresponding visits to U_j create statistical temperature gradients of $\delta\beta_{j-1} < 0$ and $\delta\beta_{j+1} > 0$, which form an outgoing probability flux from U_j and assist the system to escape. Consequently, the system moves freely through the accessible energy range.

C. Correspondence between the temperature estimate and the entropy estimate

When the simulation has converged, thermodynamic quantities can be calculated by determining $\tilde{S}(U)$ from Eq. (2). However, note that $\tilde{T}(U)$ is only defined on the grid points U_j, while the force scaling factor $\gamma(U)$ requires a continuum description. Furthermore, the integrand $\tilde{\beta}(U)$ in Eq. (2) shows a very steep variation at low temperatures, so a direct numerical integration for $\tilde{S}(U)$ is undesirable. Here we present two interpolation methods which provide off-grid values of $\tilde{T}(U)$, yielding continuum entropy estimates by analytic integration.

1. Staircase temperature estimate

We first applied the simplest staircase interpolation,

$$\tilde{T}(U) = \sum_i \tilde{T}_i \theta(U - \tilde{U}_{i-1})\theta(\tilde{U}_i - U), \tag{11}$$

with θ being the Heaviside step function and $\bar{U}_i = (U_i + U_{i+1})/2$ denoting the energy midway between the grid points. Since the stepwise estimate is constant for each energy bin, the entropy estimate is straightforward,

$$\tilde{S}(U) = \int_{U_l}^{U} \tilde{\beta}(U')dU' = \sum_{j=l}^{i-1} \frac{\Delta}{\tilde{T}_j} + \frac{(U - \bar{U}_{i-1})}{\tilde{T}_i}, \tag{12}$$

for $U \in [\bar{U}_{i-1}, \bar{U}_i]$. Here U_l is an arbitrarily defined lower integration limit. Then, under the scaling operation for $\tilde{T}(U)$ on the visit to U_i, the entropy estimate is modified as

$$\tilde{S}'(U) = \begin{cases} \tilde{S}(U) + \delta f(U - \bar{U}_{i-2}) & \text{for } U \in [\bar{U}_{i-2}, \bar{U}_{i-1}] \\ \tilde{S}(U) + \delta f \Delta & \text{for } U \in [\bar{U}_{i-1}, \bar{U}_i] \\ S(U) + \delta f(\bar{U}_{i+1} - U) & \text{for } U \in [\bar{U}_i, \bar{U}_{i+1}] \\ \tilde{S}(U) & \text{otherwise.} \end{cases}$$

Thus, on a visit to state i, the entropy estimate always increases for the $(i-1)$th bin, is shifted by the constant $\delta f \Delta$ for the ith bin, and linearly decreases towards the original value at the $(i+1)$th bin, clearly demonstrating that our update scheme, Eq. (6), is nonlocal for $\tilde{\Omega}(U)$.

2. Linear temperature estimate

The second interpolation method is to connect successive grid points linearly,

$$\tilde{T}(U) = \tilde{T}_j + \lambda_j(U - U_j), \tag{13}$$

for $U \in [U_j, U_{j+1}]$, where $\lambda_j = (\tilde{T}_{j+1} - \tilde{T}_j)/\Delta$ is the slope of the linear segment connecting $[U_j, \tilde{T}_j]$ and $[U_{j+1}, \tilde{T}_{j+1}]$. Linear interpolation is particularly appropriate at low temperatures, where the heat capacity is nearly constant, but the sequence of consecutive interpolations also enables a faithful representation of $\tilde{T}(U)$ corresponding to a phase transition.[36] Equation (13) yields a continuum entropy estimate by an analytic integration,

$$\tilde{S}(U) = \int_{U_l}^{U} \tilde{\beta}(U')dU' = \sum_{j=l+1}^{i^*} L_j(U_j) + L_{i+1}(U), \tag{14}$$

where $i^* = i-1(i)$ for $\bar{U}_{i-1} \leqslant U \leqslant U_i (U_i \leqslant U \leqslant \bar{U}_i)$ and $L_j = \lambda_{j-1}^{-1} \ln[1 + \lambda_{j-1}(U - U_{j-1})/\tilde{T}_{j-1}]$.

Denoting $\tilde{\Omega}_i = e^{\tilde{S}(U_i)} = \Pi_{j=1}^{i} Y_j$, $Y_j = \exp\{L(U_j)\} = [\tilde{T}_j/\tilde{T}_{j-1}]^{\Delta/(\tilde{T}_j - \tilde{T}_{j-1})}$, the scaling operation of Eq. (6) upon a visit to state i modifies $\tilde{\Omega}_j$ by multiplying it by the nonuniform modification factor f_k for $k \in [-1, 2]$,

$$\tilde{\Omega}'_{i+k} = f_k \tilde{\Omega}_{i+k}, \tag{15}$$

where $f_k = \Pi_{j=i-1}^{i+k} Q_j (= Y'_j / Y_j)$, with Y'_j being evaluated at the updated $\tilde{T}'_j$. By expanding the exponent of $Y_j \simeq \exp\{\Delta/\tilde{T}_{j-1}(1 - \Delta_j/2\tilde{T}_{j-1})\}$ or $\exp\{\Delta/\tilde{T}_j(1 + \Delta_j/2\tilde{T}_j)\}$ to first order with respect to $\Delta_j = \tilde{T}_j - \tilde{T}_{j-1} \ll 1$, and using $\tilde{T}'_{j\pm 1} \simeq \tilde{T}_{j\pm 1} \pm \delta f \tilde{T}^2_{j\pm 1}$, we identified $Q_{i-1} \simeq f^{(\tilde{T}_{i-1}/2\tilde{T}_{i-2})^2}$, $Q_i \simeq f^{(\tilde{T}_{i-1}/2\tilde{T}_i)^2}$, $Q_{i+1} \simeq 1/f^{(\tilde{T}_{i+1}/2\tilde{T}_i)^2}$, and $Q_{i+2} \simeq 1/f^{(\tilde{T}_{i+1}/2\tilde{T}_{i+2})^2}$. Since both Q_{i-1} and Q_i are always greater than 1 while both Q_{i+1} and Q_{i+2} are always less than 1, the update operation creates an upward curvature in $\tilde{\Omega}'_j$ around the visited ith state. By taking the approximation $\tilde{T}_{j'}/\tilde{T}_j \simeq 1$, for $|j - j'| = 1$, Eq. (15) is further simplified to the symmetric operation for the entropy estimate, $\tilde{S}'_j = \tilde{S}_j + \frac{1}{2} \ln f$ for $j = i$ and $\tilde{S}_j + \frac{1}{4} \ln f$ for $j = i \pm 1$, demonstrating again that the update of $\tilde{T}(U)$ is intrinsically nonlocal.

D. Reweighting

The combination of the fundamental Eq. (6) and the mathematical mapping between $\tilde{T}(U)$ and $\tilde{S}(U)$ through the smoothing of Eqs. (12) and (14) allows STMD to handle continuum systems regardless of the size, Δ, of the energy bin. STMD can maintain statistical accuracy with a large Δ, which is potentially very useful for systems having a large range of $\Omega(U)$, while multicanonical-type simulations based on the accumulated histogram cannot afford a large energy bin, since the determination of the sampling weight is directly influenced by the statistical accuracy of the energy distribution, $P(U)$. The significant slowing down of the convergence of WL sampling with increasing Δ has been also pointed out in Ising spins and the Lennard-Jones fluid.[24]

For a large Δ the entropy estimate can be corrected as

$$\tilde{S}'(U) = \tilde{S}(U) + \ln P(U), \tag{16}$$

and an arbitrary canonical averaged observable $A(\beta)$ can be determined as $\int dUP_{\text{cano}}(U,\beta)A(U)$, with $P_{\text{cano}}(U;\beta) = \exp\{S(U) - \beta U\}/Z_{\text{cano}}(\beta)$ being the normalized canonical PDF.

Sometimes, it is more convenient to work with the raw data themselves instead of constructing a histogram for $\tilde{S}(U)$ or $P(U)$. This can be done by transforming the energy integral into the sum of states.[4,38] Let us suppose that N samples (configurations or states) having the distribution $P(U)$ have been generated from the simulation. The distribution is the product of the density of states and the normalized weight of the generalized ensemble. Then, the density of states is estimated as $\Omega(U) = Z(\beta_0)P(U)e^{\tilde{S}(U)}$, where $Z(\beta_0)$ is the partition function for the generalized ensemble with the weight from Eq. (7). Next, by the definition, we have $Z_{\text{cano}}(\beta) = Z(\beta_0)/N\Sigma_s e^{\tilde{S}(U_s) - \beta U_s}$, in which the energy sum has been transformed to the state sum with the identity, $P(U) = (1/N)\Sigma_s \delta(U_s - U)$. Consequently, the canonical PDF is obtained as

$$P_{\text{cano}}(U;\beta) = \sum_s \frac{\delta(U - U_s)e^{\tilde{S}(U) - \beta U}}{\sum_s e^{\tilde{S}(U_s) - \beta U_s}}, \tag{17}$$

and the canonical ensemble average for the observable A obeys

$$A(\beta) = \sum_s \mathcal{P}(s;\beta)A(s), \tag{18}$$

where $\mathcal{P}(s;\beta) = e^{\mathcal{F}(U_s)}/\Sigma_s e^{\mathcal{F}(U_s)}$, where $\mathcal{F}(U_s) = \tilde{S}(U_s) - \beta U_s$ is the weighting function for state s and $A(s)$ is the observable value.

One potential difficulty with Eq. (18) is that the numerator and denominator can become very large or very small due to the huge ranges of U and $\tilde{S}(U)$, which can cause either numerical over-or under-flow in the computation. This problem can be avoided by assuming that $P_{\text{cano}}(U;\beta)$ is localized with a Gaussian shape around the fixed point U^*, obeying $\tilde{T}(U^*)=T$. Then, the summation in Eq. (18) can be done without any numerical difficulty by subtracting the maximum value of $\mathcal{F}^*=\tilde{S}(U^*)-\beta U^*$ from the exponents of both the numerator and the denominator,

$$A(\beta)=\sum_s \frac{e^{\mathcal{F}(U_s)-\mathcal{F}^*}}{\sum_s e^{\mathcal{F}(U_s)-\mathcal{F}^*}} A(s). \tag{19}$$

E. Detailed simulation protocols

Detailed simulation protocols for STMD are outlined as follows: (i) First determine the sampling range by selecting lower and upper temperature bounds of T_l and T_h, respectively, and the kinetic inverse temperature β_0; typically, $T_0=T_h$. Choose the energy bin size Δ, and the initial modification factor $f_d=f-1\ll 1$, with f_d measuring the deviation from the identity scaling. (ii) Perform the simulation with a standard integrator and thermostat set to β_0, supplemented by the scaling operations of Eq. (6) for the temperature estimate every time step, with initial guess $\tilde{T}(U)=T_h$, until a flat energy distribution is obtained. The flatness of histogram is determined by checking that the fluctuations are less than 20% of the average histogram $\bar{H}_i$ [the histogram is the non-normalized $P(U_i)$] as $|(H_i-\bar{H}_i)/\bar{H}_i|<0.2$. In the initial stage of the simulation, the low energy end of the temperature estimate is flattened at specified time intervals as $\tilde{T}(U)=T_{\min}$ for $U<U_{\min}$, where $T_{\min}=\tilde{T}(U_{\min})=\min\{\tilde{T}(U)\}$. This boundary flattening accelerates the convergence by allowing the system to access an unexplored energy region very quickly through the canonical sampling at $T_{\min}$. The initial simulation data are not taken into the accumulation of histogram H_i until $T_{\min}$ reaches T_l. (iii) Starting from the current estimate $\tilde{T}(U)$, repeat the same procedure, with a reduced convergence factor $\sqrt{f}$, to obtain again a flat histogram. The iteration is terminated when δf is sufficiently small, e.g., 10^{-8}. (iv) Calculate thermodynamic properties using the entropy estimate, Eq. (16), or using Eq. (18).

III. APPLICATION TO THE LENNARD-JONES FLUID

We first examined STMD in a Lennard-Jones (LJ) 110-particle system, at reduced density $\rho=0.88$ and with the potential cutoff at 2.5σ, from $T_l=0.7$ to $T_h=1.8$, corresponding to a fluid region. The LJ fluid has also been used for testing different versions of WL sampling.[14,15] Three STMD simulations with different energy bin sizes $\Delta=2$, 4, and 16 have been performed at $T_0=T_h$ starting from $f_d=0.000\,25$ and using the staircase temperature estimate of Eq. (11). As expected, the temperature estimate in the case of $\Delta=2$ in Fig. 1(a) is dynamically modified and extended to reach T_l at 3.9×10^6 MD steps, and f converges such that $f_d<10^{-8}$ after 1.4×10^7 MD steps. The corresponding energy sampling in Fig. 1(b) displays a typical random walk, sweeping the interesting energy region very frequently even with the vanishing f_d.

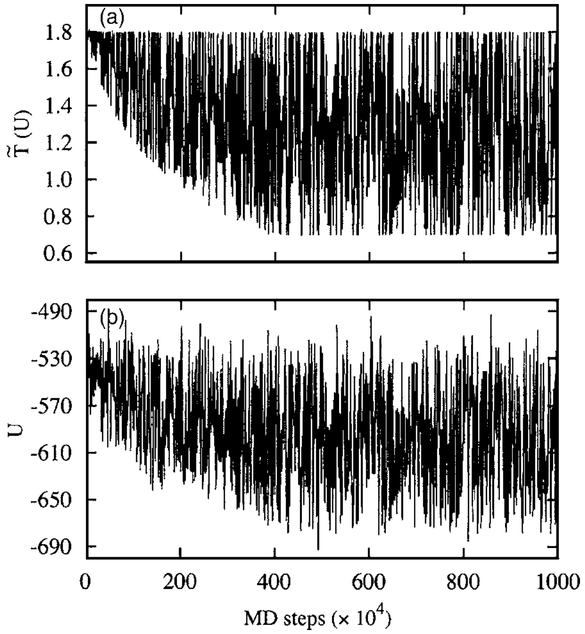

FIG. 1. (a) Temperature estimate $\tilde{T}(U)$ and (b) energy U as a function of MD steps for the initial stage of a STMD simulation of the 110-particle LJ fluid with $\Delta=2$ and initial $f_d=0.000\,25$.

During the initial stage of simulation for $T_{\min}<T_l$, the low energy end of the temperature estimate has been flattened every 2×10^5 MD steps as in Fig. 2(a) by enforcing $\tilde{T}(U)=T_{\min}$ for $U<U_{\min}$. The constant temperature estimate $\tilde{T}(U)=T_{\min}$ (recall $T_{\min}=\tilde{T}(U_{\min})=\min\{\tilde{T}(U)\}$) generates a canonical sampling at $T_{\min}$ for $U<U_{\min}$, and assists the system to access the low energy region very quickly through the extrapolation of the entropy estimate, $\tilde{S}(U)=\tilde{S}(U_{\min})+(U-U_{\min})/T_{\min}$, as in Fig. 2(b). The initial sampling speed slows down by 2.5 times without the low energy flattening. Note that the initial modification factor in STMD is very

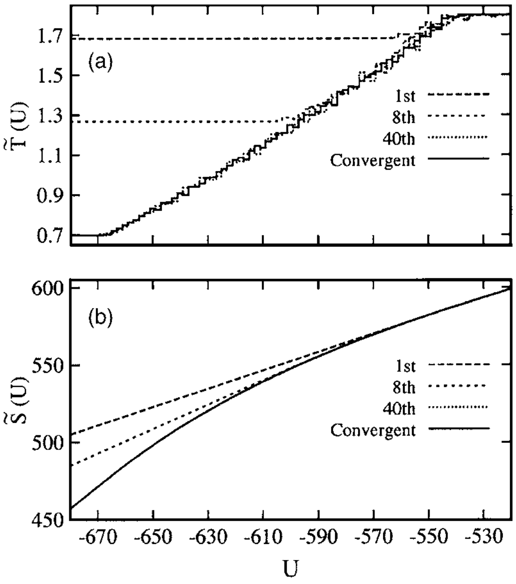

FIG. 2. (a) Temperature estimate $\tilde{T}(U)$ and (b) corresponding entropy estimate $\tilde{S}(U)$ as a function of the low energy flattening every 2×10^5 MD steps for 110-particle LJ fluid with $\Delta=2$ and initial $f_d=0.000\,25$.

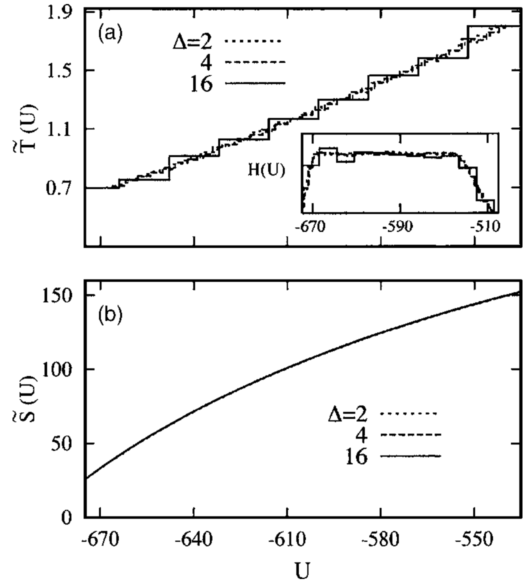

FIG. 3. (a) Convergent temperature estimates $\tilde{T}(U)$ and resulting energy histograms $H(U)$ (inset), and (b) reweighted entropy estimates for $\Delta=2$, 4, and 16 for 110-particle LJ fluid.

close to unity due to the restricted sampling range of $\tilde{T}(U)$, in contrast to WL sampling, which usually begins with $f=e$ to cover a large range of $\tilde{\Omega}(U)$. Accordingly, after the first iteration (4.4×10^6 MD steps), both $\tilde{T}(U)$ and $\tilde{S}(U)$ have almost reached their convergent values with $f_d<10^{-8}$ (1.4×10^7 MD steps) in Fig. 2(b).

When Δ increases to 16, the temperature estimate $\tilde{T}(U)$ in Fig. 3(a) shows a staircase modulation due to the discrete energy grid, which is directly reflected in the fluctuations of the energy histogram $H(U)$ in the inset of Fig. 3(a). However, the overall flatness of the histograms confirms that STMD is applicable with a large energy bin. Furthermore, the reweighting gives the same entropy estimate in Fig. 3(b) regardless of Δ. The energy PDF, $P(U)$, has been computed by collecting the simulation data of 3×10^6 MD steps with $f_d<10^{-6}$. The variation of the temperature estimates is less than 10^{-5} with this modification factor, so we assumed that the weight is fixed. The internal energy $U_{\text{ave}}(T)=\langle U\rangle_T$, with $\langle\cdots\rangle_T$ being the canonical ensemble average at T, and the heat capacity $C_{UU}(T)$ in Fig. 4 show good agreement with the canonical sampling results for 10^7 MD steps. The relative errors of the internal energy, i.e., $\epsilon=|(U_{\text{ave}}-U_{\text{cano}})/U_{\text{cano}}|$, are less than 0.0004.

The convergence of STMD is accelerated by two factors. One is the low energy flattening, which increases the initial sampling speed by allowing the system to access an unexplored energy region more rapidly through the canonical sampling. The other is the continuum description of the entropy estimate combined with an adjustable energy bin size. Since the flat histogram condition can be more easily achieved for a large Δ, the rate of convergence can be enhanced greatly without harming the statistical accuracy. We quantified the rate of convergence by plotting $\log f_d$ as a function of the number of MD steps. The flatness of the histogram has been checked every 10^5 MD steps. The time required for the first reduction of f has been shortened from 1.5×10^7 to 4.2×10^6 MD steps with the application of the flattening with the same $\Delta=2$ and $f_d=0.000\,25$ in Fig. 5(a). The effect of an enlarged Δ is also notable. By increasing Δ to 16, the rate of convergence is accelerated about 1.5 times compared to $\Delta=2$ with the same f_d.

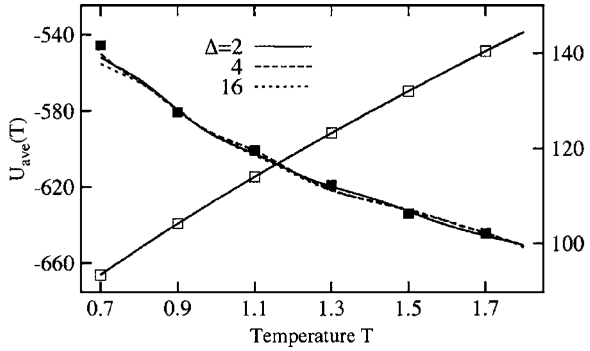

FIG. 4. Reweighted average energy $U_{\text{ave}}(T)$ and heat capacity $C_{UU}(T)$ for $\Delta=2$, 4, and 16 for 110-particle LJ fluid. For comparison, canonical ensemble results for 10^7 MD steps have been plotted at $T=0.7, 0.9, 1.1, 1.3, 1.5, 1.7$.

In the asymptotic limit of $f_d\to0$, where the dynamic modification of $\tilde{T}(U)$ is negligible, STMD reduces to the generalized ensemble sampling with the fixed weight $w(U)=\exp\{-\tilde{S}(U)\}$.[2,3] In this limit, the constant temperature estimate for each energy bin U_i produces a canonical sampling corresponding to the temperature $\tilde{T}_i$ and the resulting energy distribution is directly influenced by the staircase behavior of $\tilde{T}(U)$ for a large Δ. Thus the overall PDF is obtained as a superposition of the canonical ensemble samplings repre-

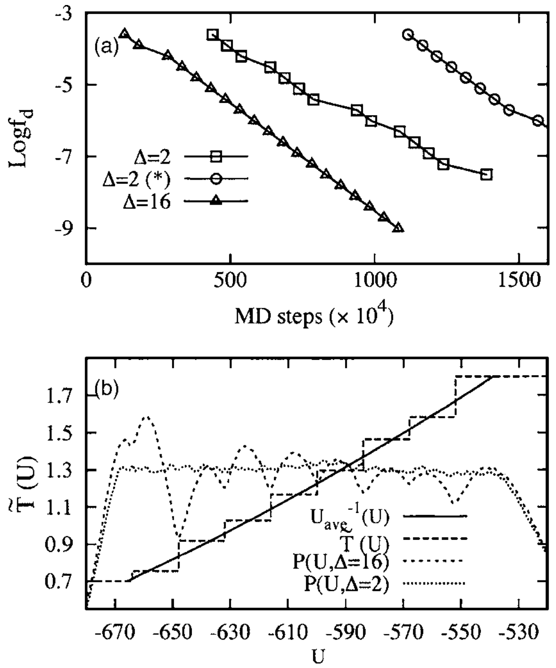

FIG. 5. (a) $\log f_d$ as a function of MD steps for various simulation conditions with the same $f_d=0.000\,25$, and (b) temperature estimate $\tilde{T}(U)$ and distributions $P(U)$ with $\Delta=2$ and 16 for 110-particle LJ fluid. The asterisk in (a) denotes the STMD simulation without low energy flattening. The exact statistical temperature $T(U)=U_{\text{ave}}^{-1}(U)$ is provided for comparison.

sented by Gaussians centered at stationary points of $\delta\beta(U_i^*)=0$, i.e., $T(U_i^*)=\tilde{T}_i$. Indeed, the energy distribution on finer grids in the case of $\Delta=16$ in Fig. 5(b) shows clear structures and a characteristic correspondence with the variation of

$$\delta T(U) = \tilde{T}(U) - T(U) = \tilde{T}(U) - U_{\text{ave}}^{-1}(U), \qquad (20)$$

where we identified the exact statistical temperature $T(U)$ by $U_{\text{ave}}^{-1}(U)$ obtained by inverting the functional relationship $U_{\text{ave}}(T)=U$ from the equivalence of microcanonical and canonical ensembles.[26]

The stationary points of $P(U)$ correspond to the crossing points of $\tilde{T}(U)$ and $T(U)$, satisfying $\delta T(U_i^*)=0$, which are also the zeros of the deterministic force $\delta\beta(U)$ in Eq. (10). The deterministic force simplifies to $-\delta T(U)/T_i^{*2}$ around the stationary points U_i^* and creates a biased probability current to the energy-decreasing (-increasing) direction for $\delta T(U) > 0$ ($\delta T(U) < 0$), which is the dynamical origin of the formation of local maxima and minima in $P(U)$ at stable and unstable zeros corresponding to $\kappa(U_i^*)=(\partial T/\partial U)/(\partial\tilde{T}/\partial U)|_{U_i^*} < 1$ and >1, respectively.[36] On the other hand, in the case of $\Delta=2$ in Fig. 5(b), the effect of the discrete energy grid on $\tilde{T}(U)$ is negligible and the sampling dynamics produces an almost uniform distribution by creating more densely distributed fixed points U_i^*.

IV. APPLICATION TO COARSE-GRAINED PROTEIN MODELS

Next, we consider the more challenging problem of protein folding. Despite recent advances,[39] simulations describing both the protein and the solvent remain computationally very demanding. Thus a simplified description, incorporating the main features of real proteins, but reducing the number of degrees of freedom significantly through a coarse graining, remains quite useful. In this study, we have chosen the off-lattice Honeycutt-Thirumalai β-barrel *BLN* model,[28] denoted by βBLN, as a testing system for STMD, since this model has been extensively studied and provides a good example of a rugged energy landscape, which cannot be correctly sampled by conventional MC or MD simulations. The primary sequence is composed of three types of beads, hydrophobic (B), hydrophilic (L), and neutral (N), and the potential energy is obtained by summing harmonic bond-stretching and bond-angle terms, torsion-dihedral potentials, and nonbonded interactions. The former three interactions determine a local, secondary structure, and the nonbonded interactions determine an overall tertiary structure. We used the same potential form and parameter set as reference.[40] Dimensionless length, energy, and temperature are defined through the collision diameter σ and the well depth parameter ϵ of the nonbonded attractions.

A. β-barrel 46-mer

The primary sequence of the βBLN 46-mer is $B_9N_3(LB)_4N_3B_9N_3(LB)_5L$ with a four-stranded β-barrel global energy minimum. It has been intensively studied[28,32,41–47] in terms of its kinetics, thermodynamics, and potential energy landscape. Due to the presence of a high degree of energetic frustration, the 46-mer has been also used as a benchmark to test various global optimization algorithms.[40,48] More recently, replica exchange MC combined with principal component analysis has been applied to explore the local structural diversity of this model protein.[49]

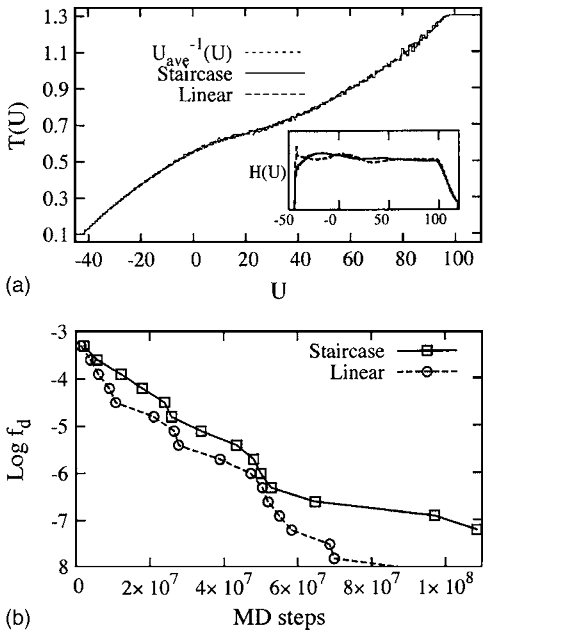

FIG. 6. (a) Convergent temperature estimates $\tilde{T}(U)$ and inverse average energy $U_{\text{ave}}^{-1}(U)$, and corresponding energy histograms $H(U)$ (inset), and (b) $\log f_d$ as a function of MD steps for both staircase and linear interpolation schemes for βBLN 46-mer.

We performed two STMD simulations using the different interpolation schemes of Eqs. (11) and (13) for the temperature range of $T_l=0.1$ to $T_h=1.3$ at $T_0=T_h$ with the initial modification factor $f=1.0005$ and the bin size $\Delta=1$. The temperature estimates in Fig. 6(a) converge to $f_d < 10^{-7}$ after 10^8 and 5.8×10^7 MD steps, for the cases of staircase and linear interpolations, respectively, realizing the flat energy histograms in the inset of Fig. 6. Irrespective of the interpolation scheme, the convergent temperature estimates are quite similar and show good agreement with the inverse average energy, $U_{\text{ave}}^{-1}(U)$, corresponding to the true $T(U)$. The plot of $\log f_d$ as a function of MD steps in Fig. 6(b) reveals that the linear interpolation scheme is more efficient in achieving a flat histogram, through a more faithful representation of the statistical temperature.

The energy sampling for the case of linear $\tilde{T}(U)$ in Fig. 7(a) displays a typical random walk, with very frequent sweeps of the entire energy range, even with a vanishing modification factor $f_d < 10^{-7}$. The energy trajectory shows two separate sampling domains, corresponding to high energy extended states and low energy folded or collapsed states. STMD yields a broad sampling, even for very low temperatures down to $T=0.1$, while most previous[42,43] studies have been restricted to $T \gtrsim 0.3$ due to glassy dynamics in the rough energy landscape, except for entropic tempering[50] and replica exchange Monte Carlo.[49]

The power of STMD may be further illustrated by examining the inherent structures (IS). In Fig. 7(a), 76 450

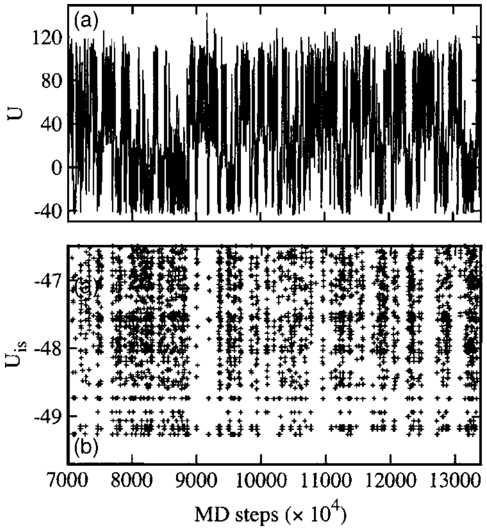

FIG. 7. (a) Energy trajectory beyond 7×10^7 MD steps with $f_d<10^{-7}$ for the linear interpolation scheme, and (b) corresponding inherent structure (IS) plot for βBLN 46-mer.

equilibrium configurations, saved every 10^3 MD steps, have been mapped to 45 616 IS using conjugate-gradient minimization until the energy variation becomes less than 10^{-5}. The global minimum, or ground state, has been exactly located at $U_0=-49.2635$, which is consistent with a recent conformational space annealing (CSA) study.[40] The IS plot shows vigorous transitions between different low-lying structures, verifying that our simulation reproduces low energy folded states correctly, overcoming a quasiergodicity at low temperature. To check the performance of STMD more quantitatively, we compared the number of different IS found with energy less than $\Delta U+U_0$ with the results of CSA (Ref. 37) in Table I. Compared to these sophisticated optimization results, STMD finds more local minima as the energy increases from the ground state U_0.

TABLE I. Number of IS less than $U_0+\Delta U$ for βBLN 46-mer and 69-mer in STMD simulations, conformational space annealing (CAS) (Ref. 48), and automated histogram filtering (AHF) (Ref. 29).

	46-mer		69-mer	
ΔU	STMD	CAS	STMD	AHF
1	5	5	3	3
2	40	36	44	47
3	189	147	205	175
4	498	339	588	486
5	1045	636	1389	935
6	1805	1010	2457	1542
7	2723	1387	3596	2237

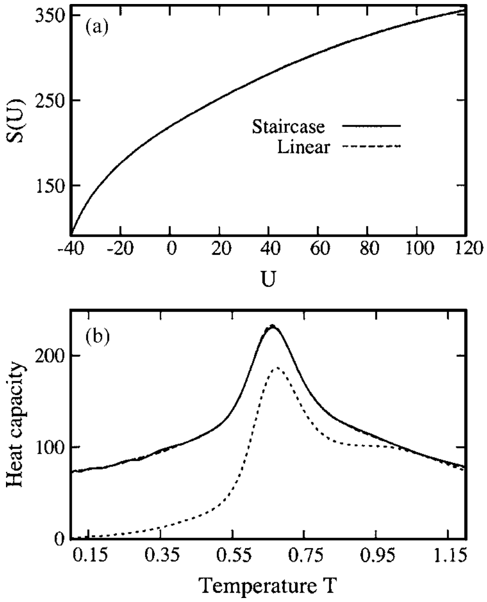

FIG. 8. (a) Entropy estimates of stepwise and linear interpolation scheme for $\tilde{T}(U)$, and (b) heat capacities C_{UU}^{all} (dashed line) determined by reweighting for 3.2×10^7 energy data and C_{UU}^{par} (Solid line) and $C_{U_{is}U_{is}}^{\mathrm{par}}$ (dotted line) determined by reweighting for 764 50 selected equilibrium and IS configurations, respectively, for βBLN 46-mer.

Previous studies[41,42,51,52] have shown that the foldability of proteins is mainly determined by two transitions. One is the collapse transition from random coil states to collapsed, but non-native, states at $T_\theta\approx0.65$ in βBLN 46-mer.[42,50] This transition is identified by the peak in the energy fluctuations, i.e., the heat capacity $C_{UU}=\langle(\delta U)^2\rangle/k_BT^2$. Here we defined $C_{XY}(T)=\langle\delta X\delta Y\rangle/k_BT^2$, $\delta X=X-\langle X\rangle$. The other is the folding transition from collapsed states to the native state originally estimated at $T_f\approx0.35$,[42] which is indicated by a peak in the order parameter fluctuations, $\Sigma_{QQ}=\langle Q^2\rangle-\langle Q\rangle^2$. The order parameter, Q,[42,51] measuring the structural similarity of a configuration to the ground state or global energy minimum, is

$$Q=\frac{1}{M}\sum_{i,j>i+4}^{N}\theta(\epsilon-|r_{ij}-r_{ij}^0|), \tag{21}$$

where r_{ij} and r_{ij}^0 are the relative distances between beads i and j in the instantaneous configuration and the ground state, respectively. Here M is the normalization constant and $\epsilon=0.2$ is the threshold value to take into account thermal fluctuations around the ground state.

The microscopic definition of the native state, which corresponds to a specific three-dimensional structure performing its biological function, is nontrivial in the potential energy landscape. It is not the minimum energy configuration, since thermal fluctuations always exist. The IS formalism automatically groups configurations which drain to the same IS, i.e., those that belong to the same basin of attraction, so it seems clear that configurations belonging to IS_0 are representative of the native state. It must also be considered that multiple IS, notably those near the bottom of the "folding funnel," can share the structural motif of the folded state.

One advantage of STMD is that, once we determine the entropy estimate $\tilde{S}(U)$, any thermodynamic quantities can be calculated at an arbitrary temperature using the reweighting. Figure 8(a) shows the reweighted entropy estimates of both interpolation schemes, obtained by collecting simulation data with $f_d<10^{-7}$, and they are indistinguishable. As expected, the collapse transition, signaled by the slope variation of the convergent $\tilde{T}(U)$ around $U\approx20$ in Fig. 6(a), is exactly associated with the peak at T_θ in the heat capacity C_{UU}^{all} in Fig. 8(b), which has been determined by the reweighting of 3.2

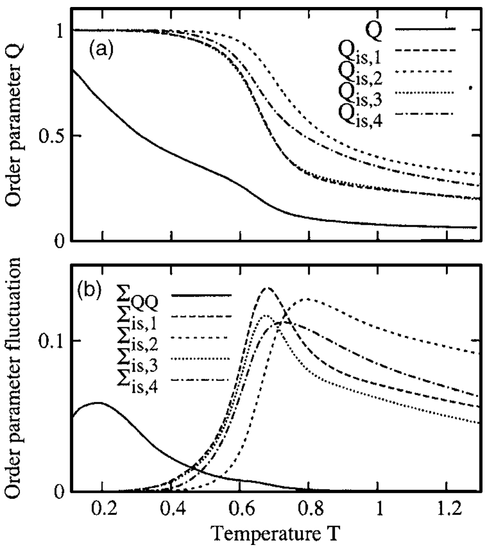

FIG. 9. (a) Average order parameters $Q(T)$ and $Q_{is,i}(T)$ for ith beta strand, and (b) order parameter fluctuations $\Sigma_{QQ}(T)$ and $\Sigma_{Q_{is,i}Q_{is,i}} \equiv \Sigma_{is,i}(T)$ for ith beta strand for βBLN 46-mer.

$\times 10^7$ simulated energy data. We also calculated a partial average, C_{UU}^{par}, using the energy data of 76 450 selected equilibrium configurations. In contrast to the smooth behavior of C_{UU}^{all} the partial heat capacity C_{UU}^{par} shows a small ruggedness at low temperatures, but the overall coincidence for the whole temperature region confirms that the selected date set is well equilibrated.

The transitions occurring in protein folding can also be seen in inherent structure thermodynamics quantities, e.g., through the IS energy fluctuation heat capacity, $C_{U_{is}U_{is}}$, in Fig. 8. Since the mapping from equilibrium configurations to local minima eliminates irrelevant thermal fluctuations, the thermodynamic signature of the collapse transition is more clearly demonstrated in $C_{U_{is}U_{is}}^{\text{par}}$ rather than in C_{UU}^{all} at the same T_θ. In addition to the main peak around T_θ, a small shoulder is observed in $C_{U_{is}U_{is}}$, but is smoothed out in C_{UU}^{par}, at $T \approx 0.95$. The shoulder is associated with an early formation of secondary structures of the second and fourth β strands, which is identified in the partial IS order parameter $Q_{is,i}$ and its fluctuations, $\Sigma_{Q_{is,i}Q_{is,i}}$, in Figs. 9(a) and 9(b), respectively, for the ith β strand comprised of consecutive beads specified in Table II.

In contrast to the sharp collapse transition consistent with the previous study,[42] the folding transition is not clearly seen. The total order parameter Q in Fig. 9(b) is monotonically increasing crossing T_θ and T_f, with no saturation behavior indicating a dominant occupation of the native state. The order parameter fluctuations Σ_{QQ} are also monotonically increasing, with a small shoulder at T_θ and a very broad peak at $T \approx 0.2$, far below the previously determined T_f. The folding transition has also been identified by thermal features such as a shoulder or peak in the heat capacity.[42] However, Fig. 8(b) shows that the thermodynamic signature of the folding transition is almost entirely suppressed in C_{UU}^{all} and is only detected as a small shoulder in $\partial C_{UU}^{\text{all}} / \partial T$ (not shown). Recently, the folding temperatures have been reassigned to ≈ 0.27 in a replica exchange Monte Carlo study,[49] based on the heat capacity and the order parameter fluctuations. Indeed, we find that the heat capacity C_{UU}^{par} determined by a smaller data set shows similar behavior at low temperature, as in the magnified view of Fig. 10. However, one must take care with a procedure that amounts to deliberately using incomplete sampling. The smooth variation of the heat capacity at low temperatures and the suppression of the folding signature have also been observed with simulated tempering[53] for the same system with a smaller bond-stretching constant.[31]

TABLE II. Classification of sequences of secondary structures in both 46-mer and 69-mer.

β strand	1st	2nd	3rd	4th	5th	6th
46-mer	1–9	13–20	24–32	36–46		
69-mer	1–9	13–20	24–32	36–43	47–55	59–69

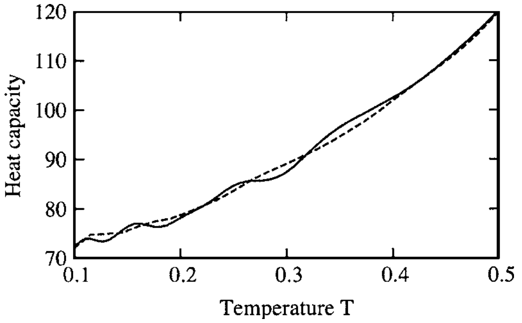

FIG. 10. Magnified view of heat capacities: C_{UU}^{all} (dashed line) determined by reweighting of 3.2×10^7 simulation data and C_{UU}^{par} (solid line) determined by reweighting of 76 450 selected equilibrium configurations for βBLN 46-mer.

With a database of IS in hand, a scatterplot in the order parameters (U_{is}, Q_{is}) [Fig. 11(a)] clearly reveals the multifunnel structure of the landscape. We propose the scatterplot as an alternative to the disconnectivity analysis.[46,47,54] Below $U_{is} < -45$ there exists an apparent energy barrier separating the folding funnel leading to the global minimum, in which conformations with $Q_{is} > 0.7$ have been grouped together, from misfolded conformations with $Q_{is} < 0.65$, which are more easily accessible from the main branch of collapsed states with $U_{is} > -30$. In our view, the low-lying IS in the folding funnel constitute the native state.

Low-lying non-native IS are found at termini of the misfolding funnel. They differ from the native state primarily through hydrophobic mismatches between β strands due to variations in the loop regions. Folding means a dominant occupation of the native state, but the order parameter distribution $P(Q,T)$ in Fig. 11(b) reveals that there are still significant populations for non-native IS even at very low $T = 0.11$. In addition to the peak at $Q \approx 1$, corresponding to the native state occupation, $P(Q,T)$ shows several sharp peaks at $Q \approx 0.3$, 0.43, and 0.65, at both $T = 0.11$ and 0.17, corresponding to the low-lying non-native IS seen in Fig. 11(a). Furthermore, the largest occupation probabilities after collapse lie in the misfolding funnel, ranging from $Q \approx 0.2$ to 0.6, and population of the folding funnel begins below $T \approx 0.3$.

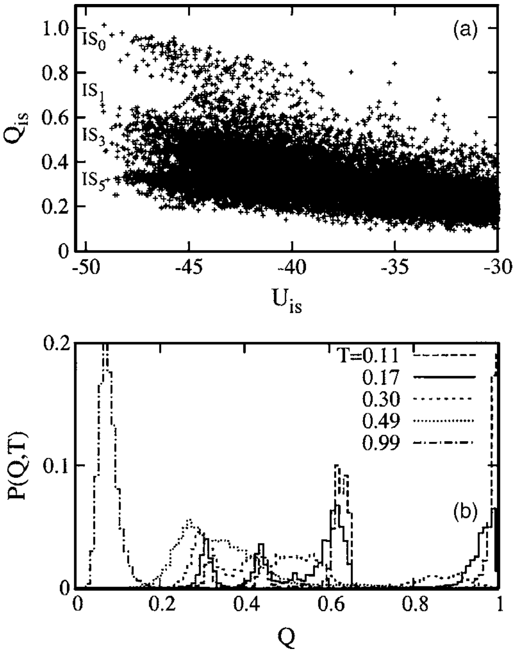

FIG. 11. (a) Multifunnel energy landscape visualized by IS scatterplot in (U_{is},Q_{is}), and (b) reweighted order parameter distributions $P(Q,T)$ at several temperatures.

The order parameter Q and the potential energy play central roles in characterizing the protein folding process. However, we have just seen that, in a multifunnel landscape, averaging over substantially occupied, structurally dissimilar, low-lying IS reduces the folding signatures significantly[31] in $Q(T)$, Σ_{QQ}, and C_{UU}^{all}. The averaging will not occur unless the algorithm employed can overcome quasiergodicity. Our prior work showed[24,31] that, below the collapse temperature T_θ, thermodynamics is dominated by a finite number of low-lying IS. Then, with the usual indicators not working well, the occupation probabilities for individual IS can be particularly useful to elucidate a folding transition. To pursue these ideas we have calculated the canonical average occupation probabilities $p_i(T)$ by reweighting,

$$p_i(T) = \sum_s \mathcal{P}(s;\beta)\,\delta_{\mathrm{IS}}(s-i), \tag{22}$$

where s is the index for configurations taken from the STMD trajectory and $\delta_{\mathrm{IS}}(s-i)$ is nonvanishing only when configuration s belongs to the basin attraction of IS. A somewhat related analysis based on macroscopic thermodynamic states determined by top-down free-energy minimization has been used to find hierarchical properties of the energy landscape in a small peptide.[55]

The results are shown in Fig. 12. The main observations are that (i) the system begins to occupy low-lying IS with a finite probability as T falls below the collapse temperature T_θ; (ii) the ground state occupation p_0 is less than p_1 and p_2 over a "misfolding interval" extending from T_θ down to a new characteristic temperature which we have denoted[31] by $T_{p_0}\approx 0.34$, such that p_0 is the largest individual p_i for $T<T_{p_0}$; and (iii) there are still non-negligible populations for excited states down to $T\approx 0.1$.

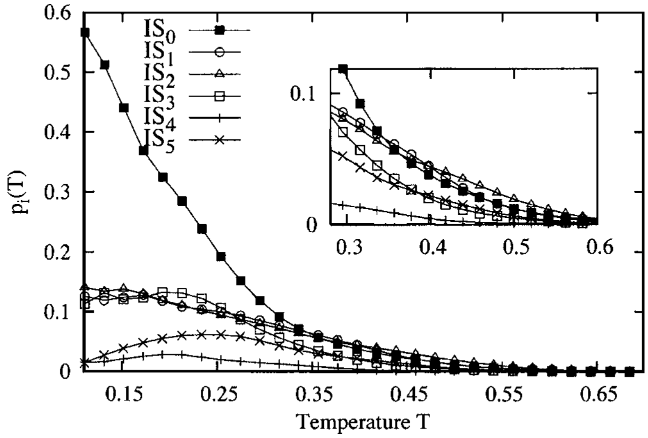

FIG. 12. Average occupation probabilities $p_i(T)$ $(i=0-5)$ for low-lying IS as a function of the temperature and magnified view around the collapse temperature $T_\theta\approx 0.65$ (inset) for βBLN 46-mer.

The analysis of $p_i(T)$ confirms our expectation of blurred folding signatures, i.e., the monotonic increase of $Q(T)$ and the nonexistence of a peak in Σ_{QQ} crossing T_f. Furthermore, the dominance of p_i $(i=1,2)$ over p_0 below the collapse transition explains why global optimization of the 46-mer so often fails, leading to non-native or misfolded states. Indeed, IS_1 and IS_2 belong to "megabasins"[1] associated with the misfolding funnel, which are more frequently visited in simulated annealing than the native state.[40]

The slowing down of the folding process by trapping in non-native IS is strongly correlated with ergodicity breaking of the system. We found that conventional canonical MD cannot correctly sample the configurational space around and below T_f and shows an initial condition dependence due to the ruggedness of the PEL. In Fig. 13(a) and 13(b), we plot the evolution of the IS of two independent canonical trajectories starting from initial configurations belonging to the ground $(\mathrm{IS}_0=-49.2635)$ and second excited states $(\mathrm{IS}_2=-49.1488)$ at temperatures $T=0.3$ and 0.4, respectively. At $T=0.3$, the trajectory starting from IS_0 samples a restricted

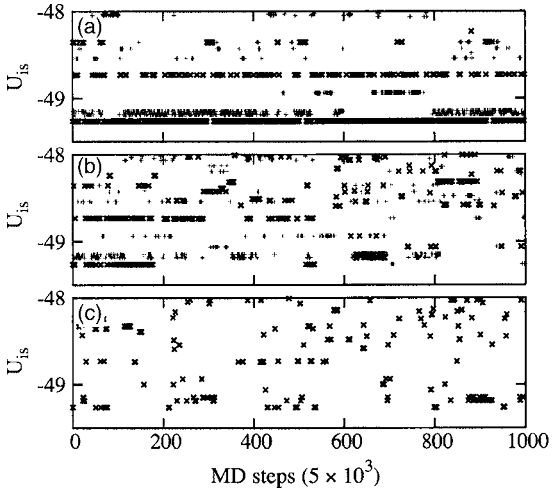

FIG. 13. IS plots for canonical MD starting from different initial configurations of the ground state IS_0 (black) and second excited state IS_2 (gray) at (a) $T=0.3$ and (b) $T=0.4$, and (c) canonical MD starting from IS_2 at $T=0.5$.

megabasin only, $B_{\rm fold}=\{{\rm IS}_0,{\rm IS}_8,{\rm IS}_{22},\ldots\}$, belonging to the folding funnel with $Q_{is}>0.7$. On the other hand, the trajectory starting from $\rm IS_2$ stays within another megabasin associated with the unfolding funnel with $Q_{is}<0.7$, $B_{\rm unfold}=\{{\rm IS}_1,{\rm IS}_2,{\rm IS}_{11},{\rm IS}_{17},\ldots\}$, during 5×10^6 MD steps. At the elevated temperature $T=0.4$, both trajectories show a short transient trapping and infrequent visits to the other megabasin. As the temperature increases to $T=0.5$ in Fig. 13(c), the dynamics shows more frequent IS transitions crossing a strong barrier between $B_{\rm fold}$ and $B_{\rm unfold}$. This implies a broken ergodicity, with preferred accessibility to the non-native $\rm IS_1$ and $\rm IS_2$, leading to a substantial slowing down of folding below $T<0.4$.

B. β-barrel 69-mer

Recently, Rothstein and co-workers[29,56] developed the automated histogram filtering method (AHF), which generates low energy states by combining a hierarchical clustering and repeated simulated annealing. The performance of AHF was tested[29,56] for the βBLN 69-mer, an extended version of the 46-mer, with two additional beta strands. The primary sequence is $B_9N_3(LB)_4N_3B_9N_3(LB)_4N_3B_9N_3(LB)_5L$, for a six-stranded β barrel global minimum. Since energetic frustration in the βBLN model arises primarily from conformal diversity generated by hydrophobic mismatches of β strands and variations in the loop regions, the increased complexity of the 69-mer presents a more stringent test for STMD. The structural diversity and the thermodynamics of the 69-mer have also been studied in a massive replica exchange Monte Carlo simulation using 97 replicas ranging from $T=0.14$ to 2.0 with 5×10^7 MC sweeps for each replica.[49]

Due to an increased energy range, we used a relatively large energy bin, $\Delta=2$, for the same temperature range used for the 46-mer, from $T_l=0.1$ to $T_h=1.3$, with $T_0=1.3$ and $f=1.000\,25$. The linear interpolation scheme for $\tilde{T}(U)$ was applied to accelerate the convergence. The temperature estimate $\tilde{T}(U)$ in Fig. 14(a) is in good agreement with the inverse average energy $U_{\rm ave}^{-1}(U)$ with the simulation converged to $f_d<10^{-6}$ after 2.1×10^8 MD steps [Fig. 14(b)], and the flat energy histogram is shown in the inset of Fig. 14(a). The corresponding energy sampling in Fig. 15(a) displays several folding-unfolding transitions during 5×10^7 MD steps with a vanishing modification factor $f_d=10^{-8}$.

The subsequent IS analysis in Fig. 15(b) locates the global minimum of the 69-mer at −99.189, in agreement with the AHF study,[29] and vigorous IS transitions confirm that STMD is very promising for sampling low-lying states, even in a more complicated PEL. We obtained 125 030 IS by minimizing 171 190 equilibrium configurations saved every 10^3 MD steps, subject to $f_d<10^{-6}$. The structures of low-lying IS differ only by the relative orientation of hydrophobic strands and loop regions. As with the 46-mer, Table I demonstrates that STMD finds more low energy IS than AHF as the energy increases from the ground state. With a rougher PEL, the 69-mer also has a larger number of local minima[30] than the 46-mer.

The thermodynamics of folding is quite similar to the case of the 46-mer. The collapse transition is identified by the main peak in $C_{UU}^{\rm all}$ in Fig. 16(a) at $T\approx0.71$, which is obtained by applying the reweighting to 1.2×10^8 energy data. For comparison we also calculated the partial heat capacities $C_{UU}^{\rm par}$ and $C_{U_{is}U_{is}}^{\rm par}$ with the energy data of selected 171 190 equilibrium and IS configurations, respectively. As observed in the 46-mer, the partial heat capacity, $C_{UU}^{\rm par}$, shows several weak thermodynamic anomalies at low temperatures, which have been ascribed to the glass and folding transition temperatures.[49] However, again, these thermal features are strongly suppressed by increasing the number of simulation data in the reweighting, as in the magnified view of Fig. 16(b). The overall coincidence of $C_{UU}^{\rm all}$ and $C_{UU}^{\rm par}$ at high temperatures implies that the discrepancies at low temperatures should be attributed to the statistics of the sampling data set.

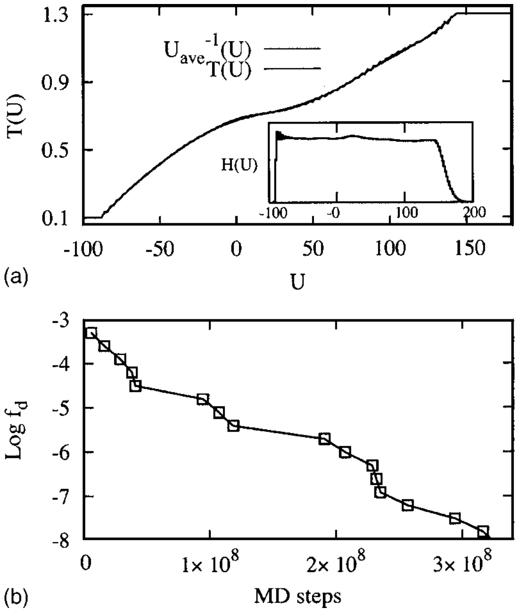

FIG. 14. (a) Convergent temperature estimates $\tilde{T}(U)$ using the linear interpolation scheme and reweighted inverse average energy $U_{\rm ave}^{-1}(U)$ and resulting histograms $H(U)$ (inset), and (b) $\log f_d$ as a function of MD steps for βBLN 46-mer.

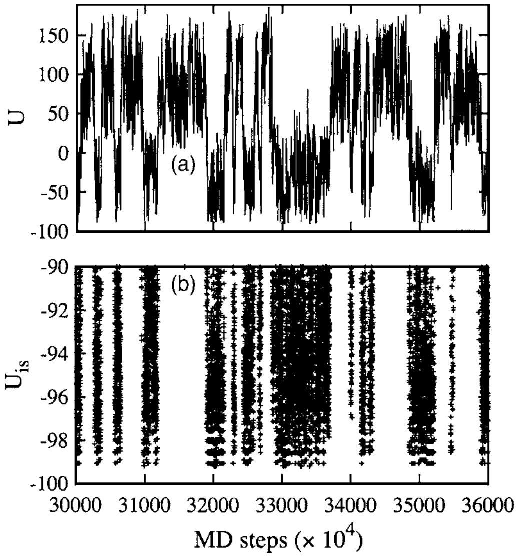

FIG. 15. (a) Energy trajectory and (b) corresponding IS plot beyond 3×10^8 MD steps, with $f_d<10^{-7}$, for βBLN 69-mer.

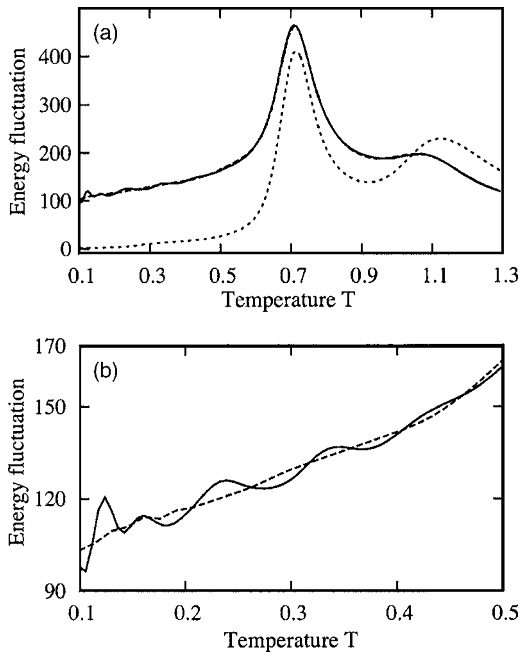

FIG. 16. (a) Specific heats C_{UU}^{all} (dashed line) determined by reweighting for 1.2×10^8 energy data and C_{UU}^{par} (solid line) and $C_{U_{is}U_{is}}^{\mathrm{par}}$ (dotted line) determined by reweighting for the energy data of 171 190 selected equilibrium and corresponding IS configurations, respectively, and (b) magnified view of C_{UU}^{par} (solid line) and $C_{U_{is}U_{is}}^{\mathrm{par}}$ (dashed line) for βBLN 69-mer.

Before the collapse a broad transition is signaled by a peak in $C_{U_{is}U_{is}}^{\mathrm{par}}$ or a shoulder in C_{UU}^{all} at $T\approx1.0$. The partial order parameter $Q_{is,i}$ and the fluctuations $\Sigma_{is,i}$ in Figs. 17(a) and 17(b) reveal that this transition is mainly due to local ordering of the second, fourth, and sixth strands. Interestingly, there exists a strong correlation between the local-ordering transition and the native state topology. In both the 46-mer and the 69-mer, the advanced formation of even-numbered strands corresponding to outer parts of native β barrels is mainly associated with the local-ordering transition with a broad peak in the heat capacity. On the other hand, the secondary structures of odd-numbered strands forming inner contacts in native barrels are established through the collapse transition.

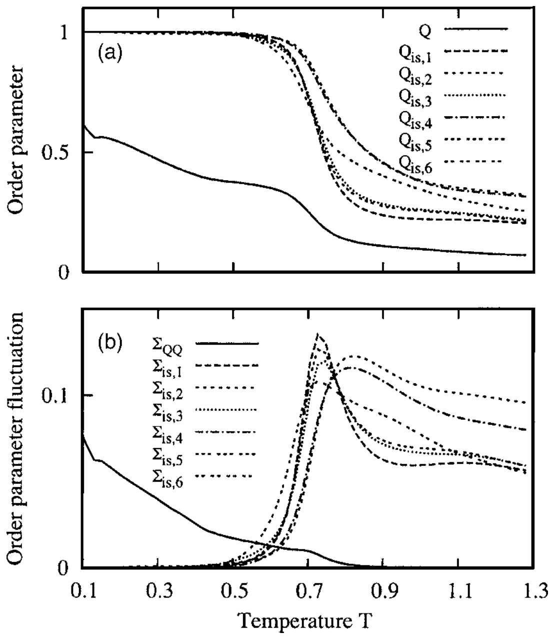

FIG. 17. (a) Average order parameters $Q(T)$ and $Q_{is,i}(T)$ for ith beta strand, and (b) order parameter fluctuations $\Sigma_{QQ}(T)$ and $\Sigma_{is,i}(T)$ for ith beta strand for βBLN 69-mer.

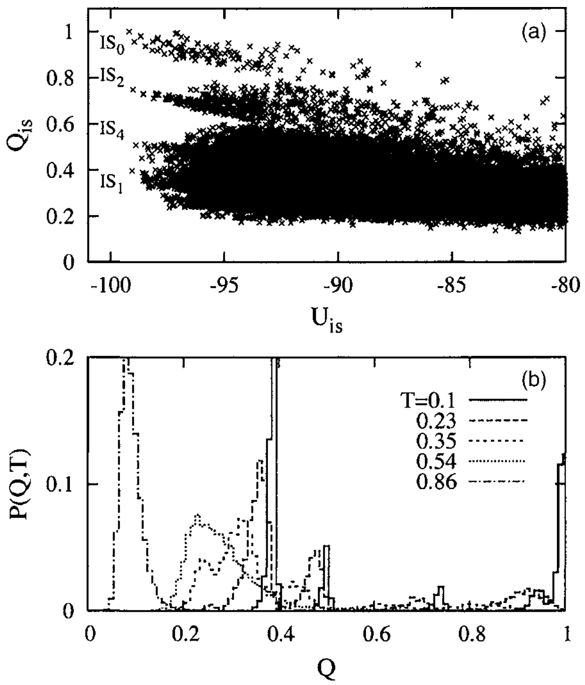

FIG. 18. (a) Multifunneled energy landscape characterized by scatterplot in (U_{is},Q_{is}) and (b) reweighted order parameter distribution $P(Q,T)$ at several temperatures for βBLN 69-mer. In (a), IS_0, IS_1, IS_2, and IS_4 correspond to the native, first excited, second excited, and fourth excited IS states, respectively.

As with the 46-mer, the thermodynamic signature associated with folding is not clearly seen in C_{UU}^{all} or $C_{U_{is}U_{is}}^{\mathrm{par}}$ in Fig. 16(a). Only a small peak is observed in $\partial C_{UU}^{\mathrm{all}}/\partial T$ and $\partial C_{U_{is}U_{is}}^{\mathrm{par}}/\partial T$ at $T\approx0.28$. Furthermore (Fig. 17), the average order parameter $Q(T)$ is monotonically increasing and the order parameter fluctuations Σ_{QQ} do not show any folding signature down to $T=0.1$, except for a small shoulder around T_θ. This is even more monotone behavior than in the 46-mer, which shows a broad peak around $T\sim0.23$ in Σ_{QQ}.

Again, the smoothing and suppression of the folding signature in our enhanced sampling are due to the contribution of low-lying, non-native IS at low temperatures. The two-dimensional scatterplot [Fig. 18(a)] in (U_{is},Q_{is}) clearly demonstrates the extraordinary complexity of the 69-mer PEL, with several misfolding funnels competing with the folding funnel below $U_{is}<-93$. Below typical energies for collapse, the landscape bifurcates into several branches leading to dissimilar IS at the bottom of each branch. The nature of averaging in a multifunneled PEL is shown by the order parameter distributions $P(Q,T)$ in Fig. 18(b). The peak in $P(Q,T)$ is localized around $Q\approx0.1$ above the collapse temperature and broadens, with a shift to $Q\approx0.3$, crossing T_θ. For $T<0.42$ the distribution spreads out over several misfolding funnels and the folding funnel. The positions of the peaks of $P(Q,0.1)$ correspond to the IS at the bottoms of the multiple

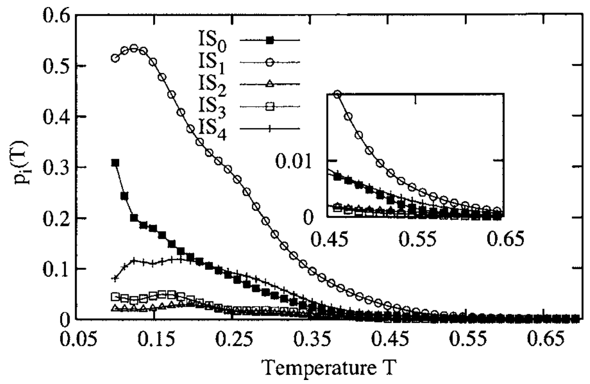

FIG. 19. Individual occupation probabilities $p_i(T)$ $(i=0-4)$ for low-lying IS as a function of temperature, and magnified view (inset) below the collapse temperature $T_\theta \approx 0.71$ for *βBLN* 69-mer.

funnels in Fig. 18(a), implying that the contribution of the non-native IS to the thermodynamics is still non-negligible at $T=0.1$.

The average occupation probabilities in Fig. 19 reveal how the folding of the 69-mer is interrupted and slowed by the presence of non-native IS. Crossing T_θ, $p_1(T)$ rises rapidly above p_0 and begins to decrease below $T \approx 0.15$; p_4 also exceeds p_0 from T_θ to $T \approx 0.25$. Remarkably, the native state occupation p_0 is still lower than p_1 down to $T=0.1$. This is in sharp contrast to the case of the 46-mer, in which p_0 surpasses p_1 at an intermediate temperature $T_{p_0}=0.34$. The persistent dominance of p_1 to p_0 down to $T=0.1$ should be attributed to the increased complexity of the PEL of the 69-mer; in this case all we can say about T_{p_0} is that it is <0.1. Both IS$_1$ and IS$_4$ correspond to termini of major misfolding funnels in Fig. 18(a), explaining why Σ_{QQ} does not show any folding signature down to $T=0.1$. The existence of such a large misfolding interval will cause a substantial slowing down of folding through a kinetic trapping in misfolded states.

V. CONCLUSIONS

In summary, our recently proposed statistical temperature molecular dynamics algorithm (STMD) has been applied to the 110-particle Lennard-Jones fluid and *βBLN* 46-mer and 69-mer protein models to examine its performance and applicability to biomolecular simulations. The tests on the LJ fluid confirm that STMD works very well even with large energy bin sizes, with a considerable acceleration of the rate of convergence, while maintaining statistical accuracy.

The enhanced sampling of STMD combined with extensive IS analysis explicitly verifies a multifunnel structure of the potential energy landscapes of the *βBLN* 46-mer and 69-mer through two-dimensional scatterplots of the IS in (U_{is}, Q_{is}), and reveals various characteristics of protein folding in terms of equilibrium and IS thermodynamic quantities. The scatterplot provides a particularly clear picture of the energy landscape and is proposed as an alternative to disconnectivity analysis.

Below the collapse temperature, the protein thermodynamics is dominated by the contributions of a finite number of low-lying IS and its estimation is affected by the efficiency of sampling due to the energetic and entropic barriers between the folding and misfolding funnels. Thermodynamic folding signatures of both the 46-mer and the 69-mer are significantly suppressed by taking into account the contributions of non-native IS. This means that the conventional thermodynamic signatures, i.e., the heat capacity or the order parameter fluctuations, might not be applicable to characterize protein folding in multifunneled energy landscapes. We recently proposed[31] the configurational entropy fluctuations as a thermodynamic indicator of folding under such circumstances.

The analysis of the average occupation probability, $p_i(T)$, for individual IS demonstrates that there exists a "misfolding interval," $T_\theta \geqslant T \geqslant T_{p_0}$, in which the occupation for non-native IS exceeds the native state occupation, and this interval is strongly correlated with the frustration or roughness of potential energy landscape. We believe that the existence and details of the misfolding interval are intimately connected to the intermediate and poor foldabilities, respectively, of the *β*-barrel-forming 46-mer and 69-mer through a substantial slowing down of folding with a long-lived kinetic trapping in misfolded states.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Nos. CHE-0312959 (ITR) and CHE-0352026.

[1] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[2] B. J. Berne and J. E. Straub, Curr. Opin. Struct. Biol. **7**, 181 (1997).
[3] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[4] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[5] B. A. Berg and T. Celik, Phys. Rev. Lett. **69**, 2292 (1992).
[6] J. Lee, Phys. Rev. Lett. **71**, 211 (1993).
[7] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[8] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[9] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1992).
[10] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[11] S. Trebst, D. A. Huse, and M. Troyer, Phys. Rev. E **70**, 046701 (2004).
[12] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).
[13] F. Wang and D. P. Landau, Phys. Rev. E **64**, 056101 (2001).
[14] Q. Yan and J. J. Pablo, Phys. Rev. Lett. **90**, 035701 (2003).
[15] M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos, Phys. Rev. E **66**, 056703 (2002); J. Chem. Phys. **119**, 9406 (2003).
[16] N. Rathore and J. J. de Pablo, J. Chem. Phys. **116**, 7225 (2002).
[17] N. Rathore, T. A. Knotts IV, and J. J. de Pablo, J. Chem. Phys. **118**, 4285 (2003).
[18] V. Varshney and G. A. Carri, Phys. Rev. Lett. **95**, 168304 (2005).
[19] Q. Yan, T. S. Jain, and J. J. de Pablo, Phys. Rev. Lett. **92**, 235701 (2004).
[20] P. Poulain, F. Calvo, R. Antoine, M. Broyer, and Ph. Dugourd, Phys. Rev. E **73**, 056704 (2006).
[21] A. Troster and C. Dellago, Phys. Rev. E **71**, 066705 (2005).
[22] D. Jayasri, V. S. S. Sastry, and K. P. N. Murthy, Phys. Rev. E **72**, 036702 (2005).
[23] C. Zhou, T. C. Schulthess, S. Torbrugge, and D. P. Landau, Phys. Rev. Lett. **96**, 120201 (2006).
[24] J. Kim, J. E. Straub, and T. Keyes, Phys. Rev. Lett. **97**, 050601 (2006).
[25] N. Nakajima, H. Nakamura, and A. Kidera, J. Phys. Chem. B **101**, 817 (1997); U. H. E. Hansmann, Y. Okamoto, and F. Eisenmenger, Chem. Phys. Lett. **259**, 321 (1996).
[26] K. Huang, *Statistical Mechanics* (Wiley, New York, 1972).
[27] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[28] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**,

3526 (1990).

[29] S. A. Larrass, L. M. Pegram, H. L. Gordon, and S. M. Rothstein, J. Chem. Phys. **119**, 13149 (2003).

[30] F. H. Stillinger and T. A. Weber, Phys. Rev. A **28**, 2408 (1983); Science **225**, 983 (1984).

[31] J. Kim and T. Keyes, J. Phys. Chem. B **111**, 2647 (2007).

[32] Z. Guo, D. Thirumalai, and J. D. Honeycutt, J. Chem. Phys. **97**, 525 (1992).

[33] F. Gulminelli and Ph. Chomaz, Phys. Rev. E **66**, 046108 (2002).

[34] B. J. Schulz, K. Binder, M. Müller, and D. P. Landau, Phys. Rev. E **67**, 067102 (2003).

[35] E. J. Barth, B. B. Laird, and B. J. Leimkuhler, J. Chem. Phys. **118**, 5759 (2003).

[36] J. G. Kim, Y. Fukunishi, and H. Nakamura, Phys. Rev. E **67**, 011105 (2003).

[37] J. G. Kim, Y. Fukunishi, and H. Nakamura, J. Chem. Phys. **121**, 1626 (2004); J. G. Kim, Y. Fukunishi, A. Kidera, and H. Nakamura, *ibid.* **121**, 5590 (2004).

[38] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon, Oxford, 1999).

[39] J. Norberg and L. Nilsson, Q. Rev. Biophys. **36**, 257 (2003).

[40] Y.-H. Lee and B. J. Berne, J. Phys. Chem. A **104**, 86 (2000).

[41] Z. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).

[42] Z. Guo and C. L. Brooks III, Biopolymers **42**, 745 (1997).

[43] H. Nymeyer, A. E. Garcia, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921 (1998).

[44] P. Amara and J. E. Straub, J. Phys. Chem. **99**, 14840 (1995).

[45] B. Vekhter and R. S. Berry, J. Chem. Phys. **110**, 2195 (1999).

[46] M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).

[47] T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G. J. Rylance, R. L. Johnston, and D. J. Wales, J. Chem. Phys. **122**, 084714 (2005).

[48] S. Y. Kim, S. J. Lee, and J. Lee, J. Chem. Phys. **119**, 10274 (2003).

[49] P. W. Pan, H. L. Gordon, and S. M. Rothstein, J. Chem. Phys. **124**, 024905 (2006).

[50] F. Calvo and J. P. K. Doye, Phys. Rev. E **63**, 010902 (2000).

[51] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).

[52] D. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).

[53] J. G. Kim, Y. Fukunishi, A. Kidera, and H. Nakamura, Phys. Rev. E **69**, 021101 (2004).

[54] O. M. Becker and M. Kaplus, J. Chem. Phys. **106**, 1495 (1997).

[55] B. W. Church and D. Shalloway, Proc. Natl. Acad. Sci. U.S.A. **98**, 6098 (2001).

[56] H. L. Gordon, W. K. Kwan, C. Gong, S. Larrass, and S. M. Rothstein, J. Chem. Phys. **118**, 1533 (2003).